SISG 2014

SISG Module 21: Coalescent Theory

19th Summer Institute in Statistical Genetics

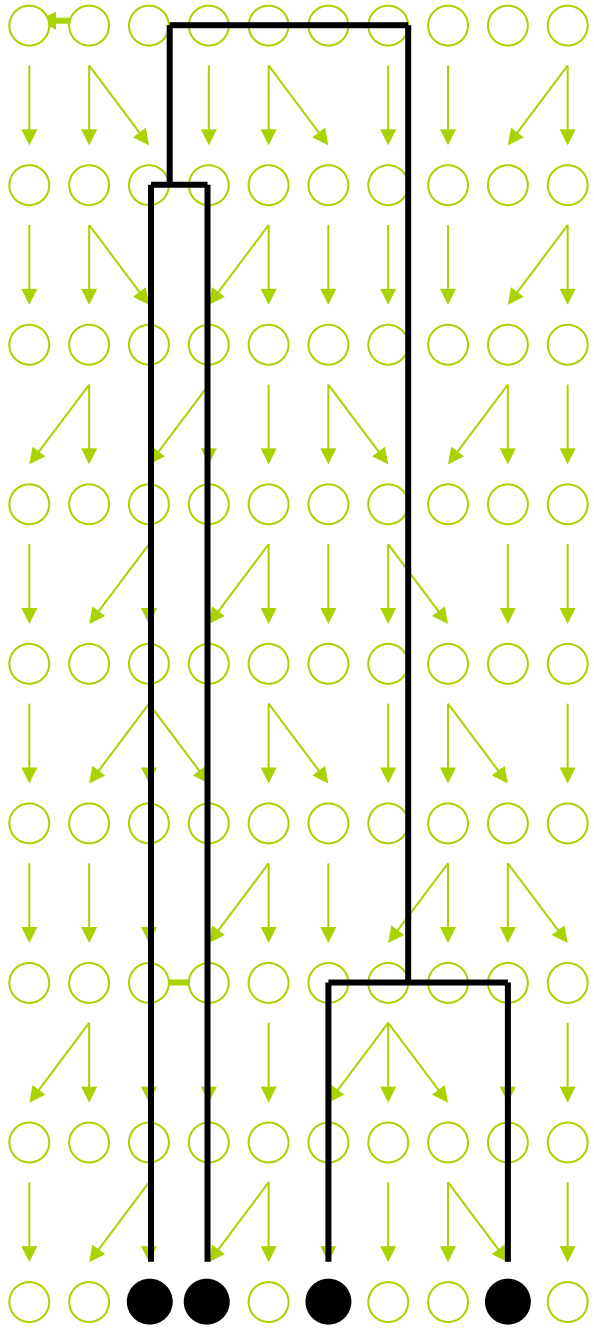
UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

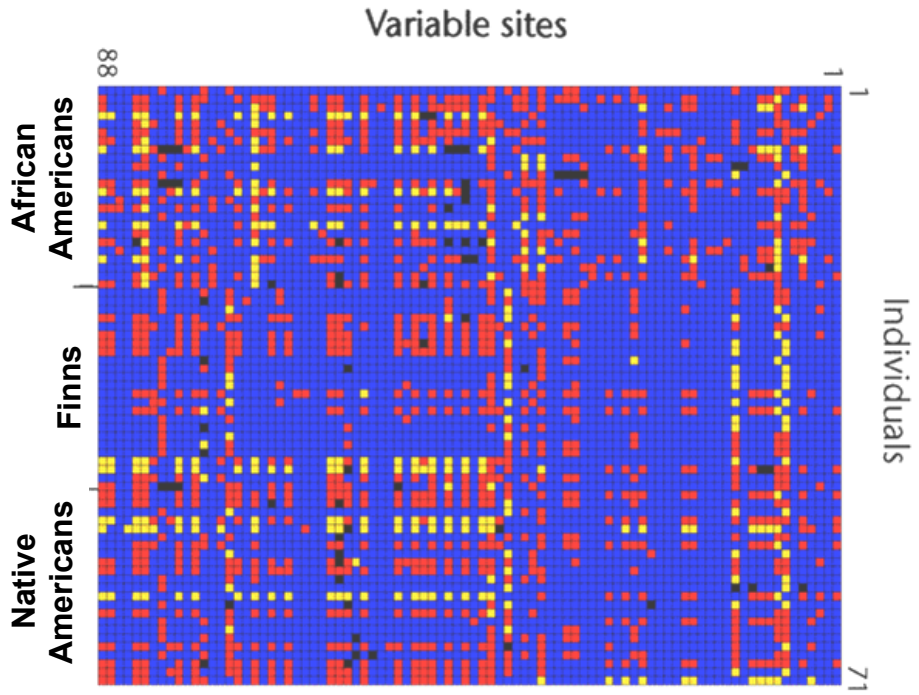
The coalescent

Understanding genetic variation

Philip Awadalla
University of Montreal



Making sense of genetic variation



- 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?

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

The aims of population genetics

- To understand the link between genetic variation and phenotypic variation
  - Is variation at this gene associated with disease susceptibility?
  - Which loci contribute the variation in hair colour?
- To investigate the evolutionary history of a species
  - How long have these populations been separate?
  - Which genes have experienced recent adaptive evolution?
- To learn about fundamental biological processes
  - How does the recombination rate vary along the genome?
  - What determines the mutation rate?

The raw material of population genetics

- Population genetics is the study of **naturally** occuring genetic variation

- Genetic variation comes in all shapes and sizes
  - Re-sequencing v. SNPs v. microsatellites
  - Recombining v. partially linked v. unlinked markers
  - Single gene v. multiple loci v. whole genome
  - Single population v. multiple population v. multiple species

- Which data type you collect (and analyse) depends on the questions you want to ask

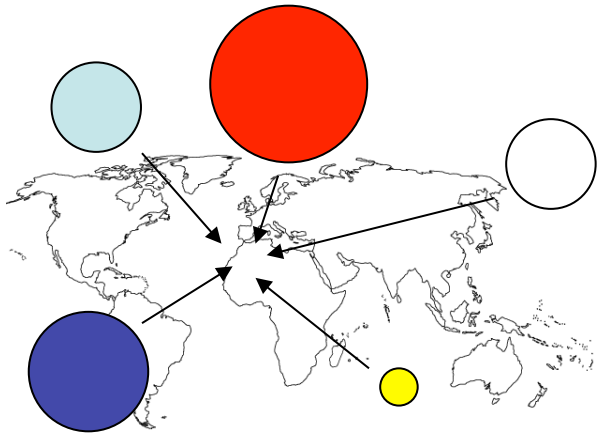
An example: structuring of human populations

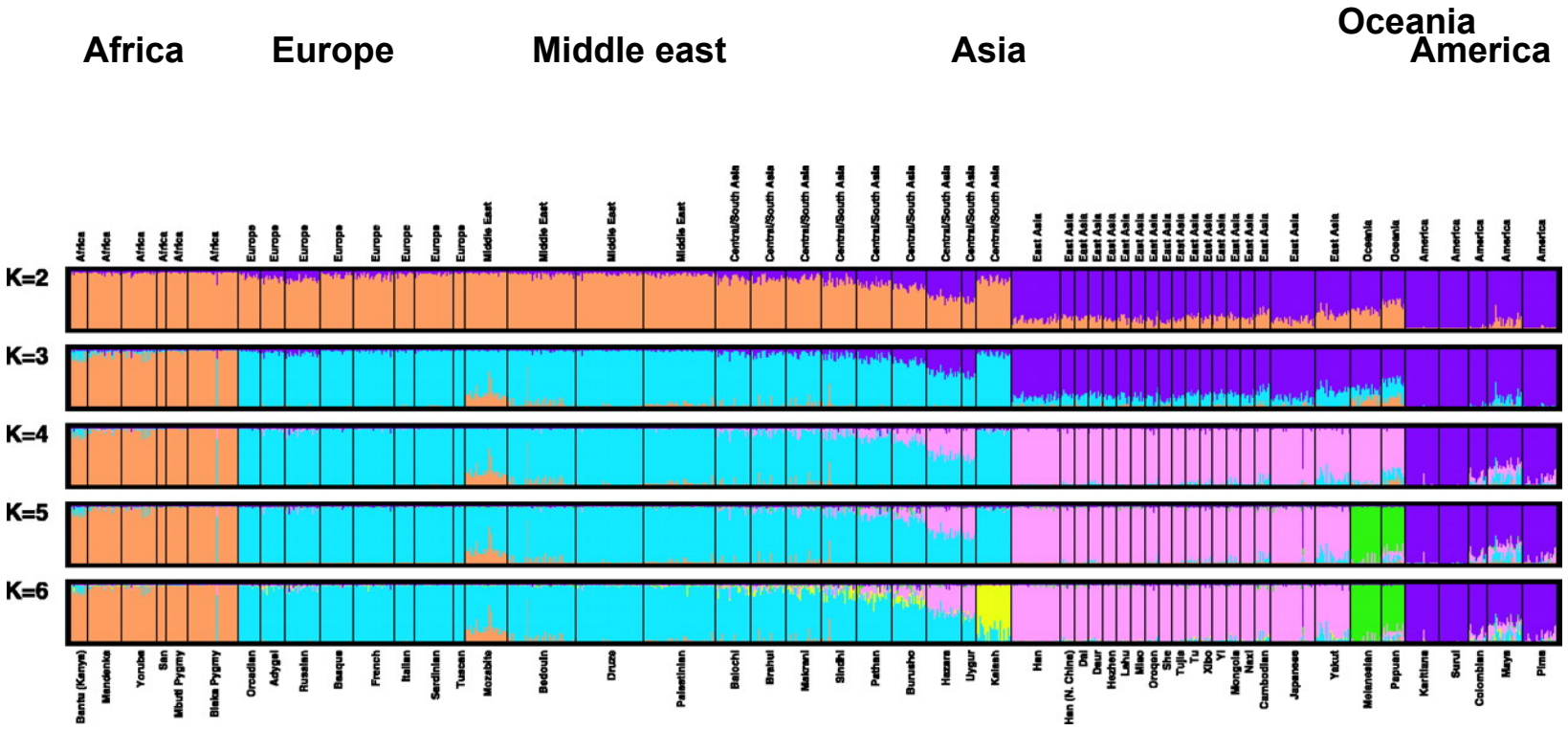
- Questions
  - Is there significant natural structuring to genetic variation in humans?
  - Does this structuring coincide with geographical boundaries?

- Data
  - 377 autosomal microsatellite loci in 1056 individuals from 52 populations. Rosenberg et al (2002)

- Model
  - K 'Hidden' populations in linkage and Hardy-Weinberg equilibrium

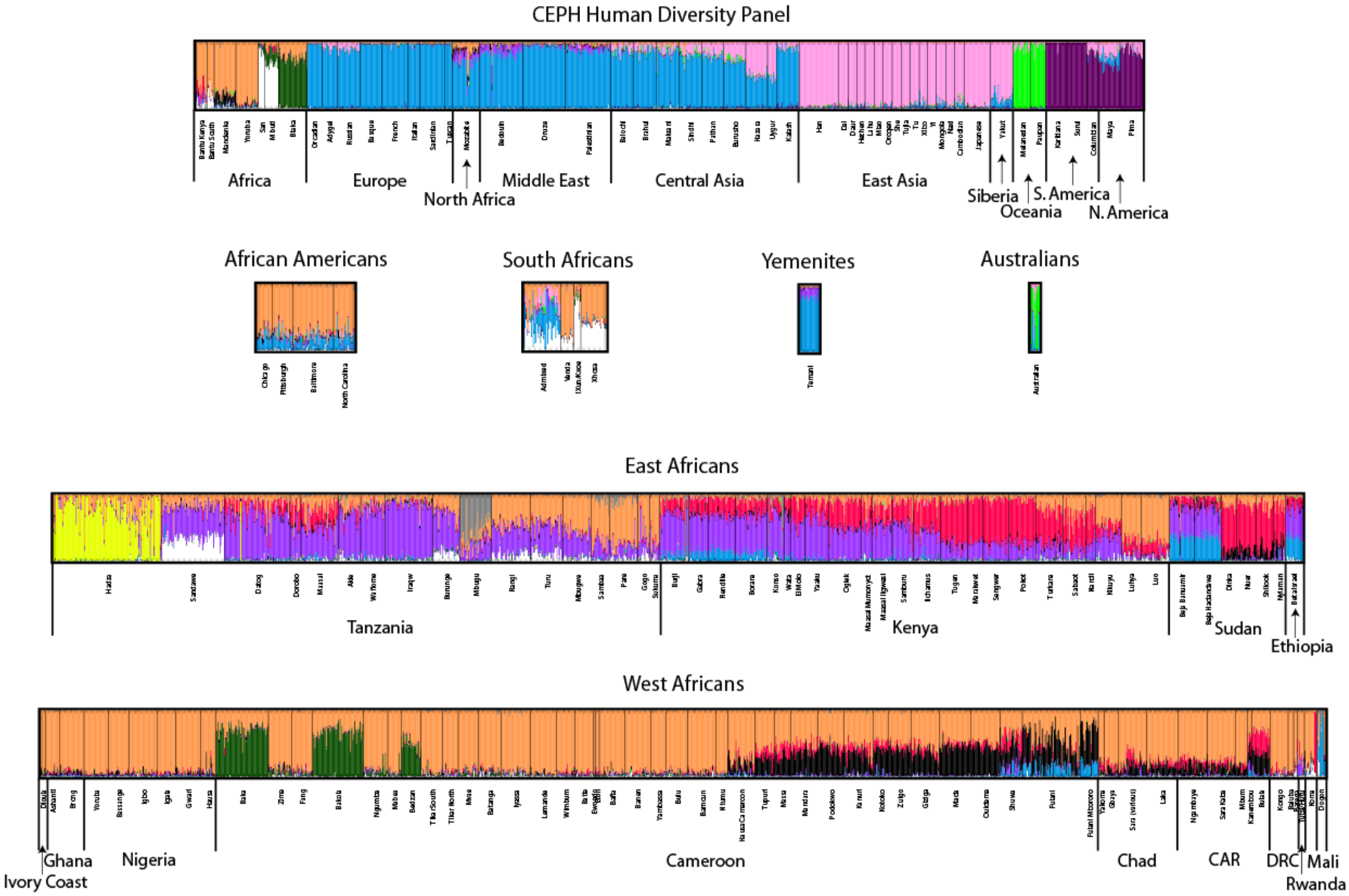
- Estimation
  - Estimate population allele frequencies
  - Most likely value of K
  - Posterior probability for each individual





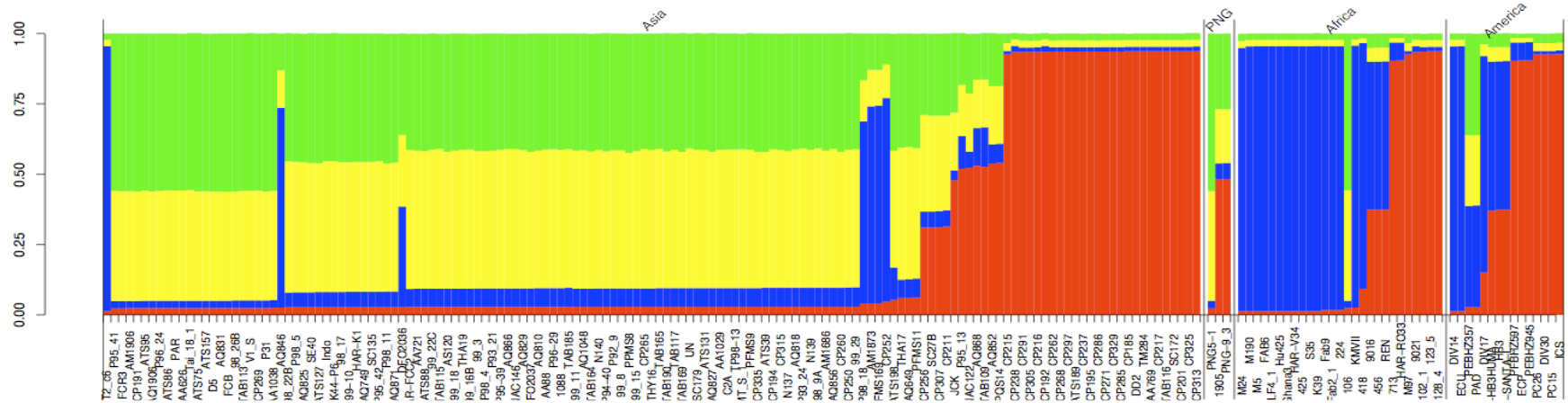
Science 298: 2381

Human Population Structure Globally

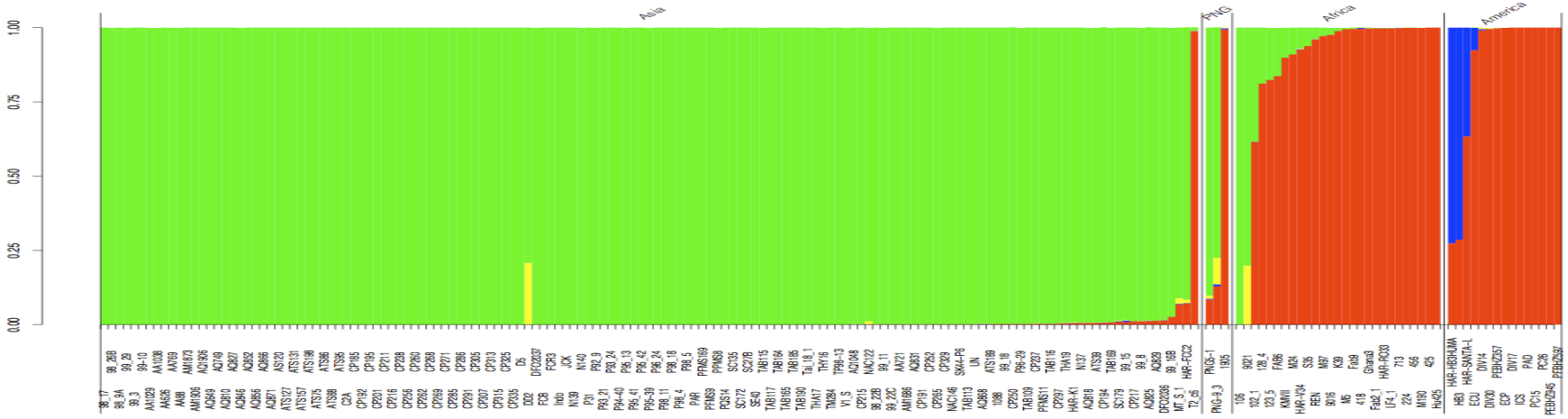


Population structure at antigens vs. genome-wide

Erythocyte Binding Antigens

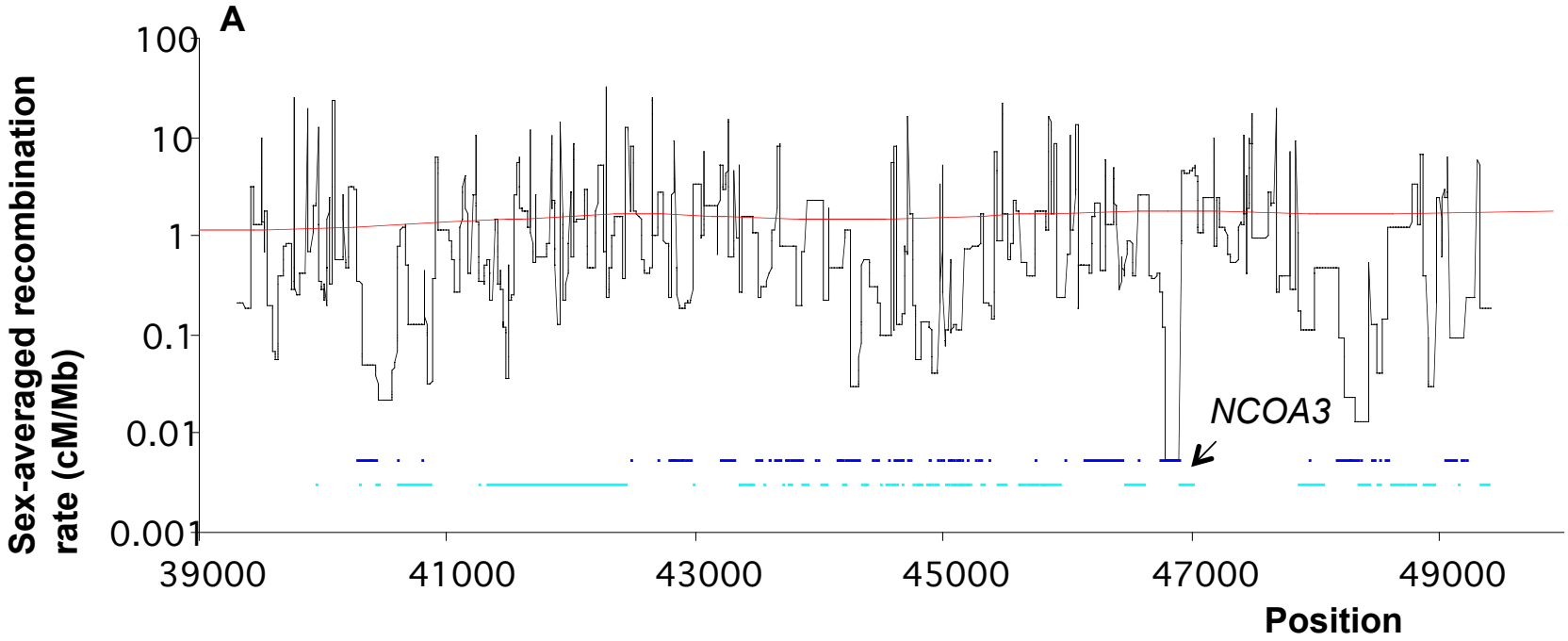


Genome-wide SNPs



An example: recombination rates in humans

- Questions
  - How does the recombination rate vary over kilobase scales in humans?
  - What is the frequency of recombination hotspots?
- Data
  - Several thousand SNPs typed across a 10Mb region of chromosome 20 (McVean et al 2004) in two populations
- Model
  - Coalescent model of recombination in a single population
- Estimation
  - MCMC scheme for estimating and representing uncertainty



Science 304: 581

How do we make sense of data?

- There are two basic ways of looking at data
- **Nonparametric methods** try to say things without relying on models
- **Parametric methods** make explicit models for the data

- For example, in phylogenetics, some people use parsimony (a nonparametric approach) and others use likelihood (a parametric approach) to estimating trees
- In population genetics, we will come across nonparametric methods (for example in looking at recombination), but most methods are model-based

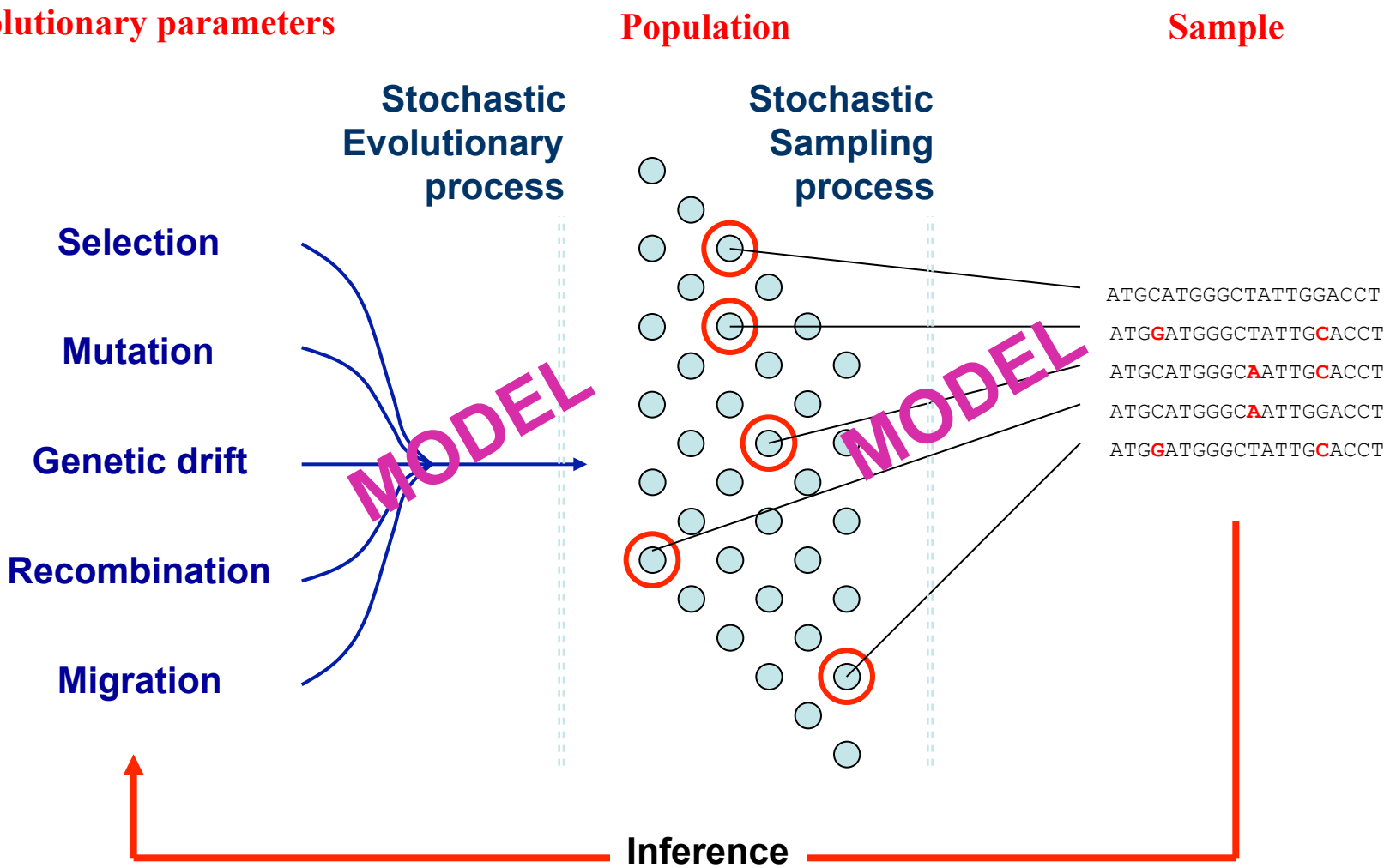
Parametric

- Explicit form of underlying distribution is assumed
- Model characterised by parameters
- Use of model allows precise statements to be made
- Inferences and predictions are wrong if model is wrong!

Non-parametric

- Explicit form of underlying distribution is NOT assumed (though certain features may be – e.g. symmetry)
- Distribution characterised by moments and empirical CDF
- Greater robustness at expense of power
- Inferences and predictions independent of exact underlying process

Population genetic inference



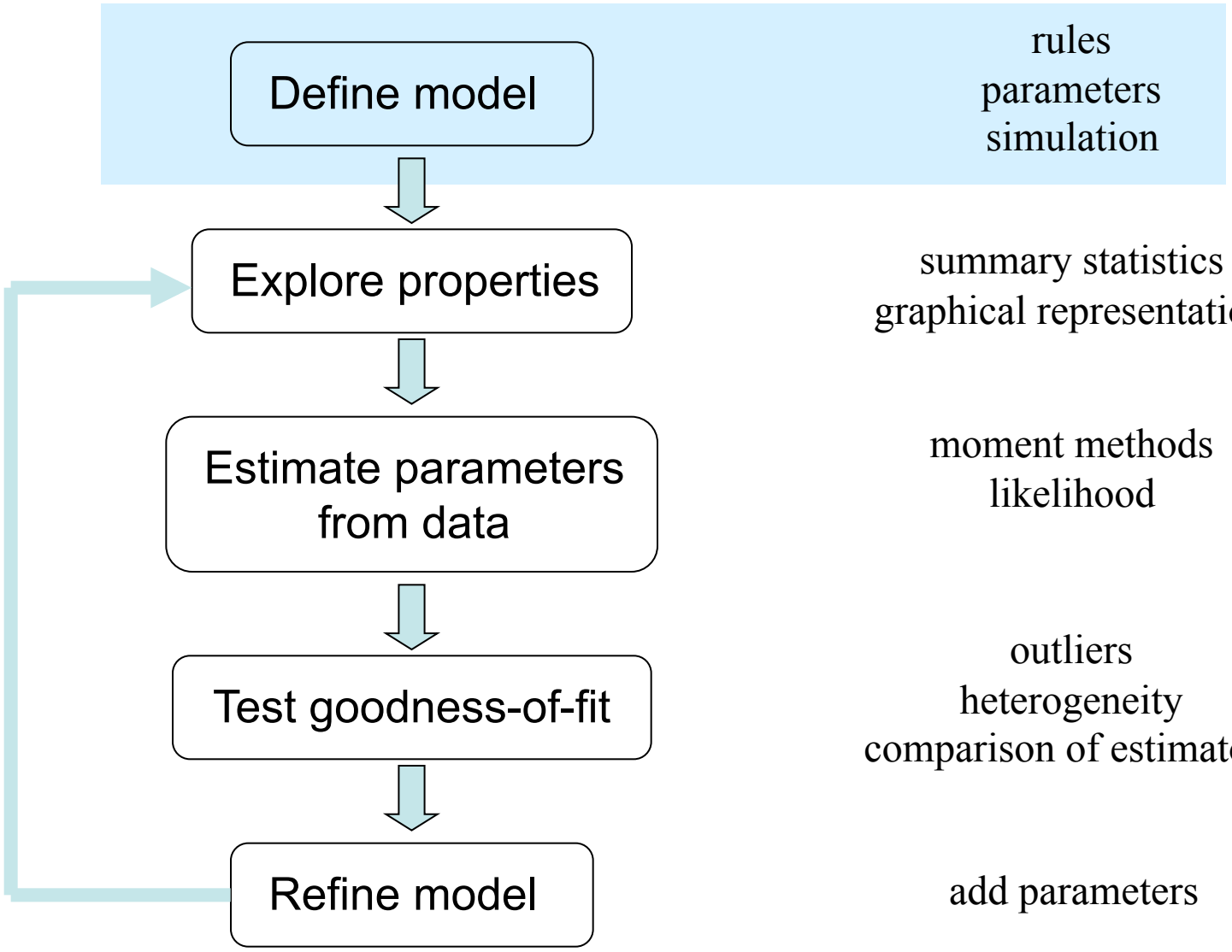
The basic model of population genetics

Nothing interesting ever happens in biology

Models in population genetics

- The natural approach to modelling is to start with the simplest possible model that you think could adequately describe the data.
  - No population structure, no recombination, no selection
- The first step is to estimate some parameters of the **null** model (assuming it is correct).
  - The mutation rate and population size
- We can then ask interesting questions
  - Is this simple model an adequate description of the data?
  - If not, what does the deviation from the null model tell us about the forces that are acting?

Statistical inference

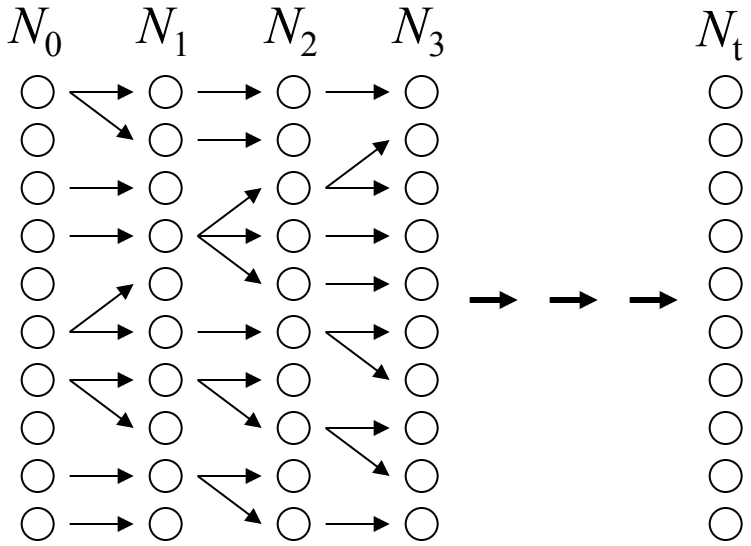


The unusual nature of population genetic data

- Conventional statistical inference
  - Many independent data points
  - Sample space of low dimensions
  - Analytical formulations for inference using all possible information often possible

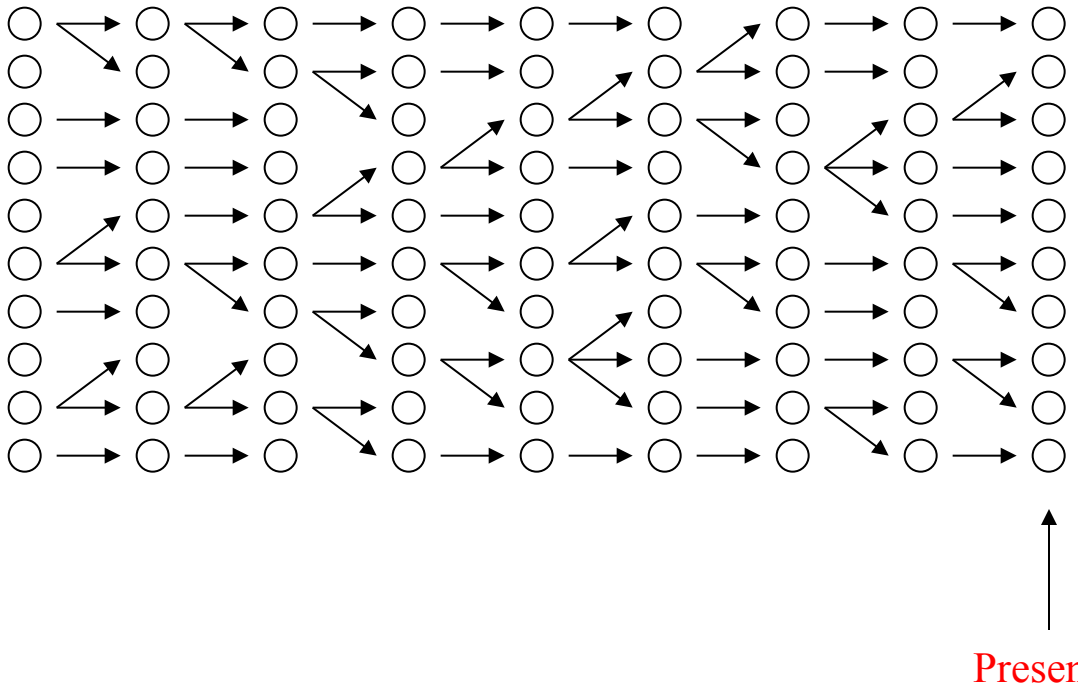
- Population genetic data
  - Typically a single draw from the evolutionary process
  - Sample space of many dimensions
  - Analytical formulations for inference using ALL possible information usually impossible to derive

The Wright-Fisher population model

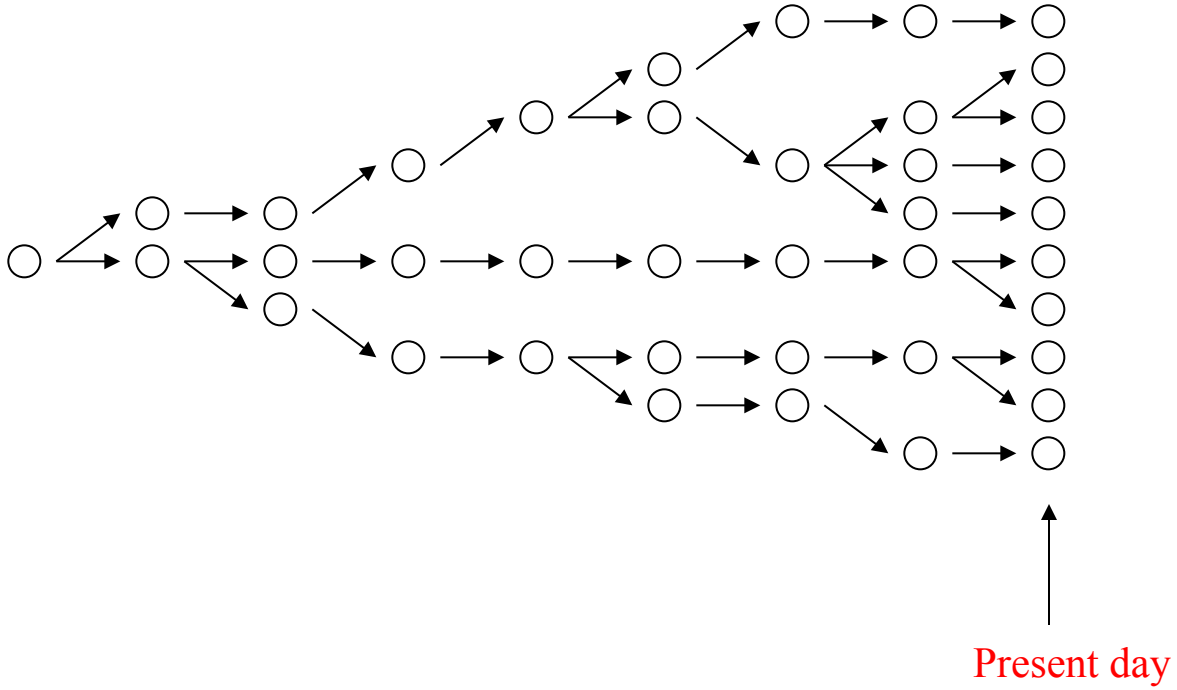


- Diploid Individuals reproduce by sexual reproduction with possibility of selfing
- Mating is random with respect to location and genotype
- Generations are non-overlapping (everyone reproduces simultaneously)
- The population size is constant of size N ($2N$ alleles)
- There is no migration or selection

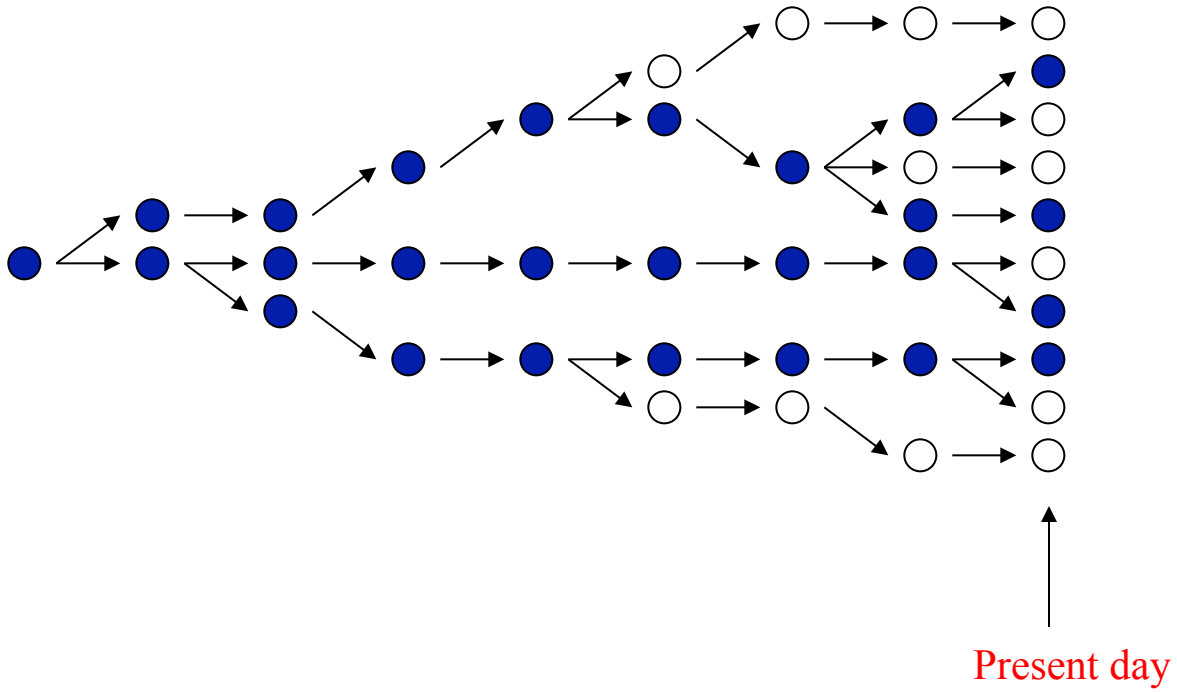
Genes in populations



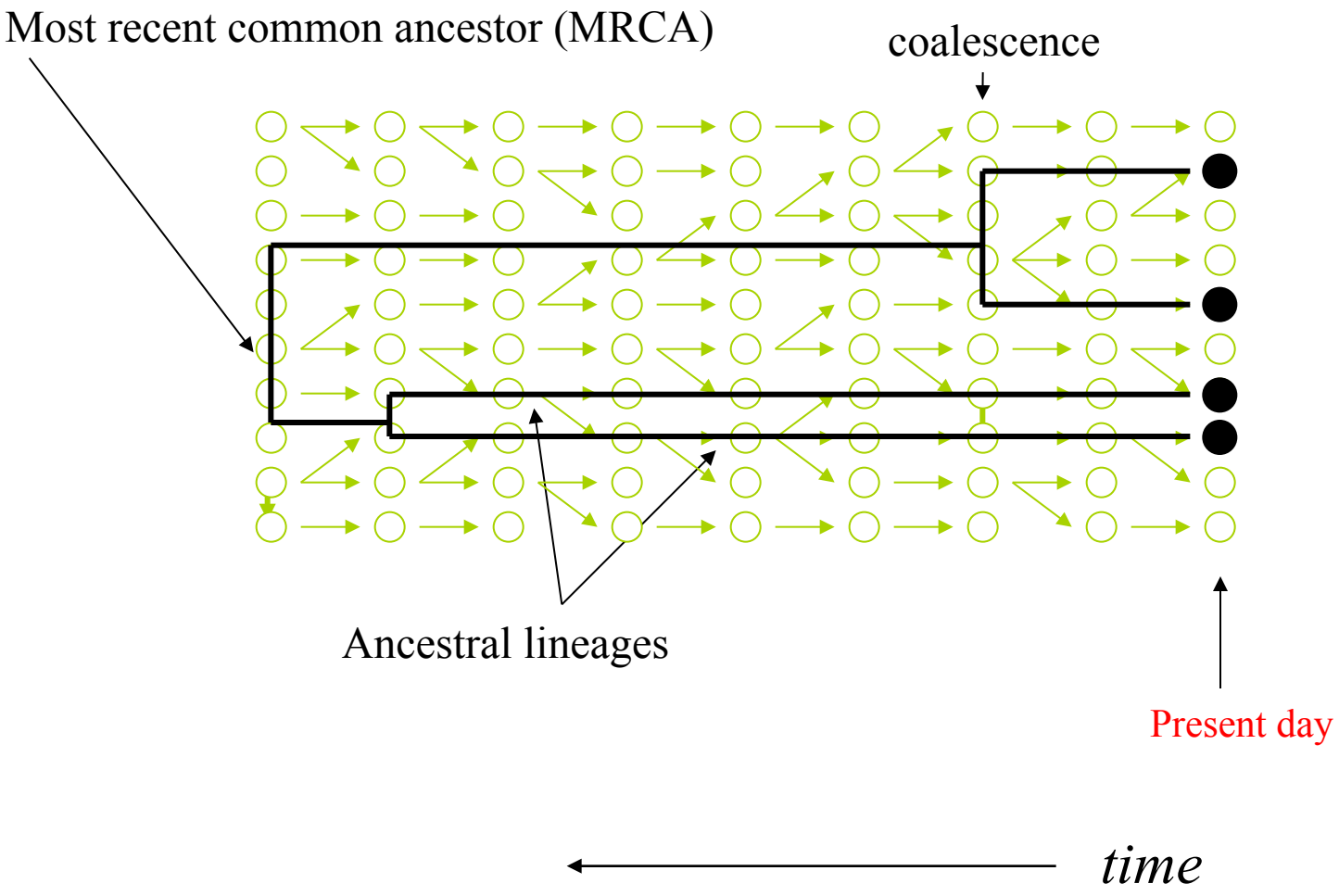
Ancestry of current population



Ancestry of sample



The coalescent: samples in populations



Characteristics of the coalescent

- The coalescent is a mathematical description of the genealogical process arising in idealised populations
- It focuses on the **genealogy** (or tree) underlying the history of a sample of chromosomes
- It is a **probabilistic** model, which implies that it describes the distribution of genealogies
- The principle idea is that the genealogy holds **all** the information we could ever know about our population and its biology
  - We don't know the genealogy, but if we did, inference would be simple

The coalescent for two sequences

- Consider the coalescent process for two sequences

- In one generation, these two sequences either came from the same parent chromosome, i.e. coalesced, (with probability 1/2N) or didn't (with probability 1-1/2N)

- The probability of coalescence t generations ago is just

$$= \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$$

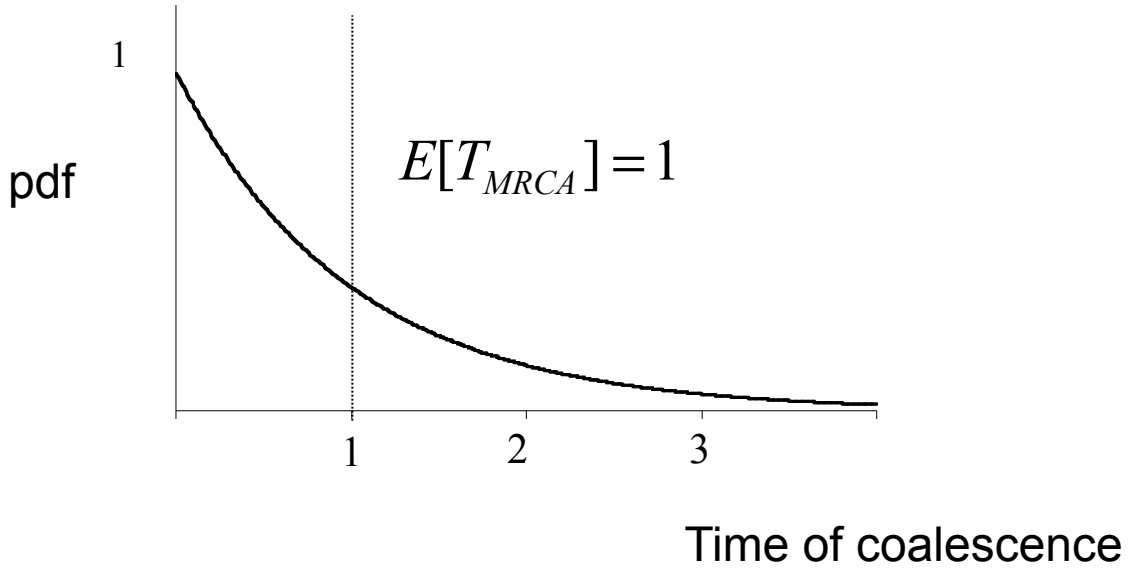
Not coalesced for first t-1 generations

Coalesce in next generation

The distribution of coalescence times

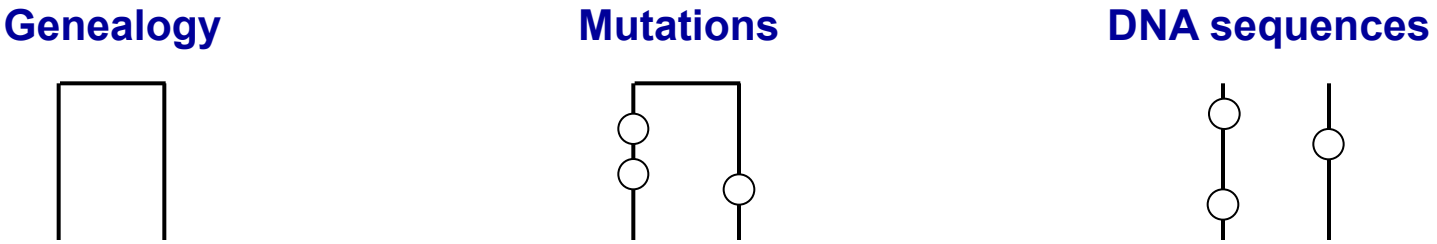
- In the case of discrete generations, the coalescence times takes a **geometric** distribution

- In the continuous time limit, which is really what coalescence theory deals with, the distribution of coalescence times takes an **exponential** distribution (time is scaled in units of 2N generations)



Adding mutations

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



- Expected number of differences between a pair of sequences

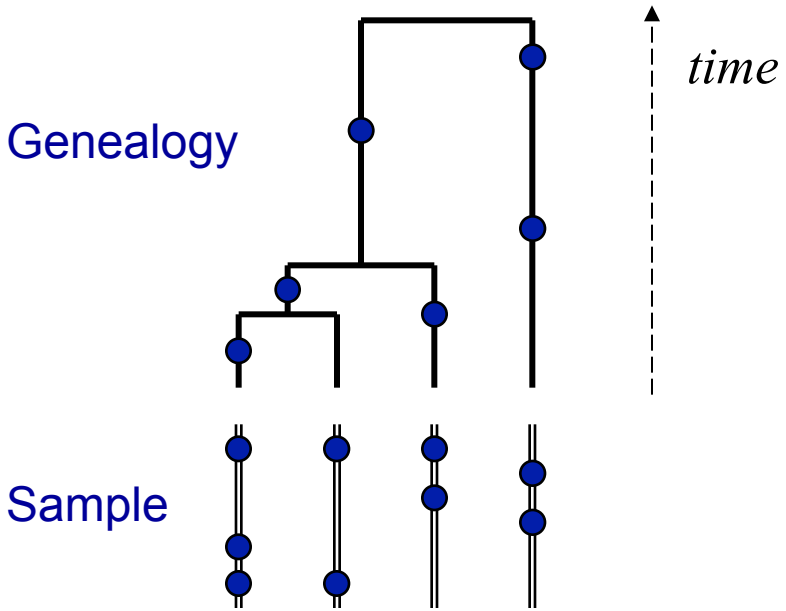
$$E[\pi] = 2 \times u \times E[T_{MRCA}]$$
$$= 4Nu$$

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

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

- When there are n sequences (as opposed to two) it is assumed that each of the $n(n\text{-}1)/2$ pairs coalesce independently

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

Number of pairs of lineages

Probability of a given pair coalescing

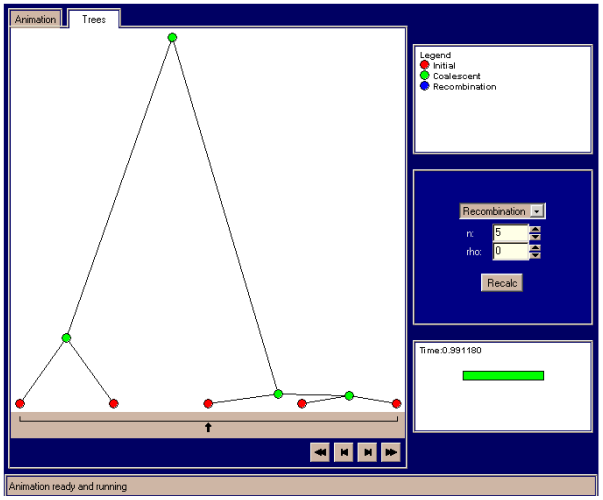
- This means that the time until the first coalescence is exponentially distributed with rate $n(n\text{-}1)/2$

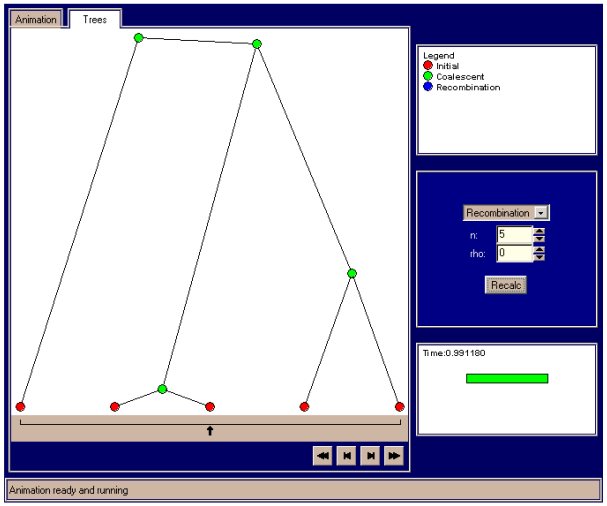
- The process is **Markov**, which means that after the first coalescence, it is like starting again for n-1 sequences

Simulating coalescent histories

- Stochastic (or Monte Carlo) simulation can be used to learn about the distribution of coalescent histories
  - www.coalescent.dk

- We learn that coalescent genealogies are very variable!

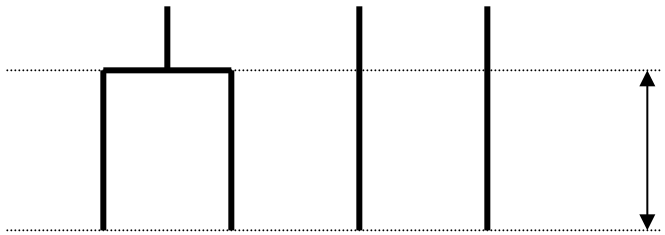




- Much of the variation in the depth of the tree is due to variation in the time for the last coalescent event to occur

Expected times and mutations

- The expected time until the next coalescence is $E[T_{co}] = \frac{4N_e}{n(n-1)}$
- So, the expected number of mutations that occur in the genealogy during this time is



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

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

Number lineages Total mutation rate

- Because it is a Markov process, we can just add up the expected number of mutations contributed by each step in the coalescent tree

Properties of the n-coalescent

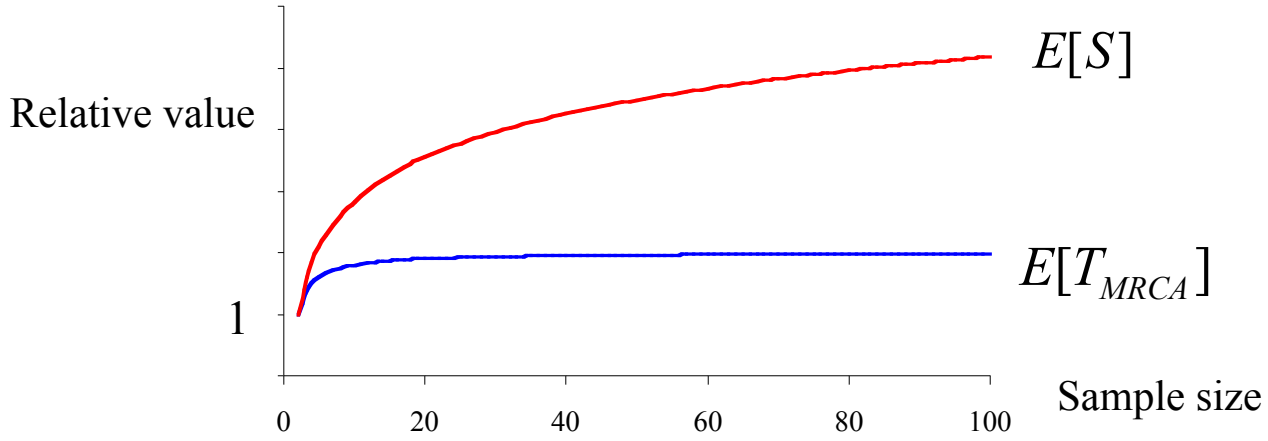
- The expected total number of segregating sites is the sum over each coalescent interval

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

Watterson (1975)

- The expected time until the MRCA for *n* sequences is

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

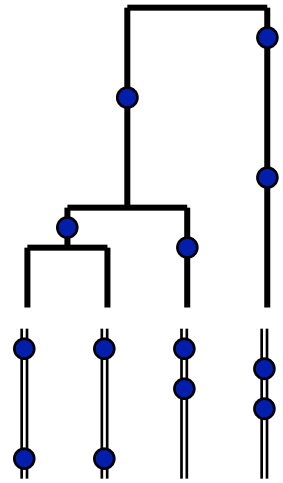


Mutations, alleles and haplotypes

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

$$E[K] = 1 + \frac{\grave{e}}{1+\grave{e}} + \frac{\grave{e}}{2+\grave{e}} + \ldots\ldots + \frac{\grave{e}}{n-1+\grave{e}}$$

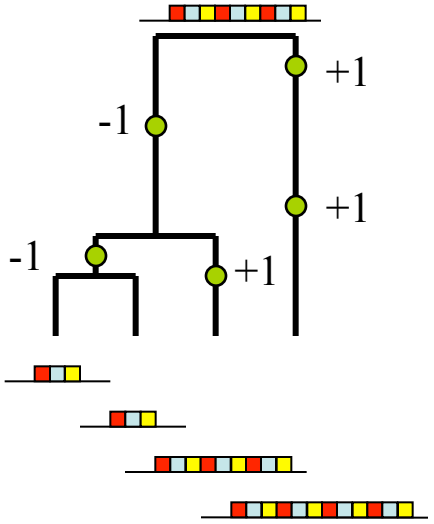
Ewens (1972)



- Microsatellites
  - Step-wise mutation model

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

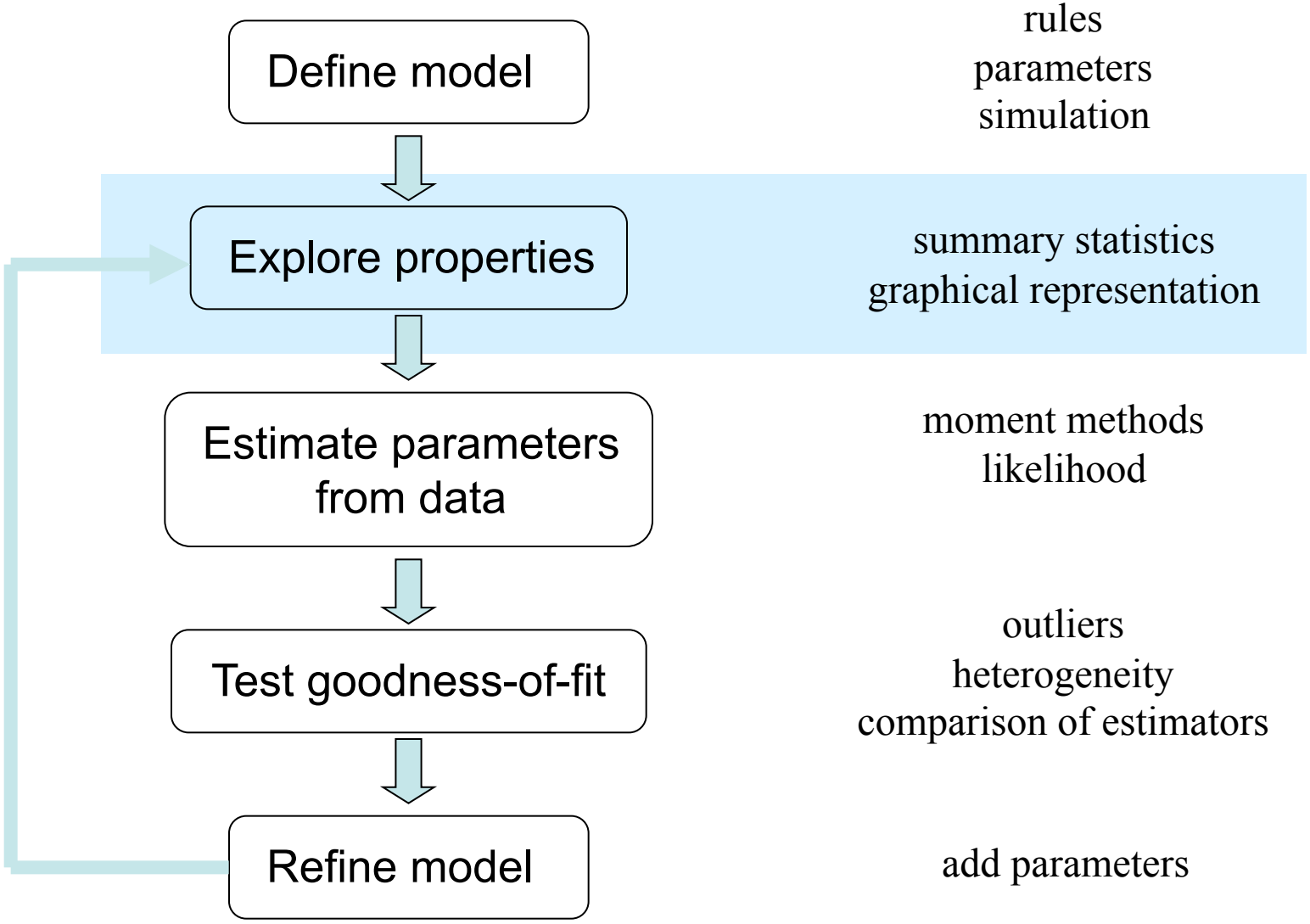
Moran (1975)
Slatkin (1995)



Summary

- Population genetics is the study of naturally occurring genetic variation
- Model-based estimation can be a powerful tool in learning about the forces important in shaping diversity
- The coalescent is a null model for describing the distribution of genetic variation in samples from unrelated individuals in idealised populations
- The key idea is that the underlying genealogy relating sequences in a sample is the fundamental quantity of interest
- The model captures the variability in gene genealogies and, by mapping neutral mutation process on top of the genealogy, patterns of variation

Statistical inference



Issues

rules
parameters
simulation

summary statistics
graphical representation

moment methods
likelihood

outliers
heterogeneity
comparison of estimators

add parameters

What next?

- We have a model for genetic variation, now we want to use it to learn things about real data

- We need to ask two central questions
  - Suppose the model is correct, what does the data tell us about the parameters of the model?
  - Does the model provide an accurate description of the data?

- How do we ask them?
  - We need to summarise the data in a way that is informative about the parameters of the model and model adequacy
  - We need to learn about the distribution of those summaries under the model (so that we can ask if our observations are unusual)

Summarising data

- Here is some data from human Xq13 (Kaessman et al 1999). A 10kb region was sequenced in 69 individuals from across the world.

Number sequences, Number of polymorphic sites

```
69 33 1
```

Locations of polymorphic sites

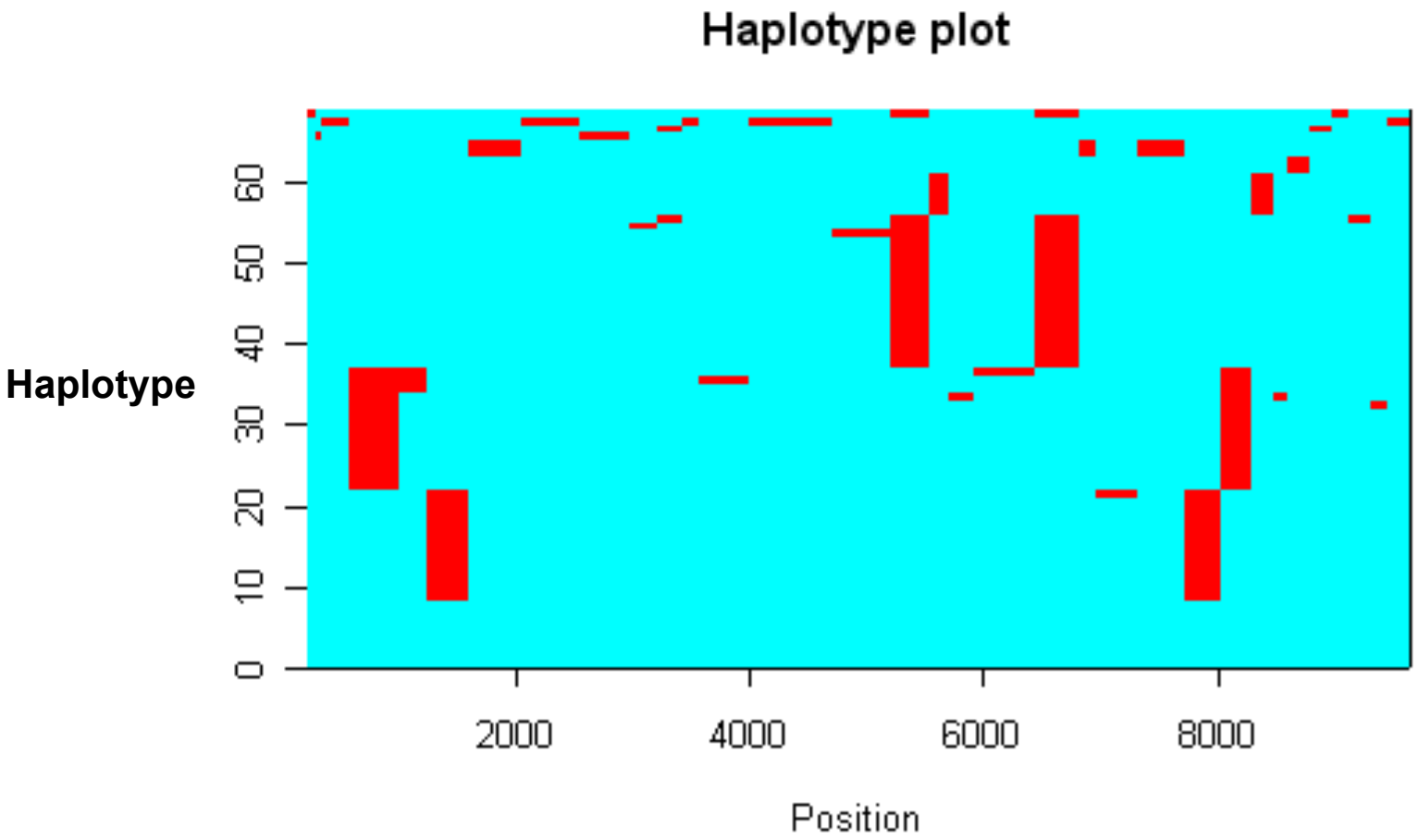
```
270 351 371 820 1230 1301 1896 2234 2890 3099 3337 3549 3583 4429 5013 5400 5688 5719 6142 6737 6882 7046 7607 7847
8220 8380 8554 8620 8965 8994 9247 9363 9541
```

Polymorphic sites represented as 0s and 1s

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
…
```

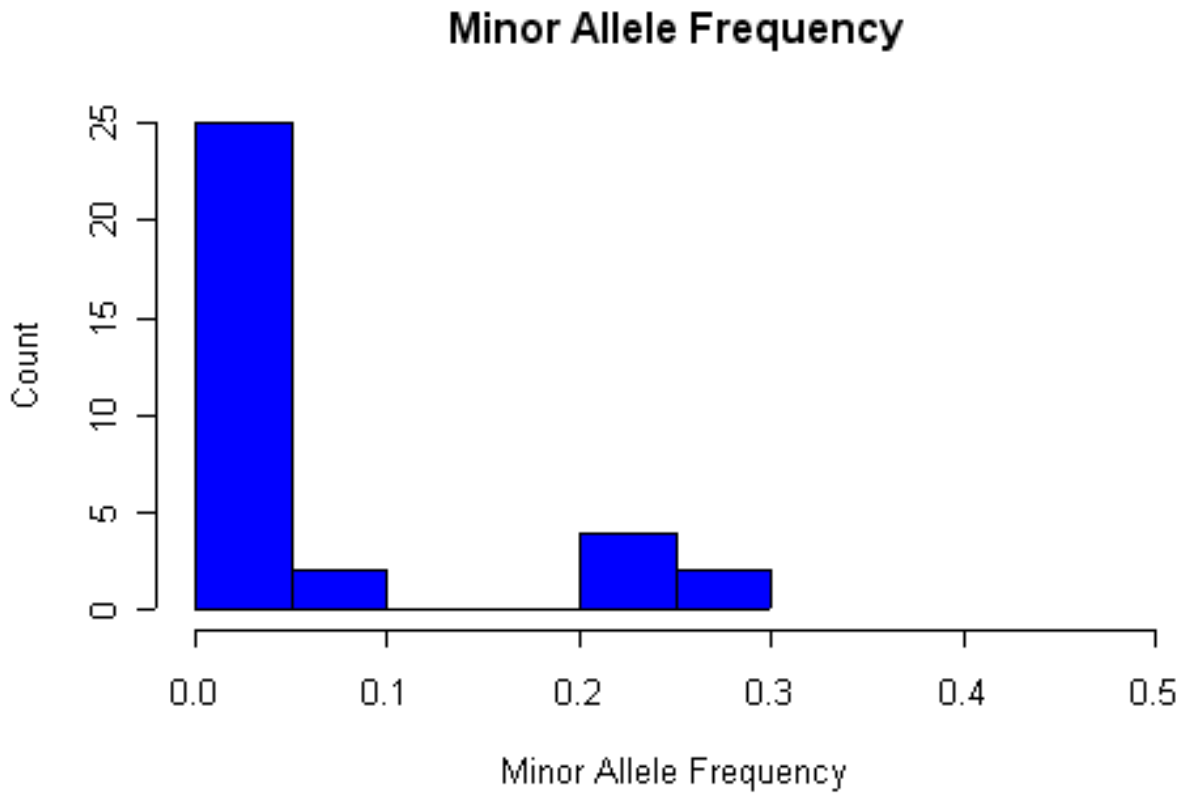
Visualising the raw data

- It can be helpful just to look at the haplotypes (genotypes)



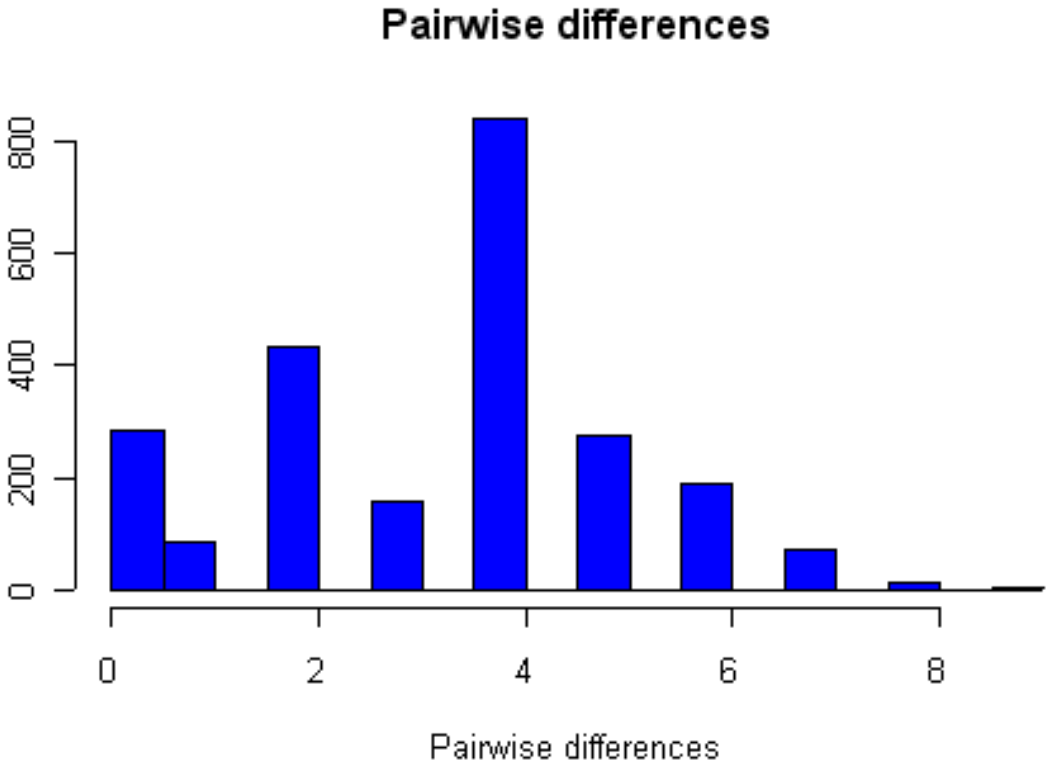
The minor allele frequency spectrum

- You will see in a bit that the coalescent tell us what to expect of the distribution of allele frequencies (or minor allele frequencies if we do not know which allele is derived)



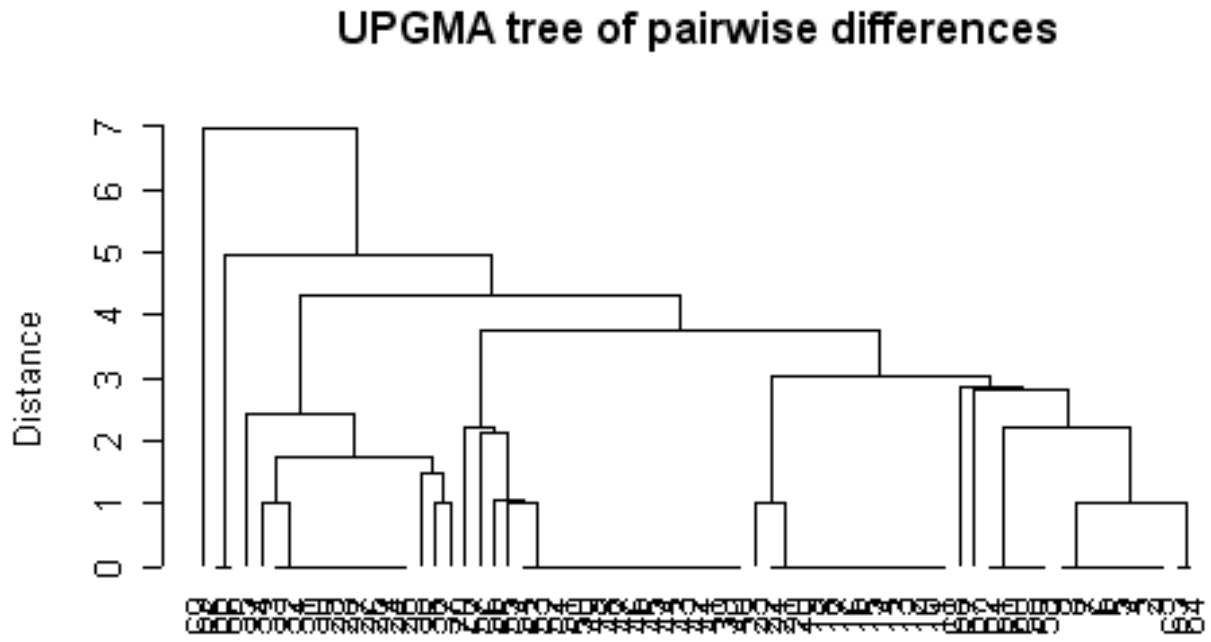
The distribution of pairwise differences

- The coalescent tell us what to expect for moments (mean, variance, etc.) of the distribution of pairwise differences



Reconstructing the genealogy

- We can also look for structure in the data by using simple clustering algorithms such as UPGMA
  - Note that this is only going to ‘reconstruct’ the genealogy if there is no recombination and lots of mutation



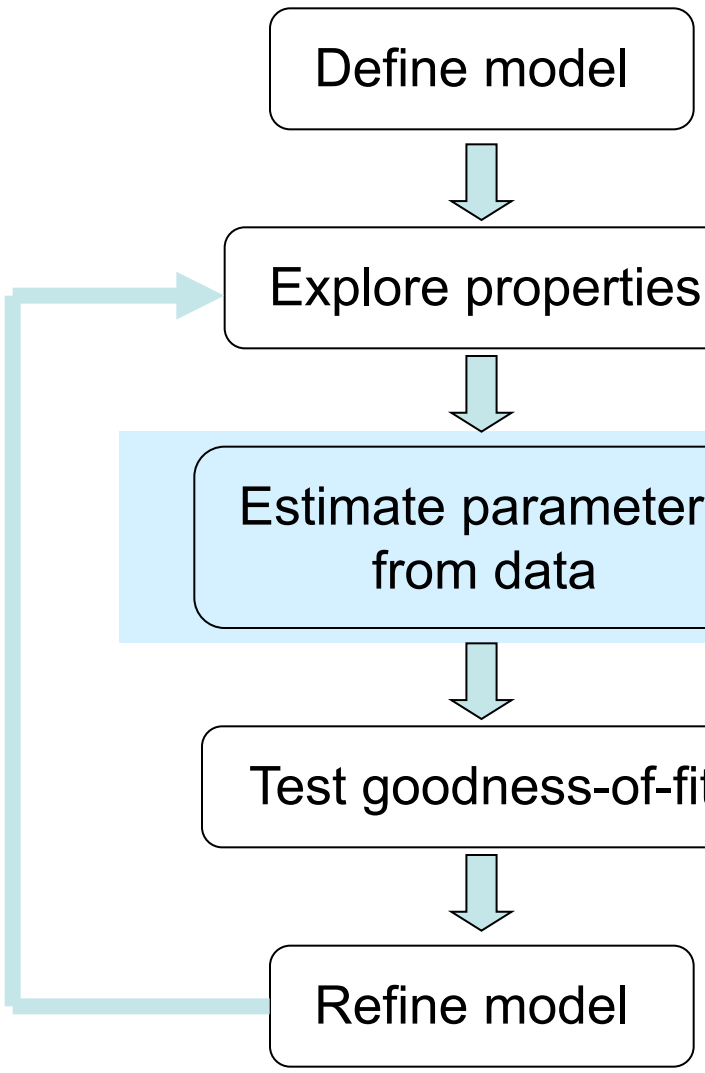
Single number summaries

- There are several useful single-number summaries (summary statistics) of the data
  - The number of segregating sites
  - The average pairwise differences
  - The number of singletons (sites at which only one individual differs from all others)
  - Neutrality statistics such as Tajima's D and Fu and Li's D*

```
Number of sequences = 69
Number of polymorphic sites = 33

Watterson's estimate of theta = 6.86919225575433
Average pairwise differences = 3.3759590792838 9
Number of singletons = 19
Tajima D statistic = -1.63320078310981
Fu and Li D statistic = -3.3243664840525
```

Statistical inference



Issues

rules
parameters
simulation

summary statistics
graphical representation

moment methods
likelihood

outliers
heterogeneity
comparison of estimators

add parameters

Estimating parameters

- The summaries of the data we have considered are good because they can be directly related to properties of the neutral coalescent model

- For example, we know that the expected pairwise differences is just the parameter q, so the sample statistic is an **unbiased** estimate of the parameter

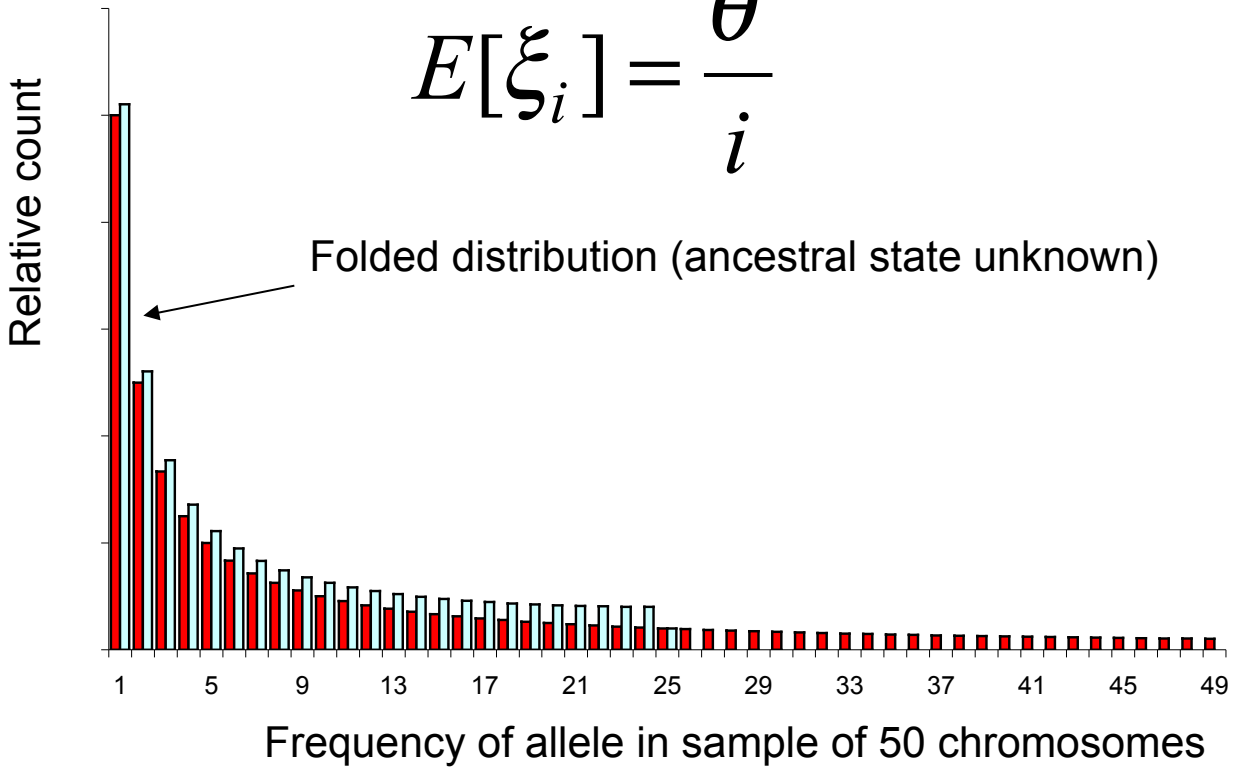
- Likewise, the expected number of segregating sites is $E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$

- So $S / \sum_{i=1}^{n-1} \frac{1}{i}$ is also an unbiased estimate of q

The expected allele frequency distribution

- The expected number of mutations with derived frequency *i* is (Tajima 1995)

$$E[\xi_i] = \frac{\theta}{i}$$



- Note, for a single (non-recombining) sample, you do not expected to see the expected distribution!

Other ways of estimating theta

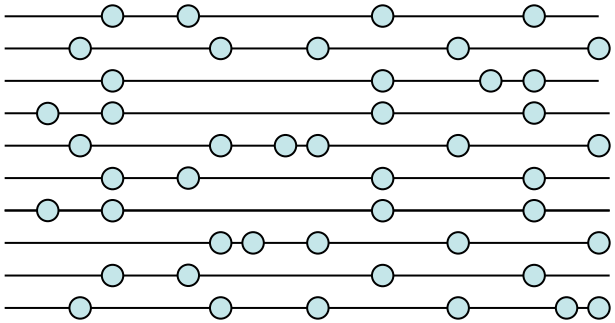
- We have considered a few **moment estimators** of theta. These compare statistics of the data to their expectations under the model

- In general, moment methods are a quick but often poor way of estimating parameters because they throw away much of the information in the data

- We can use more information by using **likelihood** methods. These calculate the probability of observing the data (or a summary of it) given the model. A natural choice of estimator is the **maximum likelihood** value

An example

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



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

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

- Note, the estimator based on the average pairwise diversity has higher variance (and is not consistent), so it is generally a worse estimator

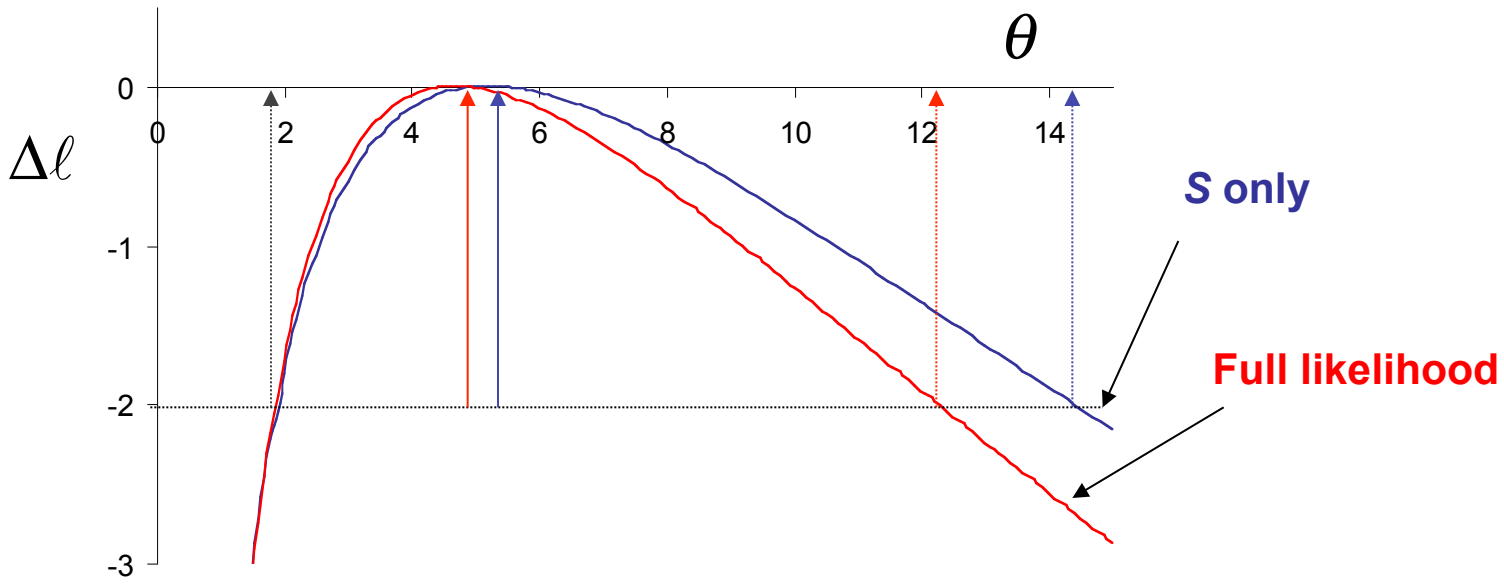
Likelihood estimation of *q*

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

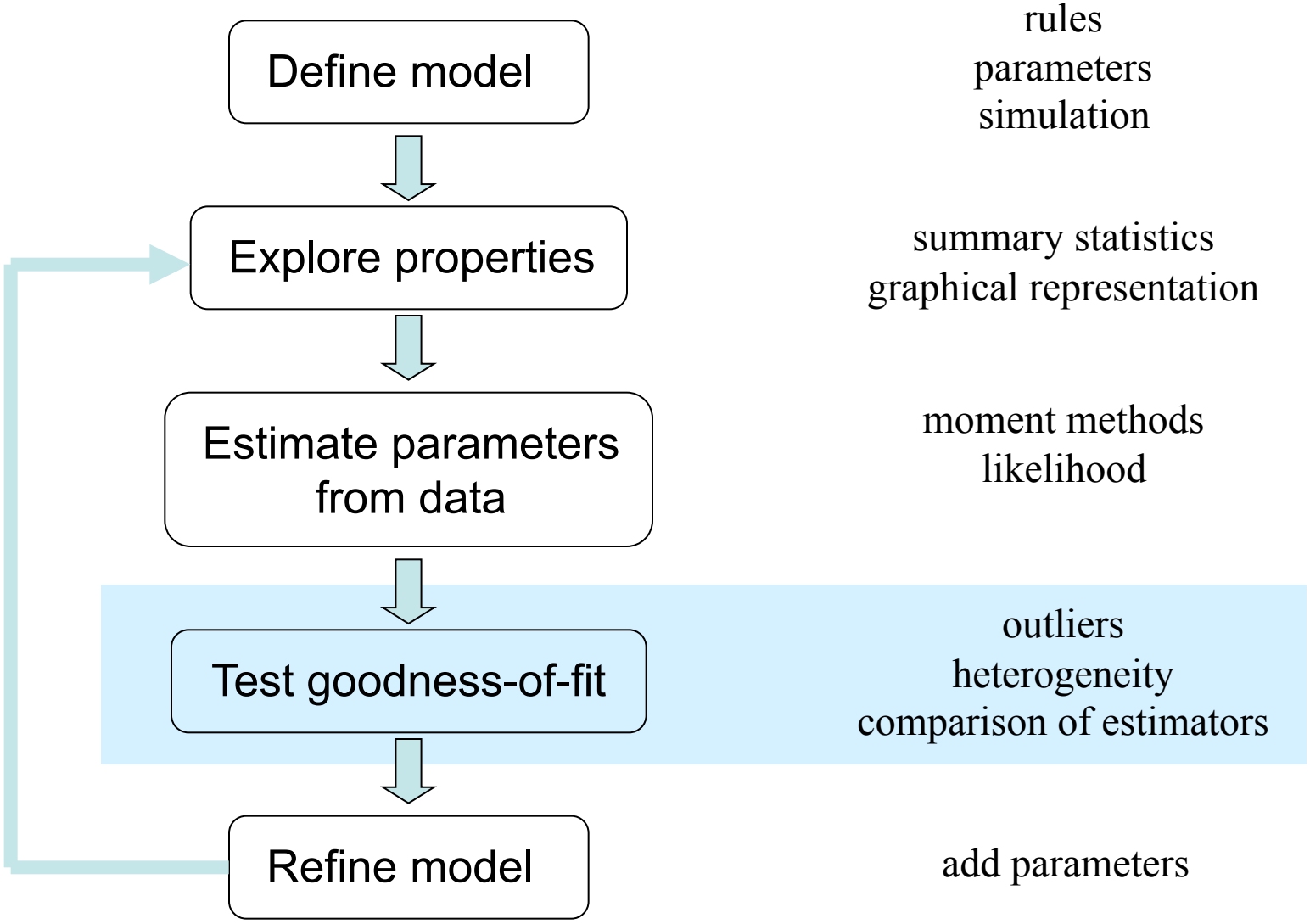
$$\hat{\theta} = 5.2$$
$$2U = 1.9 - 14.4$$

- Using the full-likelihood method of Griffiths and Tavaré (1994)
  - Implemented in GENETREE software

$$\hat{\theta} = 4.7$$
$$2U = 1.9 - 12.2$$



Statistical inference



Testing the model

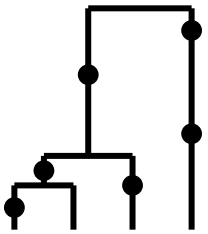
- Given our parameter estimates, we want to ask how good the model is at describing the data

- This is an example of a **goodness of fit** test. Generally, the idea is to take some other part of the data and ask how likely it is to see something as extreme as the observed value given the model (and the parameter estimates)

- A widely used approach in population genetic is to compare estimators of theta
  - Tajima D statistic compares estimators from the number of segregating sites and pairwise diversity
  - Fu and Li D* statistic compares estimates from the number of segregating sites and the number of singletons

Tajima's *D* test (1989)

- Two estimators of theta
  - Watterson's estimate: number of segregating sites
  - Average pairwise diversity: sensitive to intermediate allele frequencies



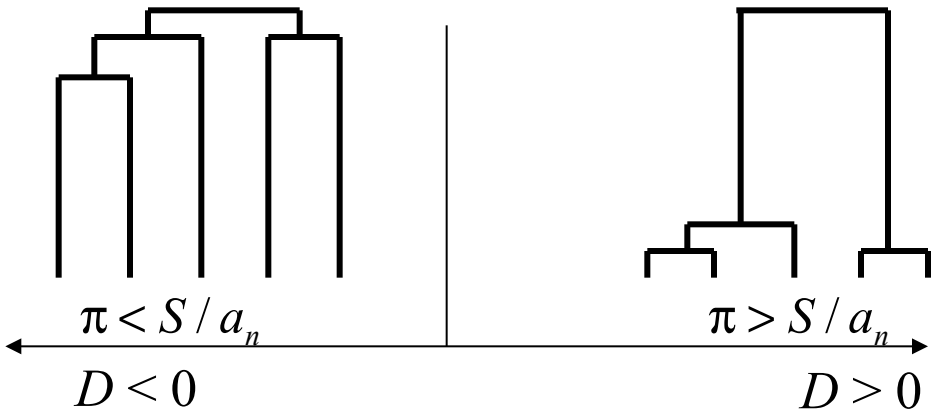
$$E[\pi] = \theta$$

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

$$D = \frac{\pi - S / a_n}{\sqrt{\mathrm{Var}(\pi - S / a_n)}}$$

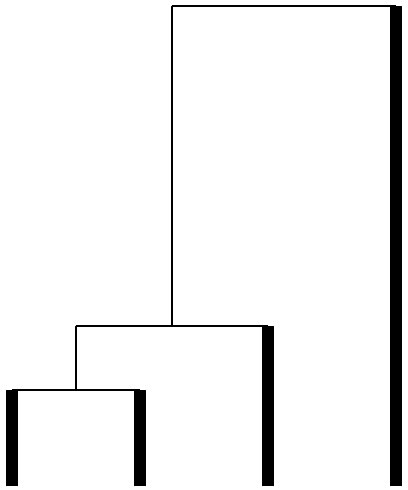
Difference normalised by s.d. like Z statistic

- Negative values of *D* indicate an excess of rare mutations, positive values indicate an excess of intermediate frequency mutations



Fu and Li D test (1993)

- Fu and Li (1993) *D* test

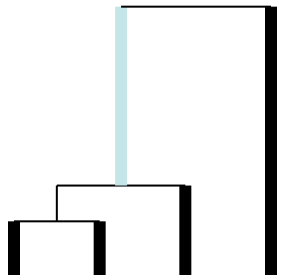


Expected number of external mutations

$$E[\eta_e] = \theta$$

$$D = \frac{S - a_n \eta_e}{\sqrt{\mathrm{Var}(S - a_n \eta_e)}}$$

Without an outgroup, the test has to be adjusted (*D**)



A general class of tests: Fu (1995)

- The expected number of mutations with a derived frequency of *i* in the sample is

$$E[\xi_i] = \frac{\theta}{i}$$

- There is a potentially endless supply of possible tests based around this result BUT
  - Much shared information
  - Large variance

- Fay and Wu (2000) suggest *H* test is powerful for detecting selective sweeps

$$H = \frac{\pi - \theta_H}{\sqrt{Var(\pi - \theta_H)}} \qquad \theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

How do we spot an unusual observation?

- We need to ask whether the statistic we have observed is unusual given the standard neutral model
- We can either do this by looking up values in a table, or simulating data (preferable)
- The basic idea is to simulate data under the null model lots (100s) of times and ask how striking an outlier the empirical observation is

More on testing the model

- There are many ways in which we can look for departures from the assumed model (see also HKA test, McDonald-Kreitman test)

- Different evolutionary forces that we have left out of the basic model will tend to have different effects on the data, so will be picked up by different ways of testing the null model

- Watterson (1977)

> "There is no single statistic which is best for testing the most general departures from neutrality"

Summary

- Genetic data should be summarised in a manner that is informative about the underlying model parameters
- There are a variety of estimators of the population mutation parameter, theta, most of which only use a small amount of information
- Summary statistics can be used to test whether the coalescent provides an adequate description of the data
- An important class of summary statistics are neutrality test statistics that compare different estimators of theta

Plan for Module 21

Wednesday 6/23	1:30-3:00	Introduction	Philip
	3:30-4:00	Introduction (continued)	Philip
	4:00-5:00	Introduction	Mary
Thursday 6/24	8:30-10:00	Recombination	Philip
	10:30-12:00	Recombination practical	Philip
	1:30-3:00	Population size and structure	Mary
	3:30-5:00	Gene flow practical	Mary
	5:00-6:00	Tutorial	Mary/Philip
Friday 6/25	8:30-10:00	Selection	Philip
	10:30-12:00	Selection practical	Philip
	1:30-3:00	Applications and study design	Mary
	3:30-5:00	Coalescent practical	Mary

Details–Wednesday 6/23

- Wednesday afternoon: Introduction to the Coalescent
  - population genetics, Wright-Fisher model
  - 2-sample coalescent
  - n-sample coalescent
  - Coalescent and sequence variation

Outline

1. What types of questions can the coalescent answer?

2. What approaches are used?

3. Case studies



What is the coalescent good for?

- We are interested in questions like
  - How big is this population?
  - When did these populations diverge?
  - Are they isolated? How common is migration?
  - How fast have they been growing or shrinking?
  - What is the recombination rate across this region?
  - Is this locus under selection? What kind?

Coalescent versus traditional population genetics

- Traditional pop gen:
  - Trace the evolutionary process $forward$ in time
  - Predict range of outcomes for a giving starting position
- Coalescent analysis:
  - Trace the evolutionary process $backward$ in time
  - Predict range of scenarios leading to given final position
- Since we know final position more often than starting position, the coalescent is useful for many questions where traditional population genetics struggles

Coalescent versus traditional population genetics

- Traditional pop gen: A neutral allele is now at 5% frequency
  - How likely is it to fix?
  - How long will that take?
  - What if it were under selection?
- Coalescent: Ten out of thirty haplotypes surveyed carry a particular variant
  - How old is the variant?
  - Is it under selection?
  - Has it been transferred among populations?

Range of applicability

The coalescent is appropriate for:

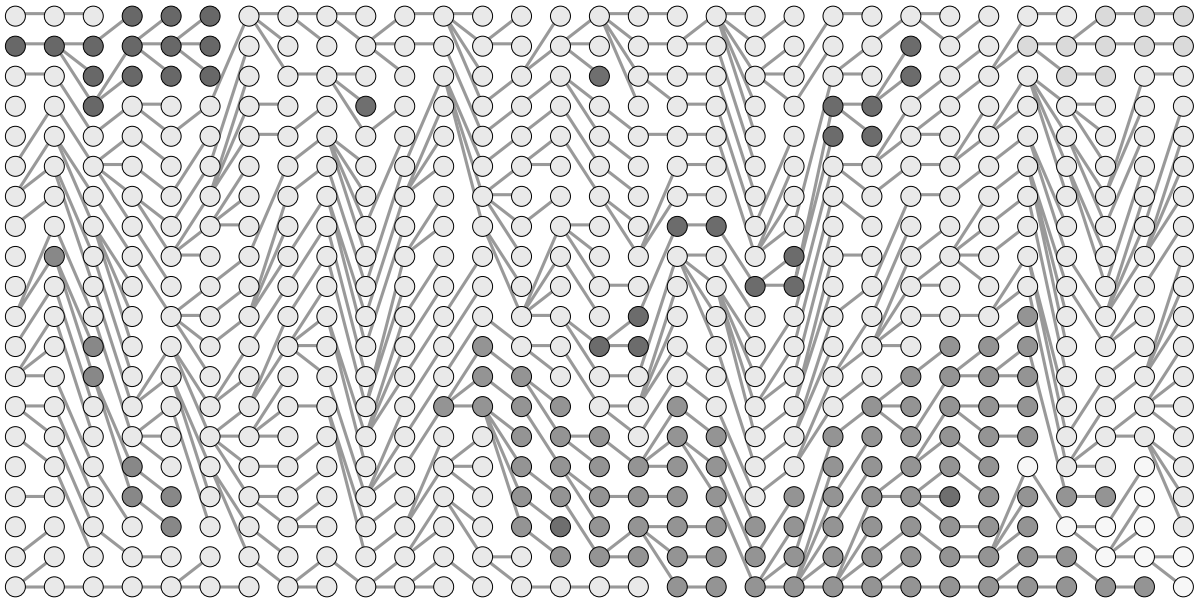
- Single populations
- Interrelated populations
- Recently diverged species

Beyond this level, other processes come into play

Key concepts in the coalescent

- In an idealized coalescent everything depends on population size
- Deviations from idealized population give us information on other parameters:
  - Difference between census size and effective size
  - Changes in size over time
  - Population subdivision
  - Population splitting (divergence)
  - Recombination
  - Natural selection

Basics: Wright-Fisher population model



All individuals release many gametes and new individuals for the next generation are formed randomly from these.

9

Wright-Fisher population model

- Population size N is constant through time.
- Each individual gets replaced every generation.
- Next generation is drawn randomly from a large gamete pool.
- Only genetic drift is changing the allele frequencies.

Other population models

- Other population models can often be equated to Wright-Fisher
- The N parameter becomes the effective population size N_e
- N_e of a cyclic population is the harmonic mean of the various sizes
- Social insect N_e is based on queens and drones only

The Θ parameter

- The n-coalescent is defined in terms of N_e and time.
- We cannot measure time just by looking at genes, though we can measure divergence.
- We rescale the equations in terms of N_e, time, and the mutation rate μ.
- We can no longer estimate N but only the composite parameter Θ.
- $\Theta = 4N_e\mu$ in diploids and $2N_e\mu$ in haploids.
- External information can allow us to separate Θ and μ.

How to separate N_e and μ?

Given an estimator of $\Theta = 4N_e\mu$:

- If one is known, the other naturally follows
- N_e is hard to estimate
- μ can be estimated from dated fossils
- Multiple observations with significant evolution between them allow separation of N_e and μ
- These could come from ancient DNA
- In fast-evolving species like viruses they can be directly observed

Utopian coalescent population size estimator

1. We get the correct genealogy from an infallible oracle
2. We know that we can calculate $\mathrm{p}(\text{Genealogy}|\mathrm{N_e})$

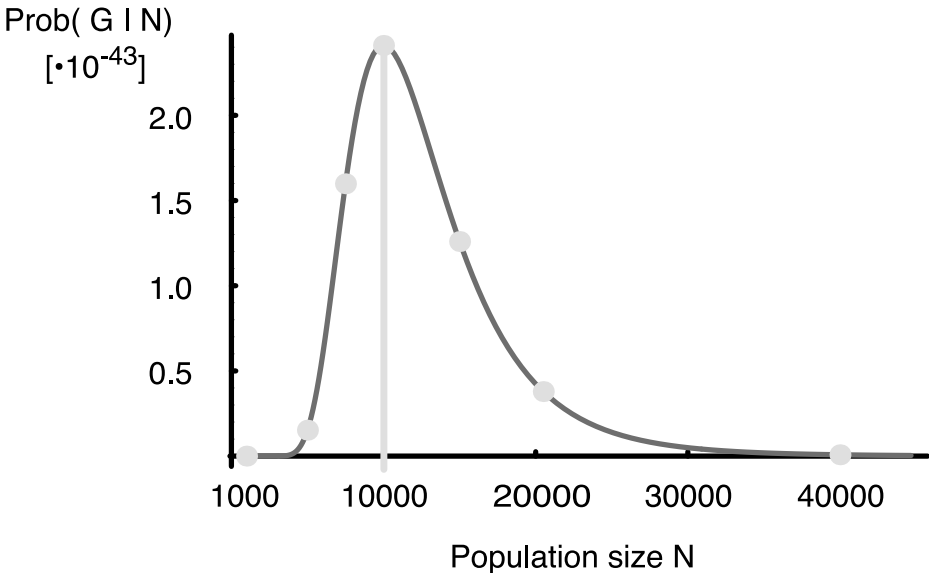
Utopian population size estimator

1. We get the correct genealogy from an infallible oracle

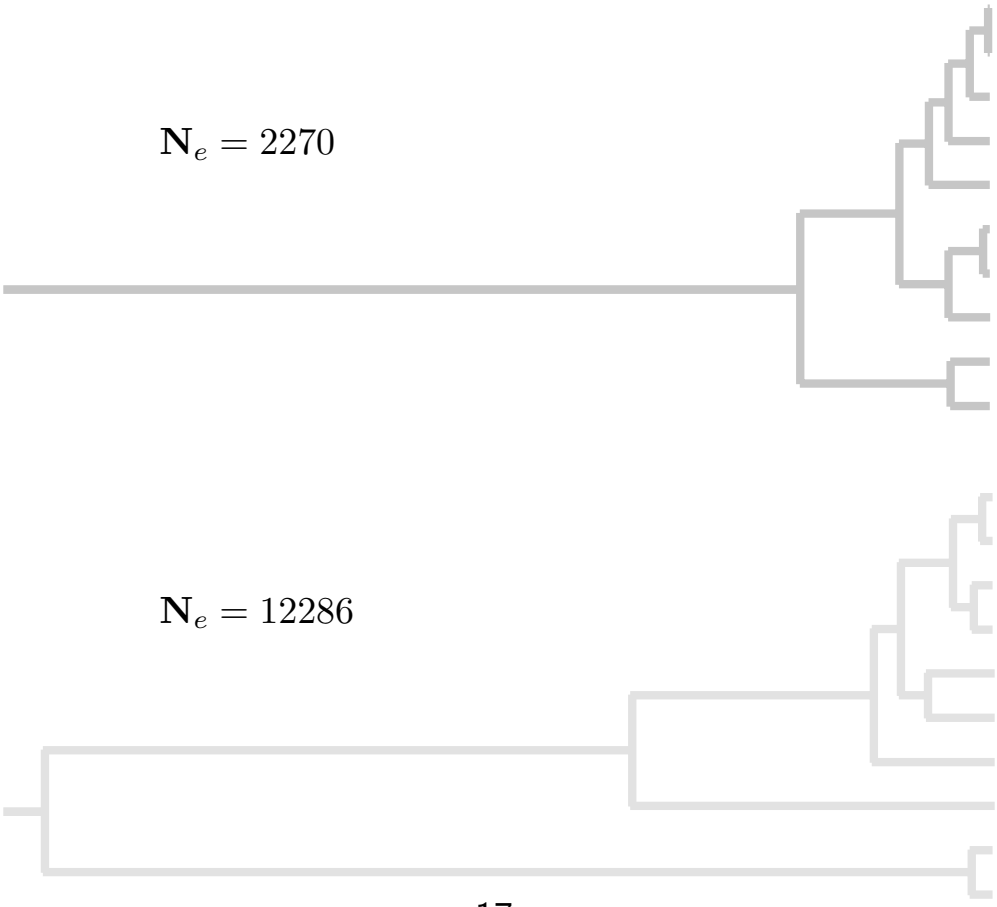
2. We remember the probability calculation

$$p(\text{Genealogy}|N_e) = \prod_{j}^{T} e^{-u_j \frac{k_j(k_j-1)}{4N_e}} \frac{1}{2N_e}$$

Utopian population size estimator



Utopian population size estimator



17

Lack of infallible oracles

- We assume we know the true genealogy including branch lengths
- We don't really know that
- We probably can't even infer it:
  - Tree inference is hard in general
  - Population data usually doesn't have enough information for good tree inference

Ways to use the coalescent

- Summary statistics
  - Watterson's estimator of θ
  - FST (estimates θ and/or migration rate)
  - Hudson's and Wakeley's estimators of recombination rate
- Known-tree methods
  - UPBLUE (Fu)
  - Skyline plots
- Few-tree methods
  - Nested clade analysis (Templeton)

These methods are conceptually easy, but not always powerful, and they can be difficult to extend to complex cases.

More ways to use the coalescent

- Approximate Bayesian Computation (ABC)
  - Simulate random coalescent genealogies with a particular set of parameters
  - Simulate data on those genealogies
  - Adjust parameters until results are "similar" to real data
  - Summary statistics define "similar"
- Full genealogy samplers
  - Find many genealogies which fit the data well
  - Make a joint estimate of the parameters from all of these genealogies

These methods are more powerful and flexible, but challenging to design and use

Use of the coalescent: three case studies

Things to look for:

- What questions were being addressed?
- What types of data were used?
- How were coalescent methods used?
- How were non-coalescent methods used?
- How were the results validated?

What is the effective population size of red drum?

Red drum, *Sciaenops ocellatus*, are large fish found in the Gulf of Mexico.

Turner, Wares, and Gold
Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish
Genetics 162:1329-1339 (2002)

What is the effective population size of red drum?

- Census population size (N): 3,400,000
- Effective population size (N_e): ?
- Data set:
  - 8 microsatellite loci
  - 7 populations
  - 20 individuals per population

What is the effective population size of red drum?

Three approaches:

1. Allele frequency fluctuation from year to year
   - Measures current population size
   - May be sensitive to short-term fluctuations
2. Coalescent estimate from $Migrate$
   - Measures long-term harmonic mean of population size
   - May reflect past bottlenecks or other long-term effects
3. Demographic models
   - Attempt to infer genetic size from census size
   - Vulnerable to errors in demographic model
   - Not well established for long-lived species with high reproductive variability

What is the genetic population size of red drum?

Assumptions of the coalescent analysis:

- Constant population size
- Mutation rate 10^{-3} to 10^{-5}
- No selection

What is the effective population size of red drum?

Estimates:

Census size (N):	3,400,000
Allele frequency method (N_e):	3,516 (1,785-18,148)
Coalescent method (N_e):	1,853 (317-7,226)

The demographic model can be made consistent with these only by assuming enormous variance in reproductive success among individuals.

What is the effective population size of red drum?

- Allele frequency estimators measure current size
- Coalescent estimators measure long-term size
- Conclusion: population size and structure have been stable

What is the effective population size of red drum?

- Effective population size at least 1000 times smaller than census
- This result was highly surprising
- Red drum has the genetic liabilities of a rare species
- "Estuary lottery" may explain results

Where to go with this finding?

- Check it experimentally–maternity testing of young fish?
- Try to find reasons for the high reproductive variance
- Be careful of this species as it may be fragile
  - Red drum were once commercially fished
  - The population responded poorly and the fishery was closed
  - Despite large numbers the species may be vulnerable
  - Are there other species like this?

What was the long-term population size of gray whales?

Alter, Rynes and Palumbi (2007) DNA evidence for historic population size and past ecosystem impacts of gray whales. PNAS 104: 15162-15167.

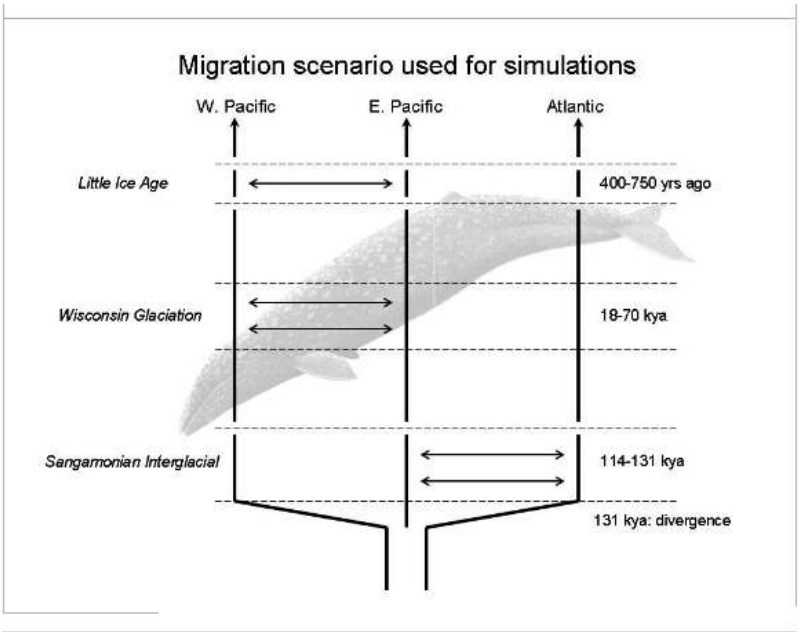
What was the long-term population size of gray whales?

- How many gray whales pre-whaling?
- Whaling ship records not conclusive
- Recent slowing of the observed growth rate may suggest recovery
- Molecular data an alternative source of information

What was the long-term population size of gray whales?

- 10 loci:
  - 7 autosomal
  - 2 X-linked
  - 1 mtDNA
- Complex mutational model with rate variation among loci
- Complex population model with subdivision and copy number
- Complex demographic model relating N_{census} to N_e

What was the long-term population size of gray whales?



What was the long-term population size of gray whales?

	Locus	n	Estimated N
Aut	ACTA	72	162,625
	BTN	72	76,369
	CP	76	77,319
	ESO	72	272,320
	FGG	72	180,730
	LACTAL	72	44,410
	WT1	80	51,972
X	G6PD	30	2,769
	PLP	52	92,655
mtDNA	Cytb	42	107,778
	All data		96,400 (78,500-117,700)
	Current census		18,000-29,000
	Previous models		19,480-35,430

What does this imply?

- Important conservation implications
- Effect on ecosystem significant:
  - Resuspension of up to 700 million cubic meters sediment
  - (12 Yukon Rivers worth)
  - Food for 1 million sea birds
- If accepted, result suggests halving gray whale kill rate
- Broadly similar results for minke, humpback, and fin whales

Should we believe this result?

- Strengths:
  - Multiple loci improve power and avoid distortions
  - Population structure taken into account
  - Does not rely on whalers' records, which may be falsified
- Weaknesses:
  - Interpretation relies on external estimate of mutation rate
  - Selection on coding loci could distort results
  - Relies on model of relationship between N and N_e

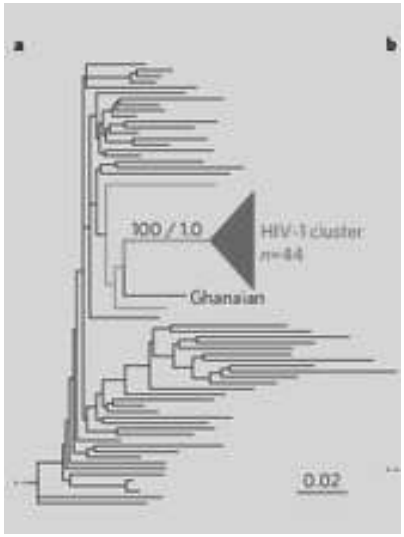
Where to go with this finding?

- Non-coding sequences
- Whale lice as corroboration? (Jon Seger's work)
- Ancient DNA?
- More sophisticated demographic models?
- Not time to de-list gray whales yet

When did the El-Fatih HIV epidemic start?

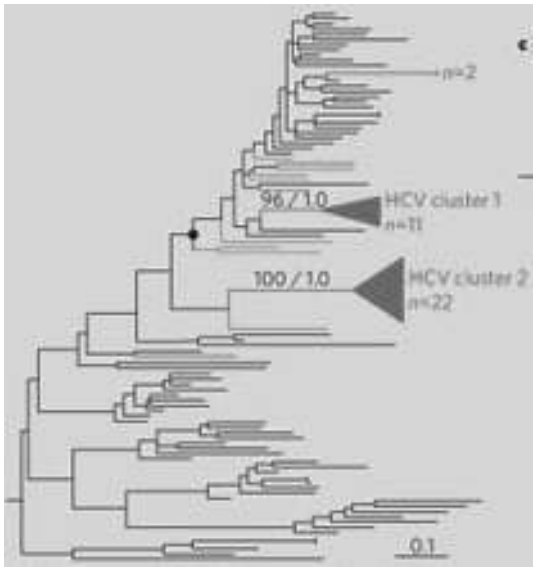
- At El-Fatih Hospital in Libya over 400 children found infected with HIV and HCV
- Libya accused foriegn medics who entered the country March 1998
- When did the epidemic start according to genetic data?

Observed relationships: HIV-1



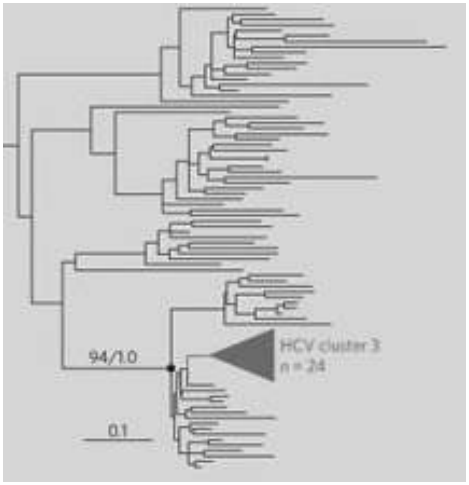
Red = hospital, blue = Cameroon
from de Oliveira et al, Nature 2006; used with permission

Observed relationships: HCV type 1



Red = hospital, green = Egypt, blue = Cameroon
from de Oliveira et al, Nature 2006; used with permission
This strain of HCV is epidemic worldwide

Observed relationships: HCV type 4



Red = hospital, green = Egypt, blue = Cameroon
from de Oliveira et al, Nature 2006; used with permission
This strain of HCV is epidemic in Egypt due to contamination of anti-worm medication in the 1970's

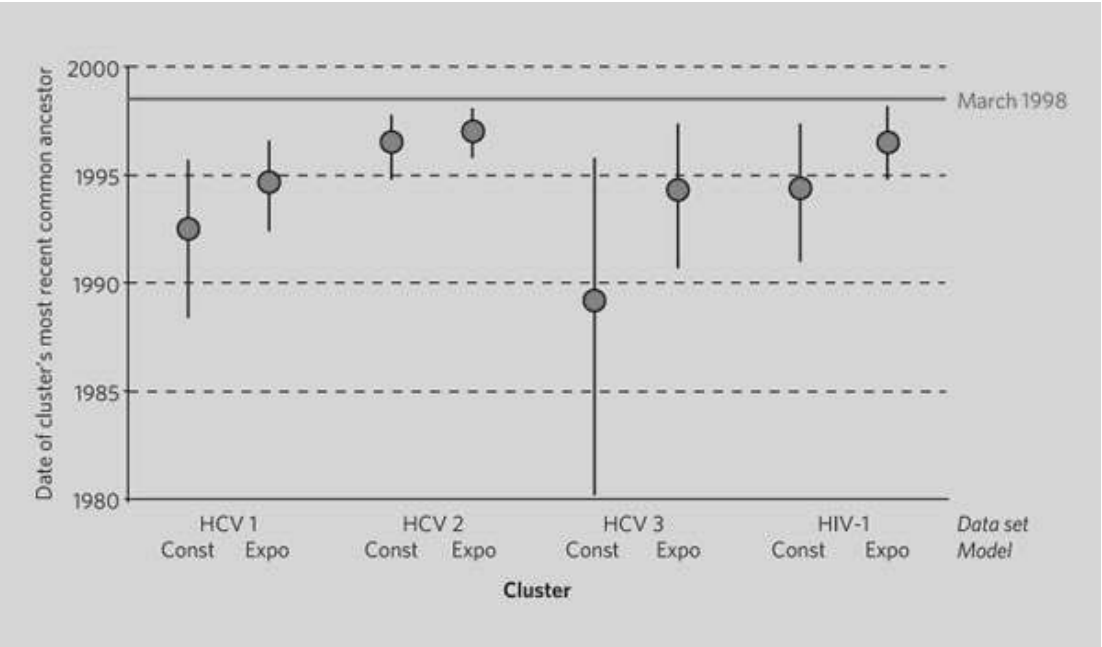
Where to go from here?

- HCV-1 looks like a community infection
- HIV and HCV-4 look like single-source; could it be the medics?
- How can we date these events?

Inferring ancestral times

- de Oliveira et al. (2006) did coalescent analysis using BEAST
- They had HIV samples from multiple time points to establish μ
- They considered huge numbers of possible genealogies:
  - Using only the best genealogy is biased
  - Considering many genealogies allows error bars

The viruses arose before March 1998



from de Oliveira et al, Nature 2006; used with permission

Limitations of this study

- Authors considered constant and exponentially growing populations
- Other growth patterns might change the answer
- This study did not consider:
  - Genetic recombination among viruses
  - Possible genetic engineering (as alleged by Libya)
  - Population structure
  - Natural selection

Validation

- A Libyan official eventually admitted that records showed infected children prior to March 1998
- A particular child who was admitted dozens of times during 1997-98 was probably Patient 0
- Due to civil disorder additional data unlikely
- Even in developed countries HIV outbreaks often cannot be traced successfully to their origins

Focus: genealogy samplers

- My practicals will focus on genealogy samplers
- Most statistically powerful way to extract information from coalescent genealogies
- Challenging to design and use
- Brief overview now, practical details later

Parameter estimation by genealogy sampling

- Mutation model: Steal a likelihood model from phylogeny inference
- Population genetics model: the Coalescent

Parameter estimation by genealogy sampling

$$L(\Theta) = P(Data|\Theta)$$

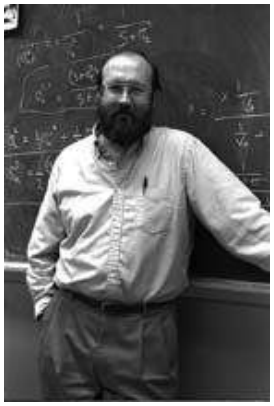
Parameter estimation by genealogy sampling

$$L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

Parameter estimation by genealogy sampling

$$L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

$P(Data|G)$ comes from a mutational model



Parameter estimation by genealogy sampling

$$L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

$P(G|\Theta)$ comes from the coalescent

Parameter estimation by genealogy sampling

$$L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

$\sum_G$ is a problem

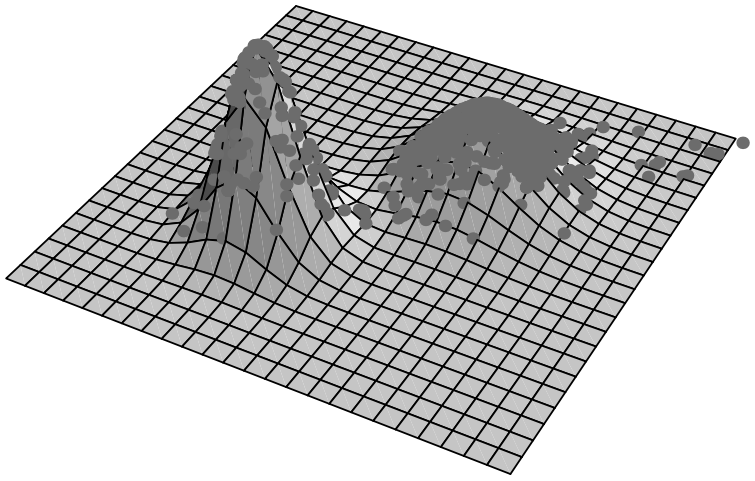
Can we calculate this sum over all genealogies?

Tips	Topologies
3	3
4	18
5	180
6	2700
7	56700
8	1587600
9	57153600
10	2571912000
15	6958057668962400000
20	564480989588730591336960000000
30	436846661310306951246468019862076389144064000000000000
40	3027333829948007356546303364551457200042939432053862501707888721920000000000000000(
50	3.28632×10^{112}
100	1.37416×10^{284}

A solution: Markov chain Monte Carlo

- If we can't sample all genealogies, could we try a random sample?
  - Not really.
- How about a sample which focuses on good ones?
  - What is a good genealogy?
  - How can we find them in such a big search space?

A solution: Markov chain Monte Carlo



A solution: Markov chain Monte Carlo

Metropolis recipe

0. first state

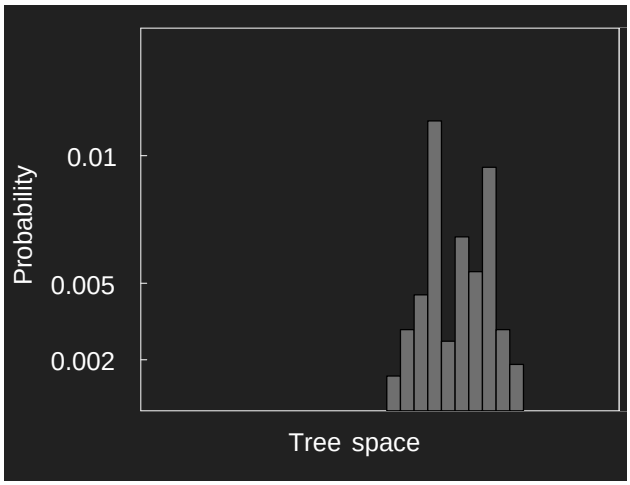
1. perturb old state and calculate probability of new state

2. test if new state is better than old state: accept if ratio of new and old is larger than a random number between 0 and 1.

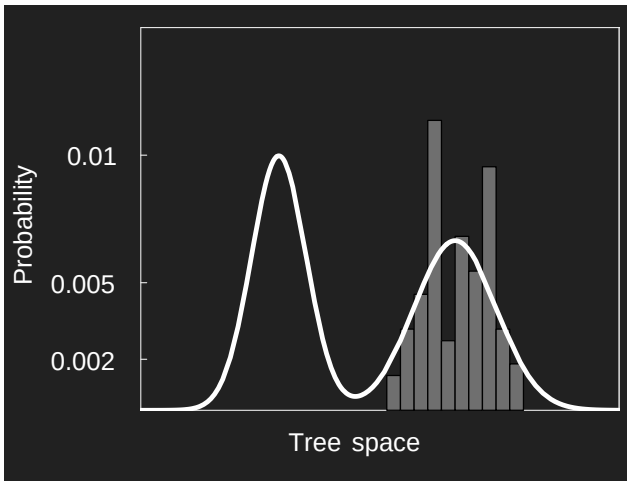
3. move to new state if accepted otherwise stay at old state

4. go to 1

MCMC walk result



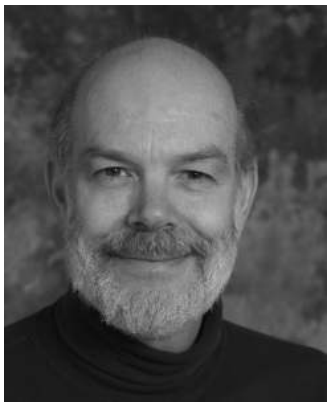
MCMC walk result



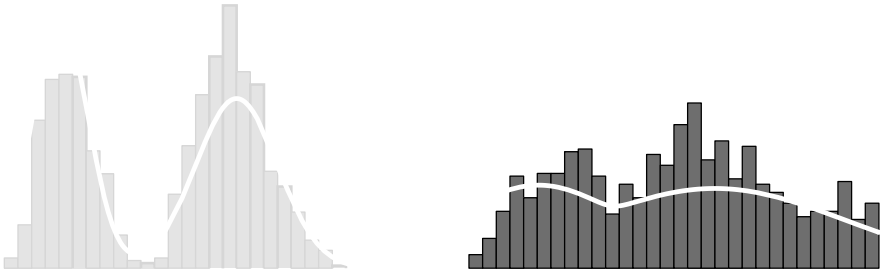
Improving our MCMC walker: MCMCMC or MC^3

Metropolis Coupled Markov chain Monte Carlo

- The "hot" searches see a flattened landscape
- Only the "cold" search is used to make the estimate
- Good solutions found by a "hot" search can be imported into the "cold" search

better MCMC walk result



Paul Lewis' MCMCRobot

This program can be found at

http://www.mcmcrobot.org

It carries out a Markov Chain Monte Carlo search on a simple surface.

MCMCRobot Experiment

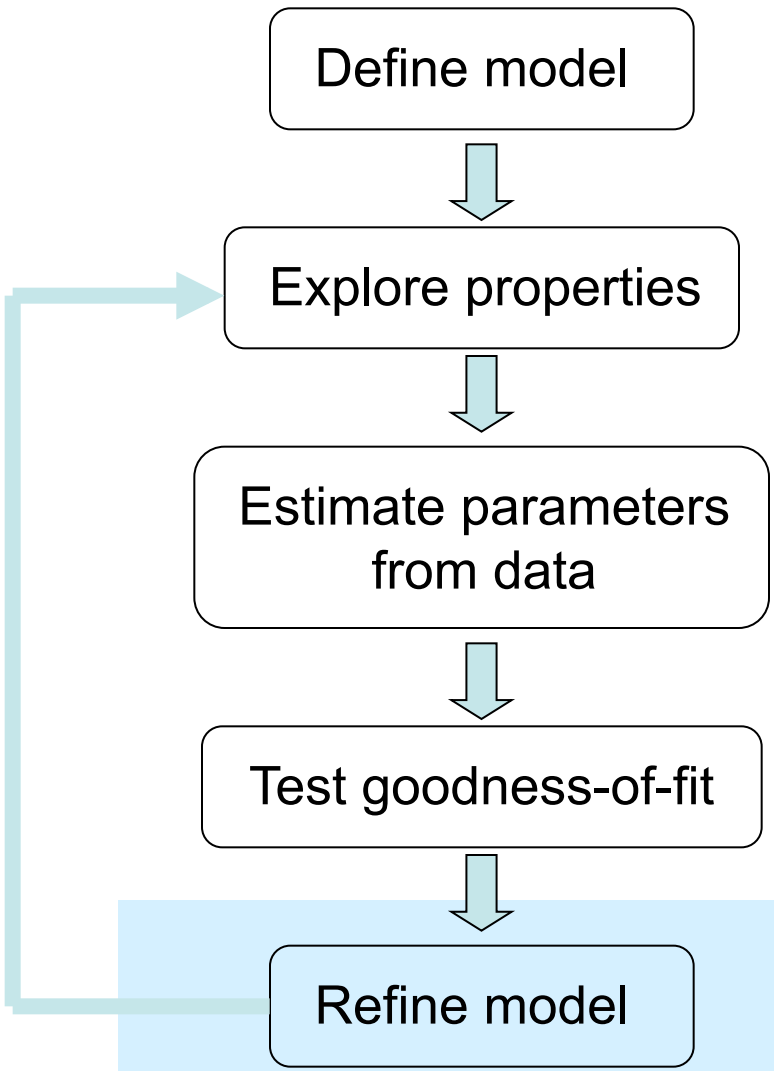
- How well does the robot search:
  - A single hill?
  - Two hills close together?
  - Two hills far apart?
- Does heating help with the far-apart case?
- Try 2, 3 and 4 chains: which seems best?

Preparation for Thursday sessions

- Data files can be downloaded from:
  - http://evolution.gs.washington.edu/lamarc/sisg-2014/demo/
- Please download these before the demonstration Thursday
- This will save time and pain during the demo. Thank you!

Extending the coalescent

Statistical inference



Issues

rules
parameters
simulation

summary statistics
graphical representation

moment methods
likelihood

outliers
heterogeneity
comparison of estimators

add parameters

What does the basic coalescent miss?

- The basic coalescent model is excellent at capturing the great stochasticity underlying genealogical processes in natural populations

- It does not, however, capture many biologically important features that we would like to learn about
  - Recombination
  - Changes in population size
  - Geographical structure
  - Natural selection

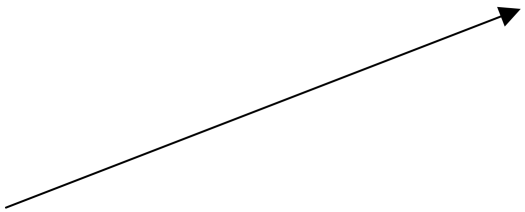
- We would like to extend the basic model to incorporate these features so that we can
  - Look for their influence (model testing)
  - Estimate their importance (parameter estimation)

Adding recombination

- Most genes in most genomes undergo recombination
  - Reciprocal cross-over
  - Gene conversion
  - Transformation
  - Template-switching

- The effect of recombination on genealogies is described by the **ancestral recombination graph (ARG)**

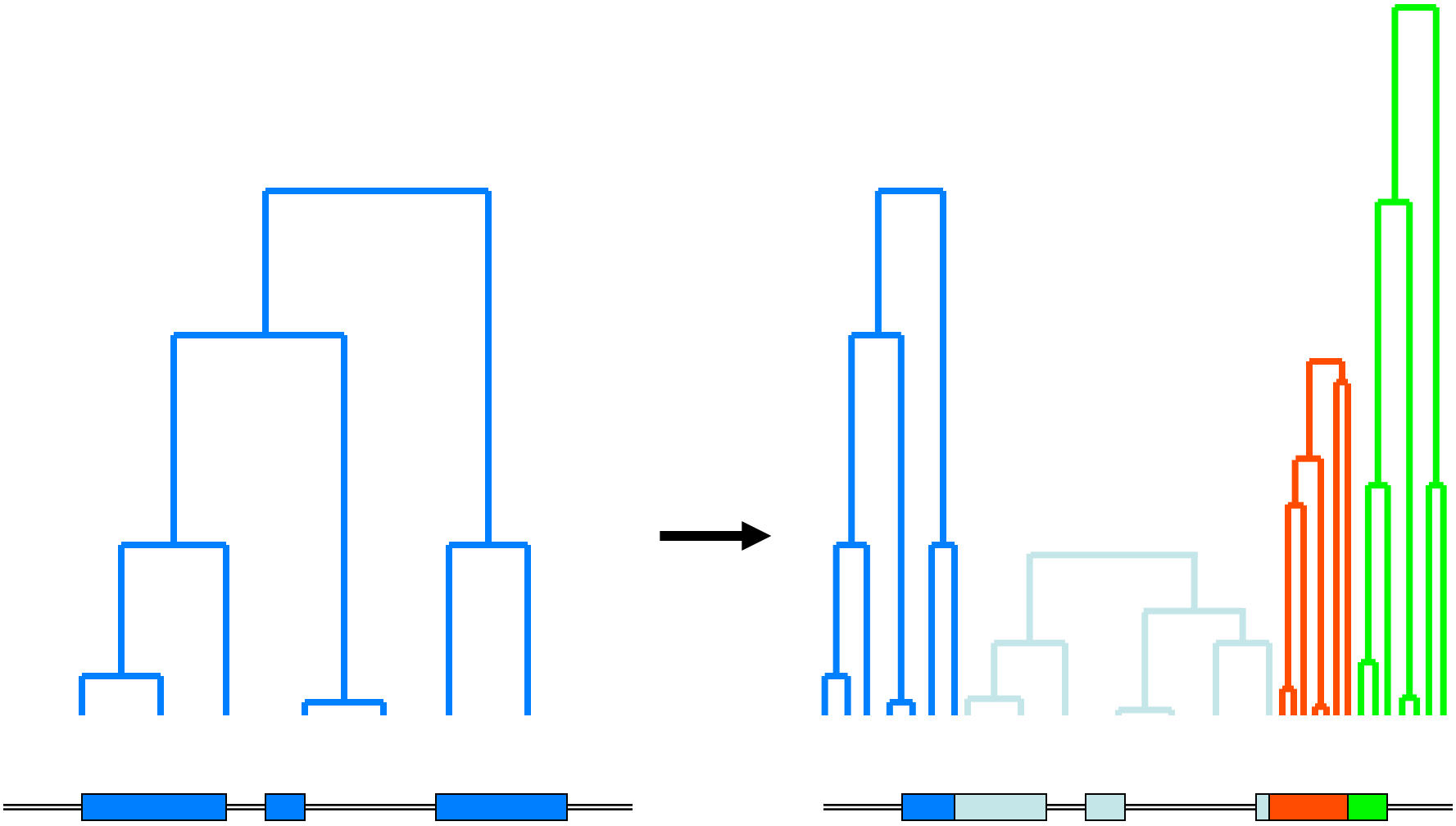
- In this lecture we will
  - Look at how to incorporate recombination in the coalescent model
  - Learn about its effects on patterns of genetic diversity
  - Look at how to learn about recombination from empirical data



Recombination is an important evolutionary processes

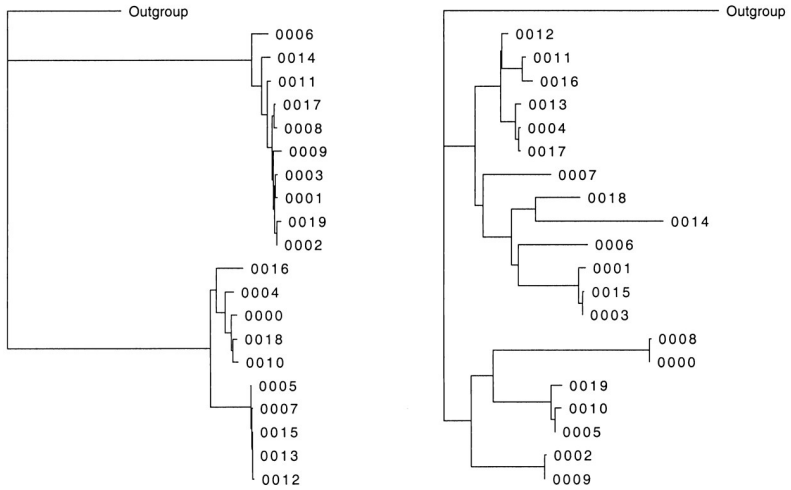
- Recombination brings together multiple beneficial mutations
  - Fisher/Muller
- Recombination brings together multiple deleterious mutations
  - Muller's ratchet, Kondrashov's synergistic deleterious mutations
- Recombination allows beneficial mutations to escape from linked deleterious ones
  - The *ruby in the rubbish*
- Recombination creates novel combinations of mutations
  - Evasion of pathogens and parasites (Hamilton)
  - Reduction in sib competition (Bell)

Recombination influences how we perform inference

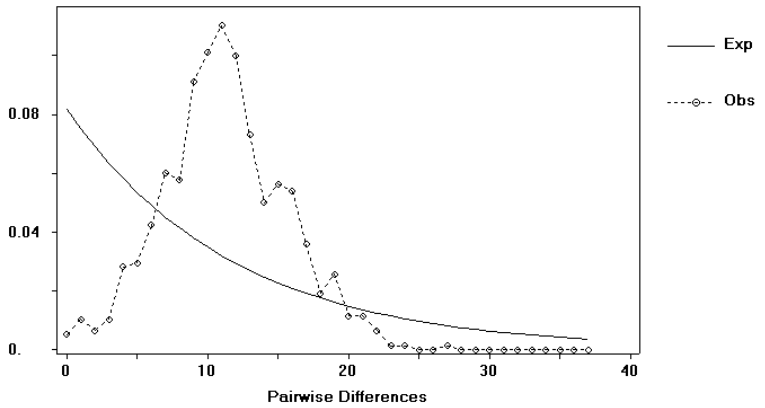


Ignoring recombination leads to problems in inference

- Many inference procedures assume no recombination
  - Presence of recombination can give misleading inferences

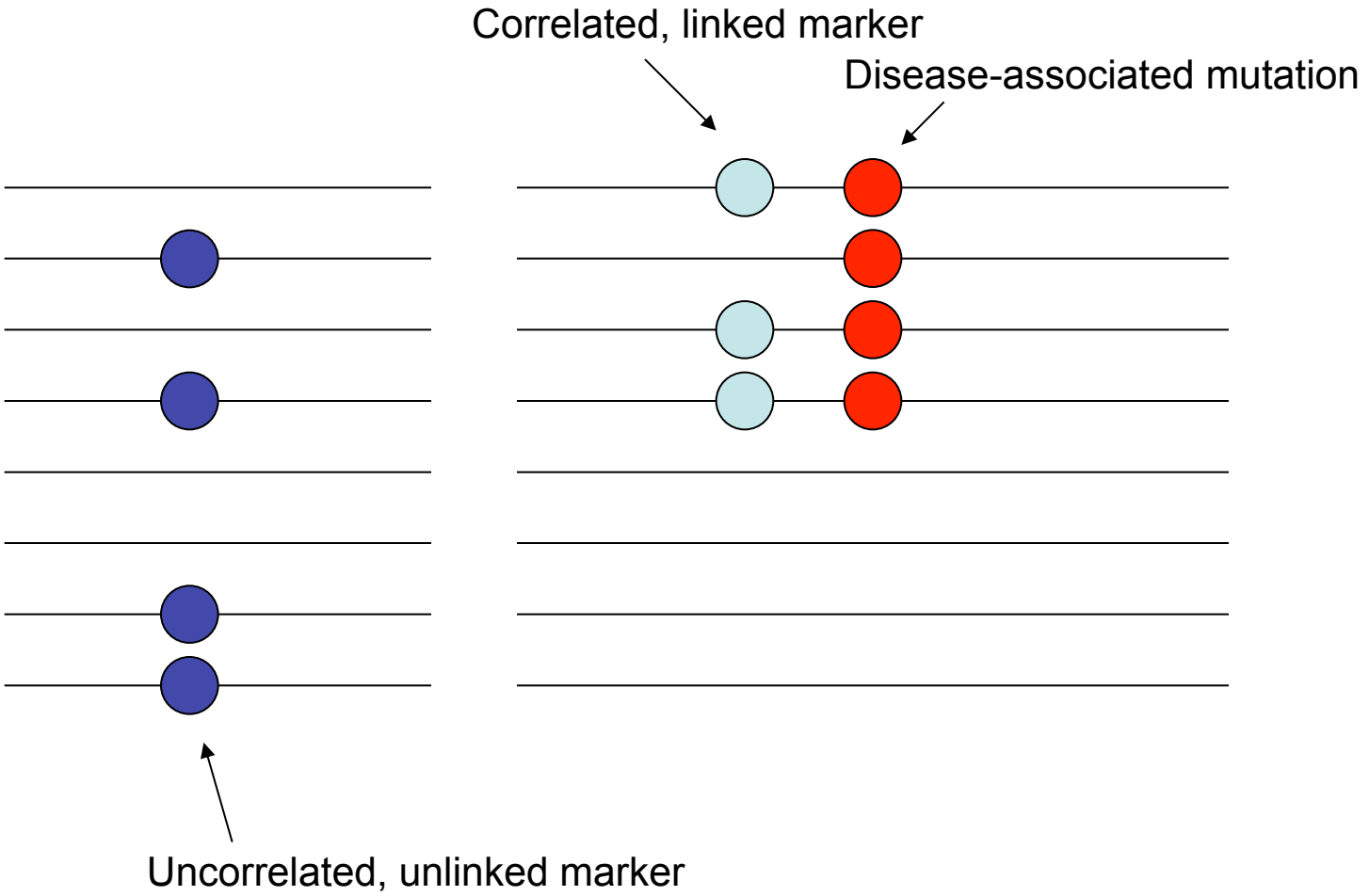


Phylogenetic tree reconstruction



Mismatch distributions

Recombination allows us to map genes by association

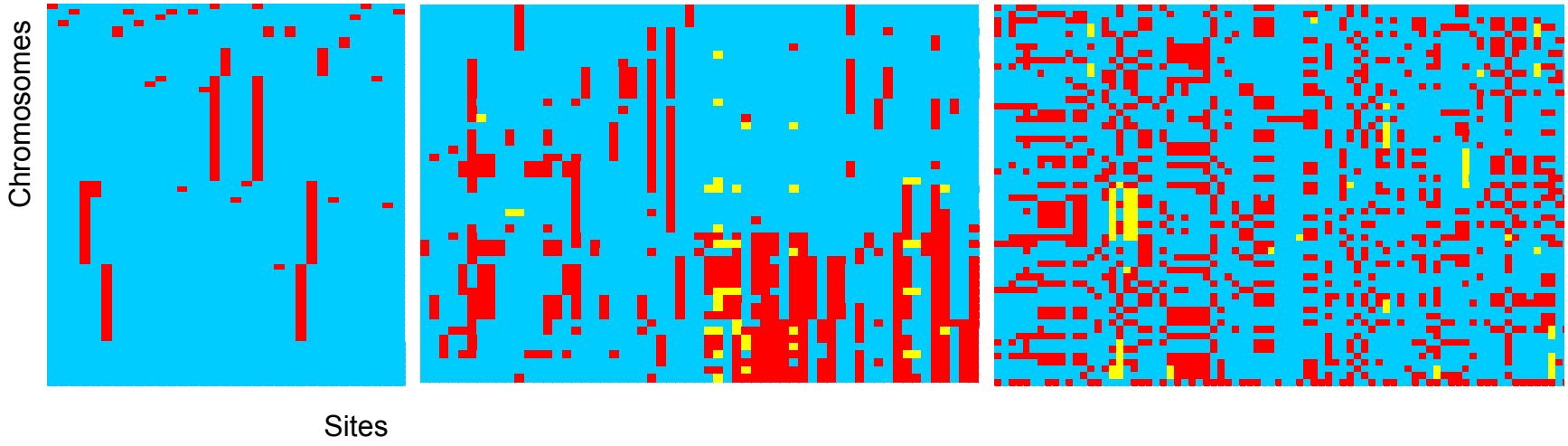


Recombination gives us greater power in inference

- A single non-recombining locus gives us a single ‘draw’ from the underlying genealogical process
- Looking at a single locus will only give us a small part of the population’s history
- Recombination means that different loci have different ‘draws’ from the genealogical process and therefore have independent information about processes we wish to learn about
- Phylogenetic trees of mtDNA have been used to make inferences about human history, but is it representative?
  - Out of Africa hypothesis
  - mtDNA eve about 200,000 years ago
  - Strong population growth

What does recombination do to genetic variation?

- Informally, recombination shuffles up genetic diversity
- We can see the effect of recombination in how ‘structured’ genetic variation is



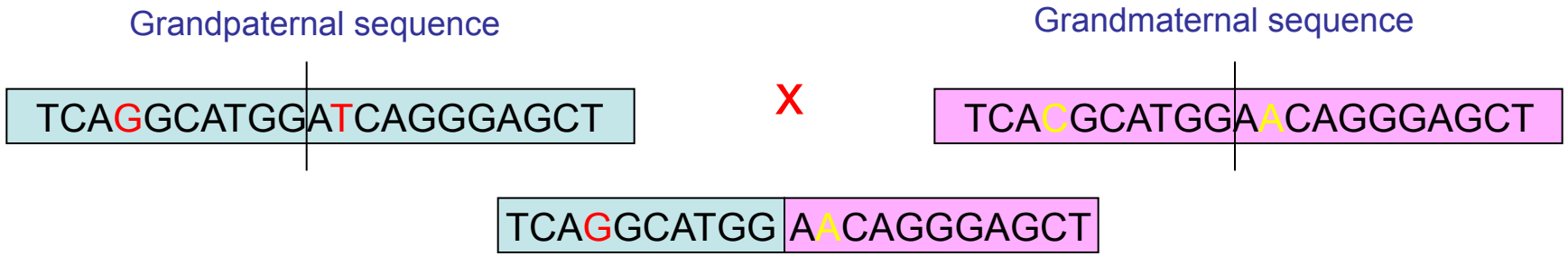
Xq13: 10kb
69 worldwide

Lipoprotein Lipase: 10kb
48 African Americans

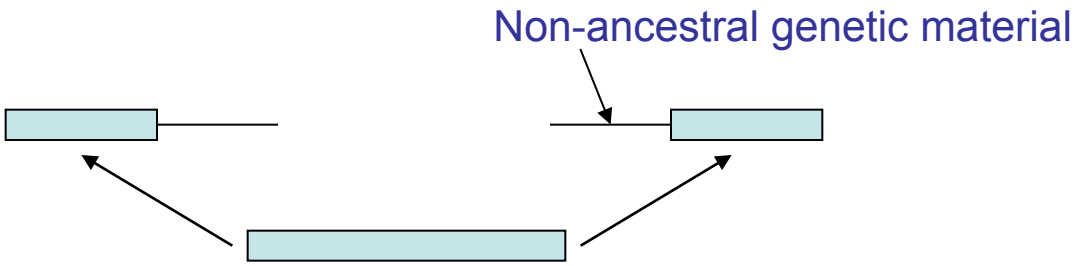
Chromosome 22: 1Mb
57 Europeans

Recombination and genealogical history

- Forwards in time

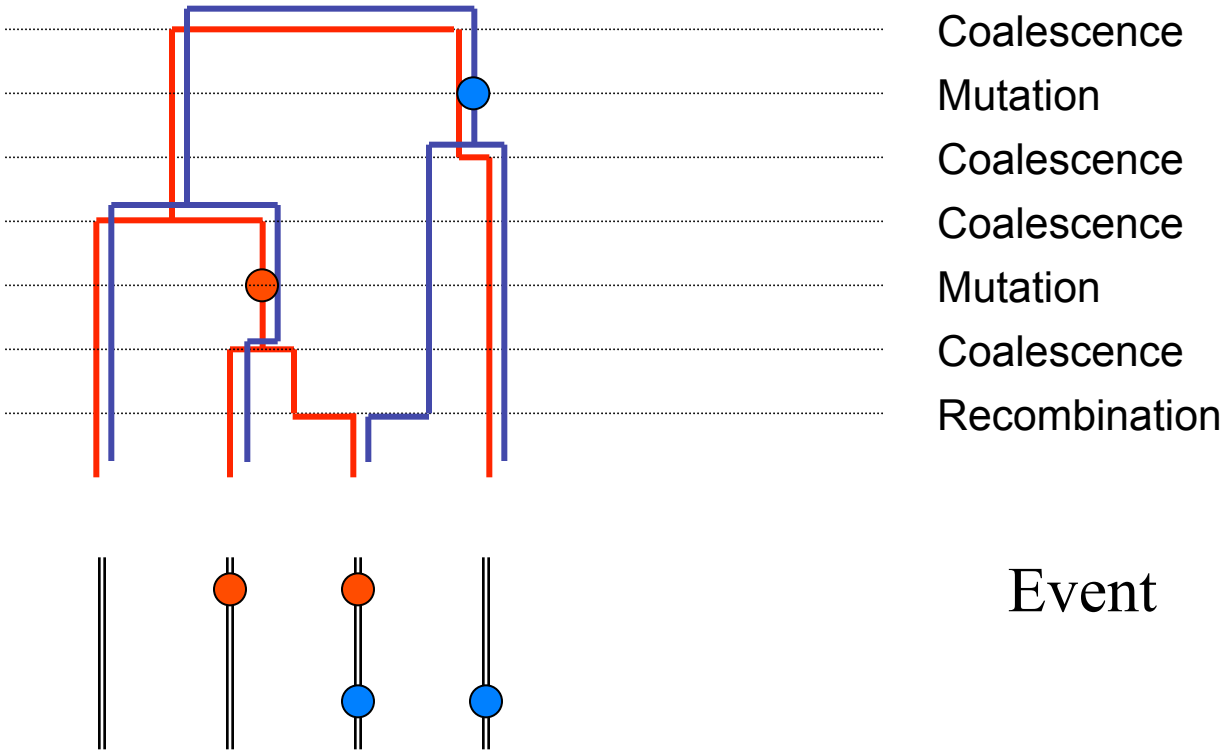


- Backwards in time

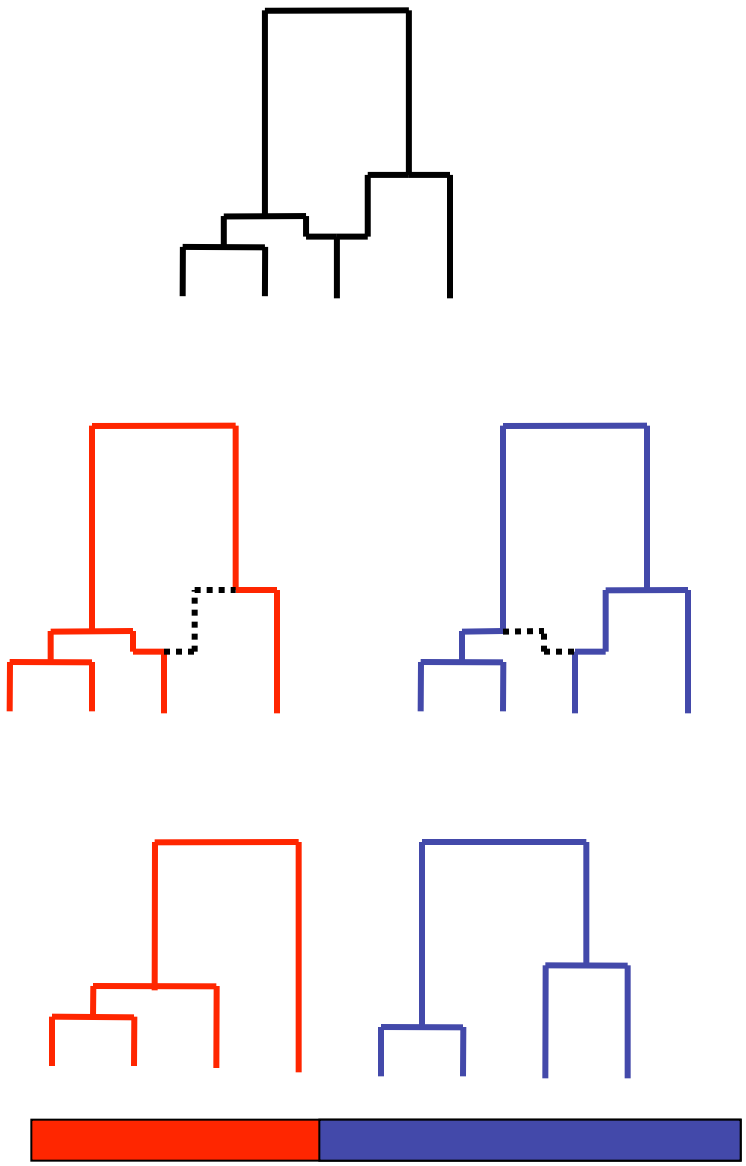


The ancestral recombination graph

- The combined history of recombination and coalescence is described by the ancestral recombination graph



Deconstructing the ARG



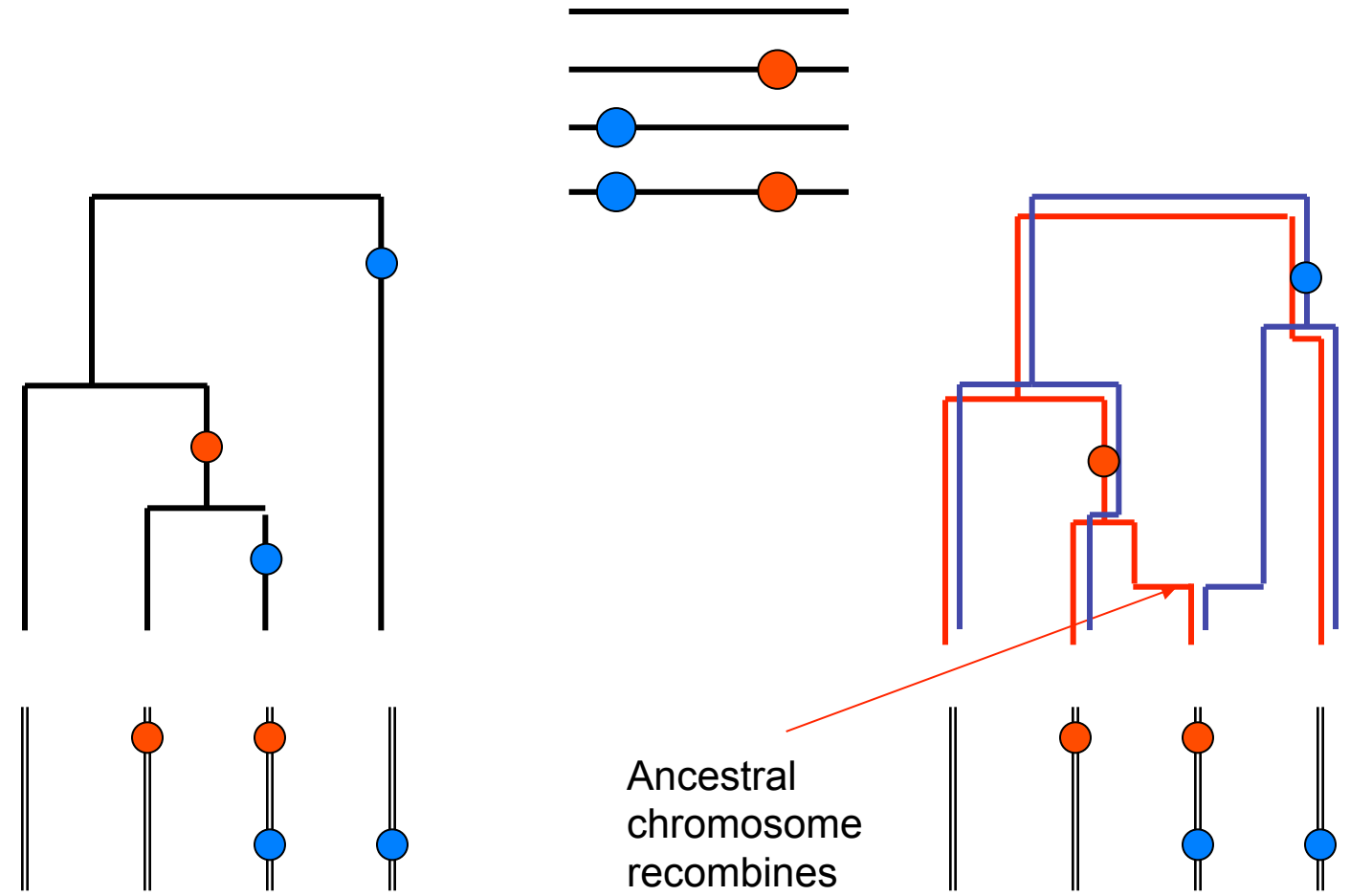
Learning about recombination

- Just like there is a true genealogy underlying a sample of sequences without recombination, there is a true ARG underlying samples of sequences with recombination

- As before, we can consider **nonparametric** and **parametric** ways of learning about recombination

- There are several useful nonparametric ways of learning about recombination which we will consider first
  - These really only apply to species, such as humans, where we can be fairly surely that most SNPs are the result of a single ancestral mutation event

The signal of recombination?

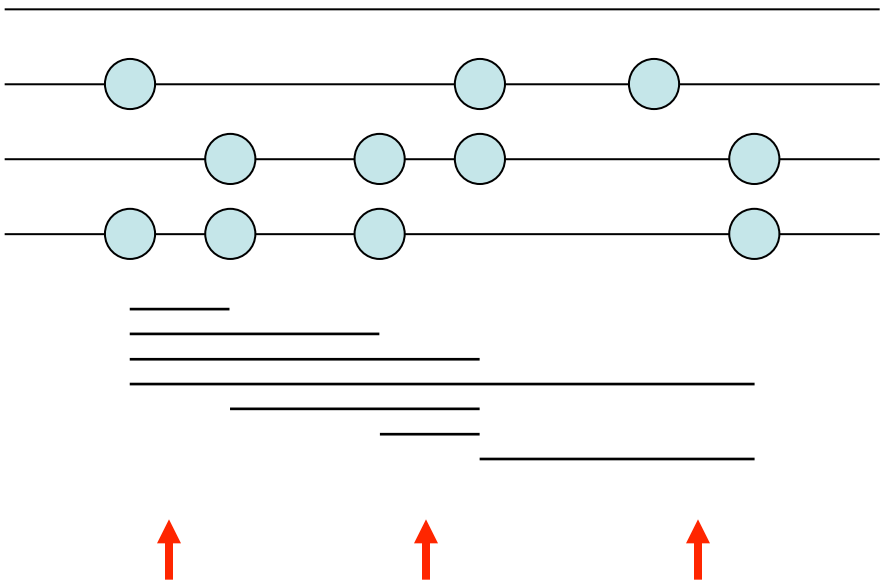


Recurrent mutation

Recombination

Detecting recombination from DNA sequence data

- Look for all pairs of compatible sites



- Find minimum number of intervals in which recombination events must have occured (Hudson and Kaplan 1985): $\boldsymbol{R_m}$

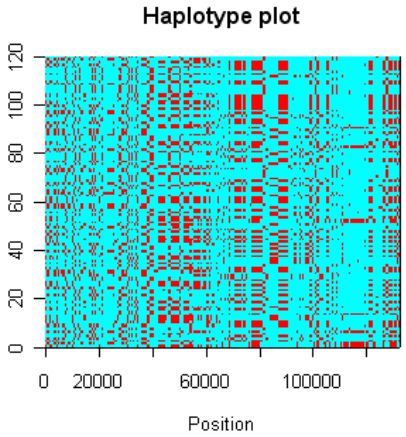
Improving the detection algorithm

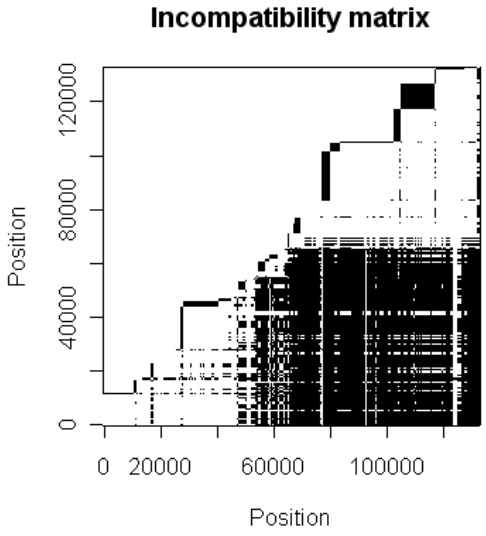
- R_m greatly underestimates the amount of recombination in the history of a set of sequences

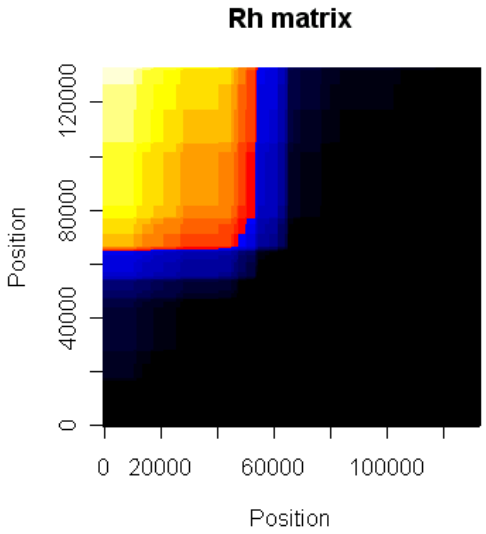
- Myers and Griffiths (2003) developed an improved way of detecting recombination events
  - Without recombination, every new mutation can create only a single new haplotype
  - With recombination, mutations can be shuffled between haplotype background, generating haplotype diversity
  - If I see *H* haplotypes with *S* segregating sites, at least *H*-*S*-1 recombination events must have occurred

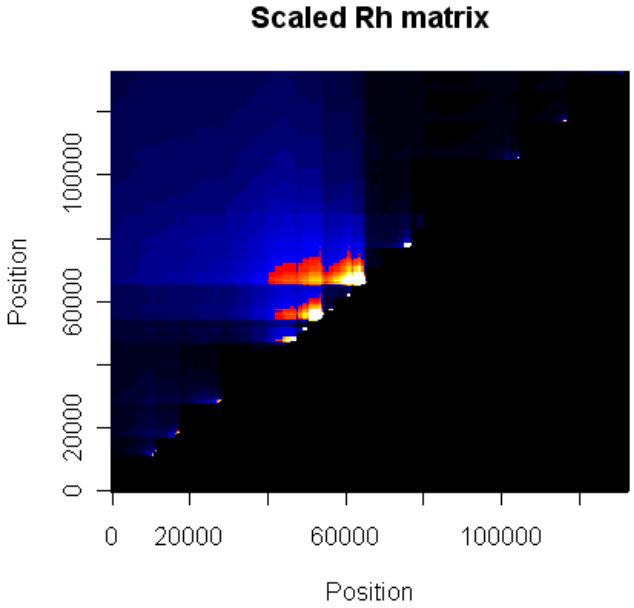
- Local bounds (e.g. just using SNPs 2, 3 and 14) can be combined across a whole sequence to generate an estimate of the minimum number of recombination events and their location

Example: 7q31





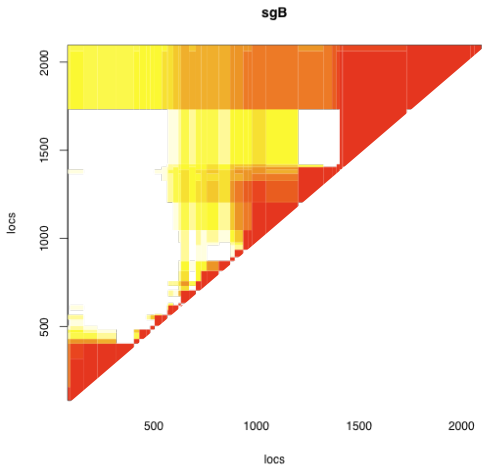




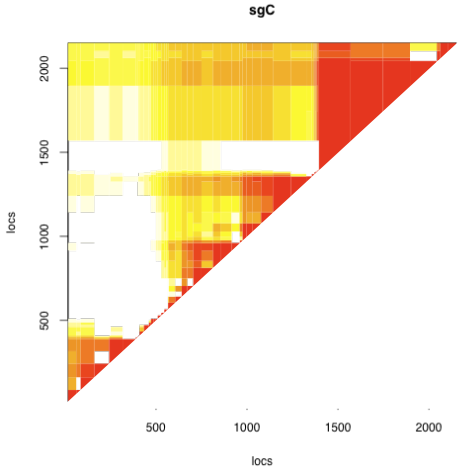
Problems with detecting recombination

- The infinite-sites model is not applicable to all species

HIV Subtype B (2kb segment)



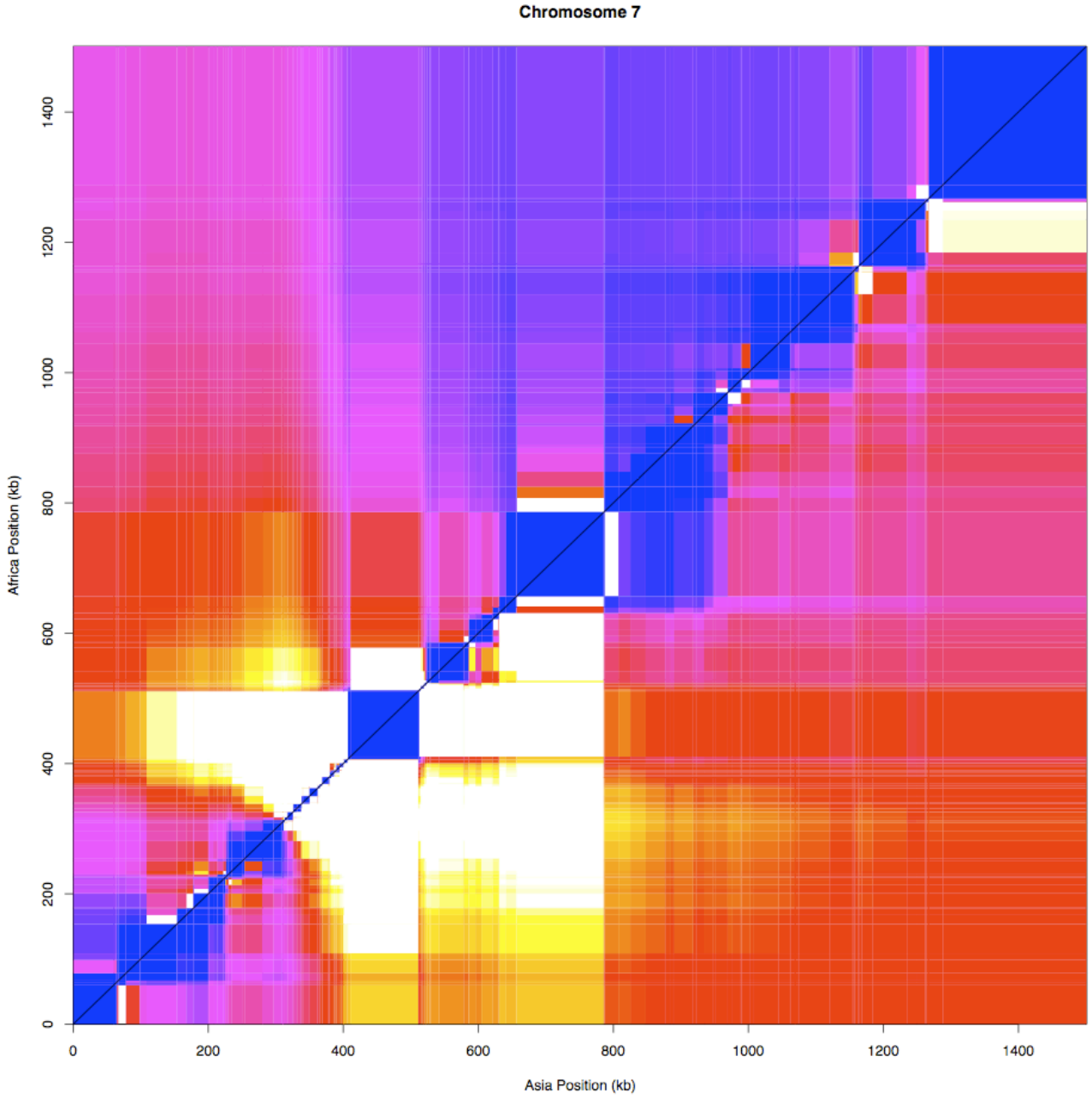
HIV Subtype C (2kb segment)



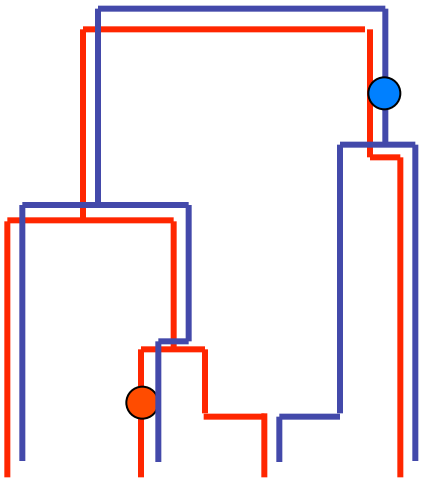
- There are many more recombination events in the history of the sample than the methods can ever detect
  - Lack of mutations in the right places
  - Recombination events completely undetectable

- We care more about recombination rates rather than events (rates are predictive)

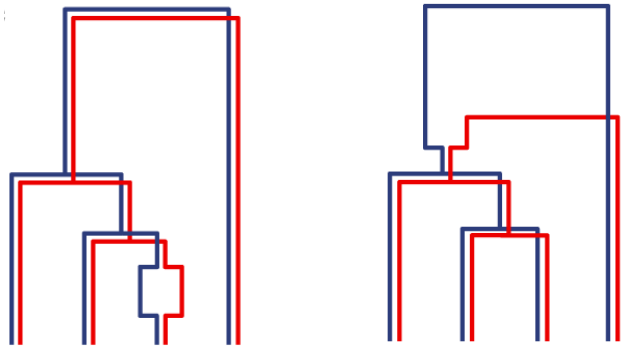
Recombination hotspots in malaria cont…



A tree-pair where we could
see recombination events, but don't



Tree-pairs where we cannot
see recombination events



The rate of recombination

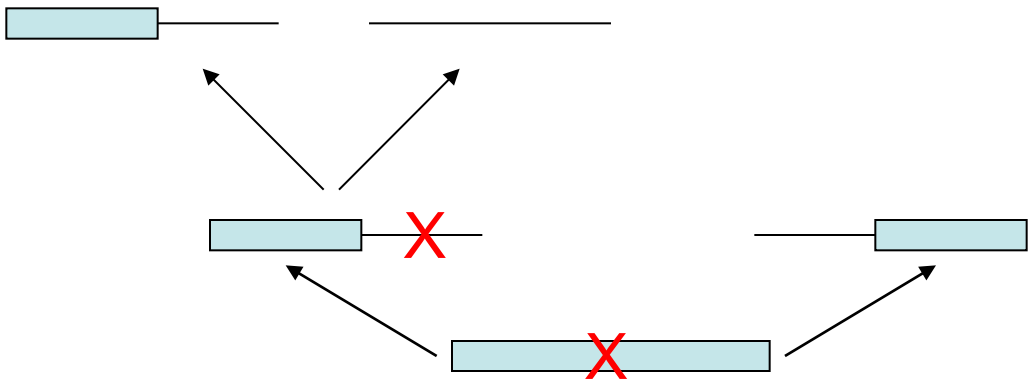
- Each generation, the probability of recombination between two loci is *r*, working in scaled time, this means that recombination occurs at rate *r*/2 per sequence where $r = 4N_e r$
- Recombination and coalescence occur independently, so looking back in time, recombination occurs as a Poisson process with rate *nr*, while coalescence also occurs as a Poisson process with rate $n(n\text{-}1)/2$
- The time until the next event is also a Poisson process with rate $nr + n(n\text{-}1)/2$, and the probability that the next event is a coalescent is

$$\Pr(rec) = \frac{n\rho}{n\rho + n(n-1)/2} = \frac{\rho}{\rho + n - 1}$$

Recombination in non-ancestral material

- Once a region has recombined, further recombination can occur in both ancestral lineages

- However, recombination in **non-ancestral** DNA cannot in anyway influence patterns of diversity (under a neutral model)



Properties of the ARG

- Unlike the basic coalescent, there are few results about the effects of recombination on gene genealogies that we can derive analytically

- For example, we cannot even calculate the expected number of recombination events in the history of a sequence
  - Though we can show it is less than infinity!

- There are some useful results about how many recombination events we can see
  - The key is that only a small minority of recombination events that occur in the history of the sample can ever be directly detected

Estimating the population recombination rate

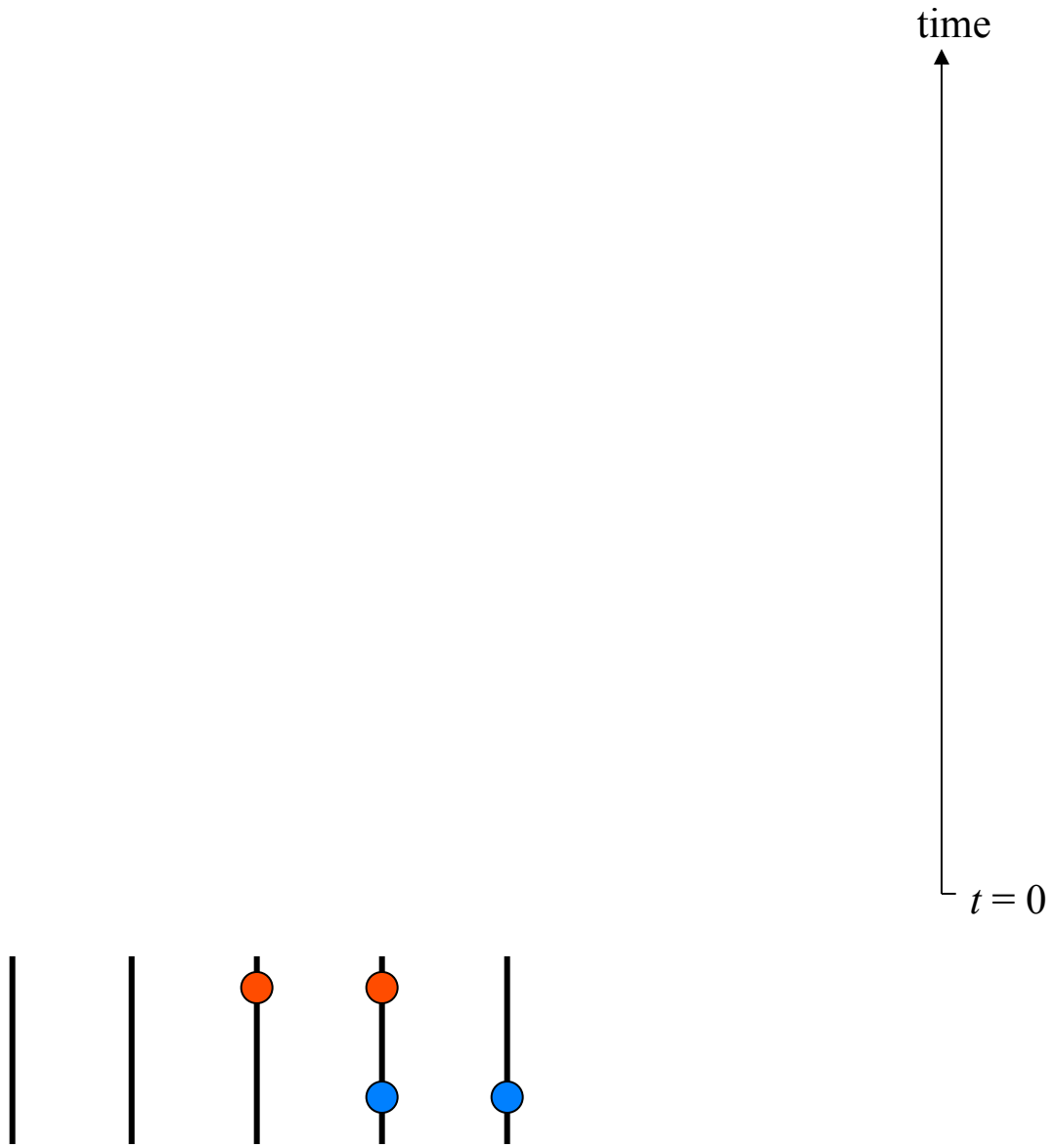
- Moment methods. Compare variance in pairwise differences to theoretical expectations under coalescent with recombination
  - Hudson (1987); later improved
  - Methods are fast but sampling variance is huge

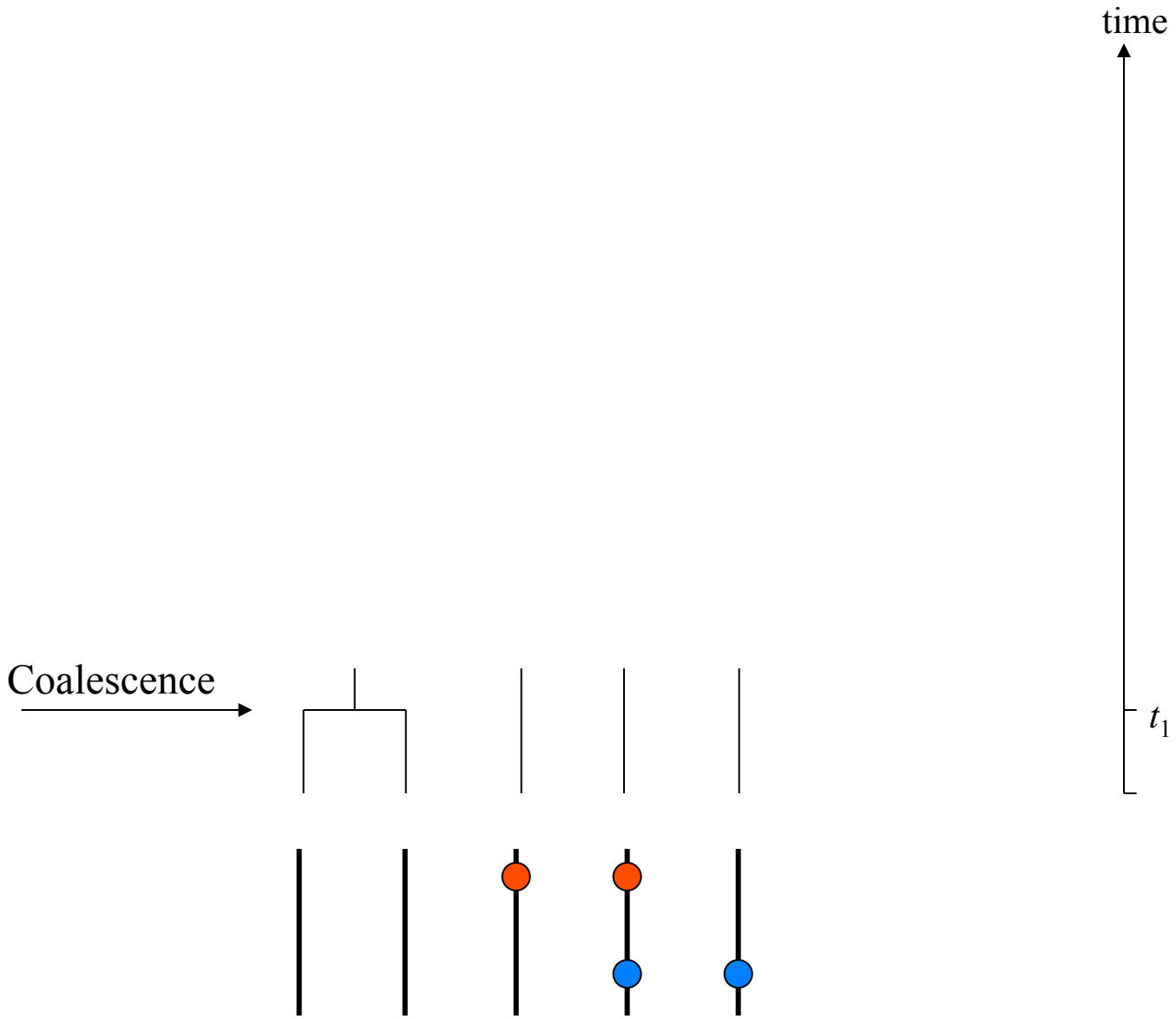
- Full-likelihood inference
  - Calculate the coalescent likelihood of observing the data for given values of theta and rho. Find most likely values
  - Uses all information in the data, but are computationally intensive

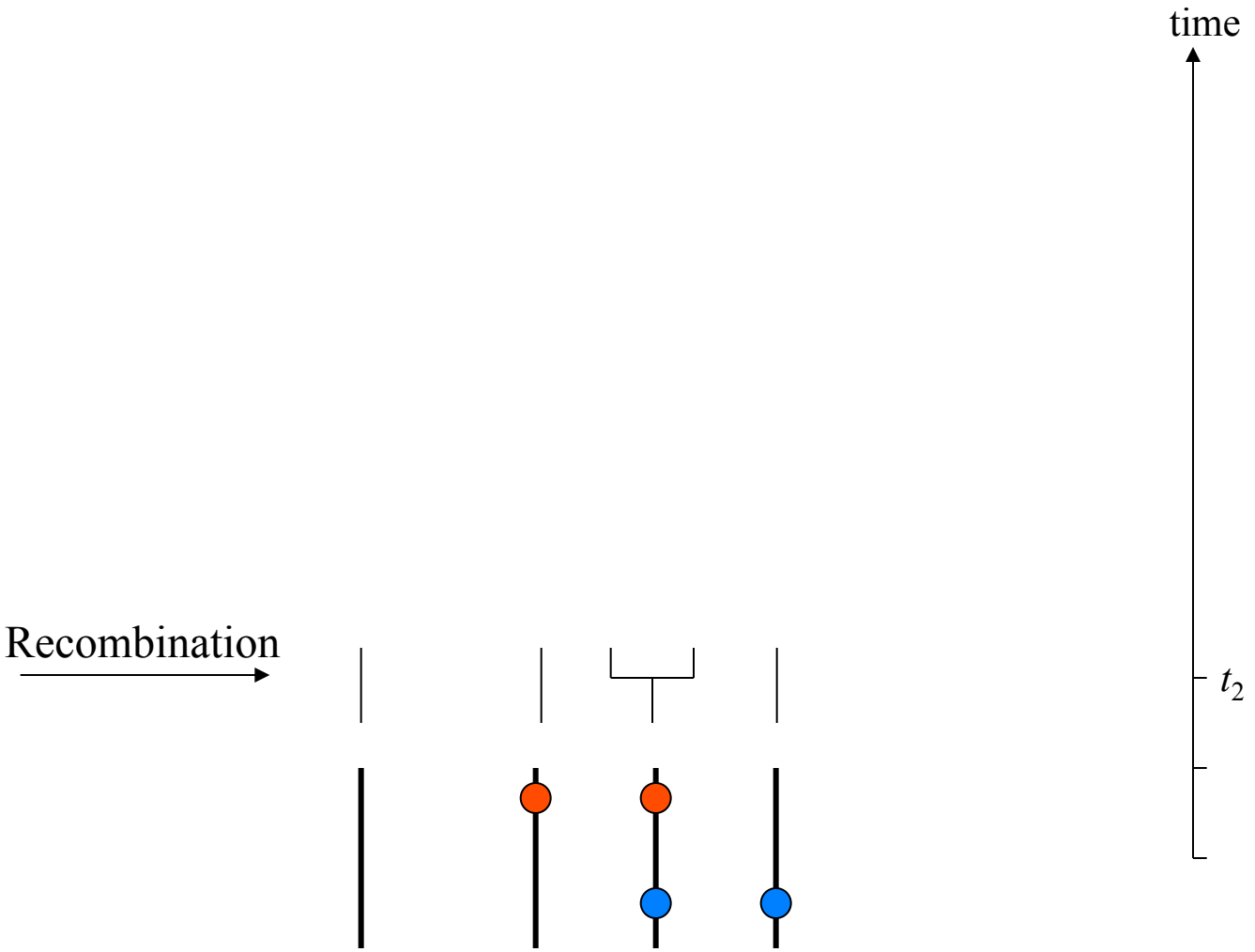
- Approximate-likelihood methods
  - Composite likelihoods
  - Approximations to coalescent with recombination
  - Likelihoods of data summaries

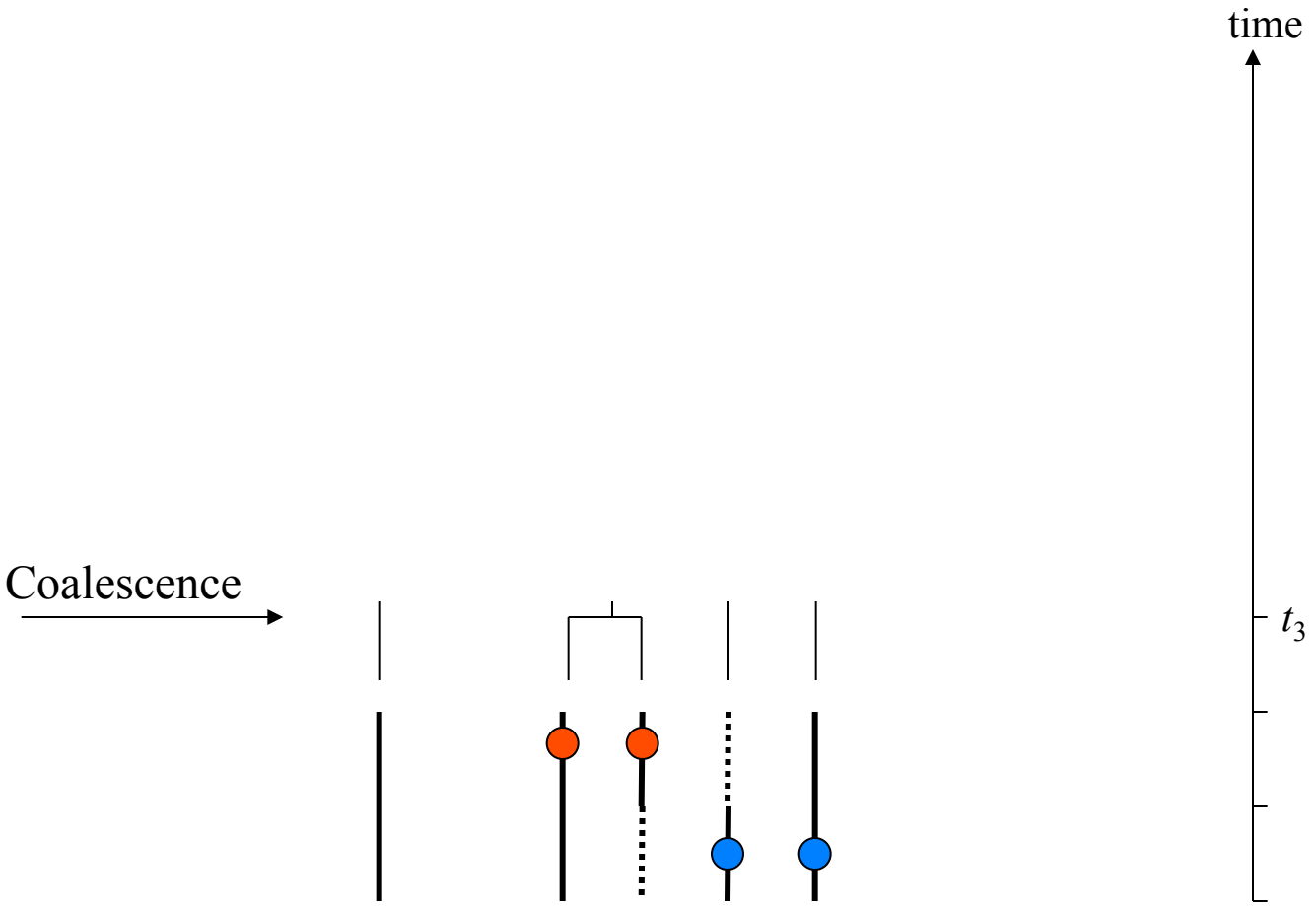
How do you calculate a likelihood?

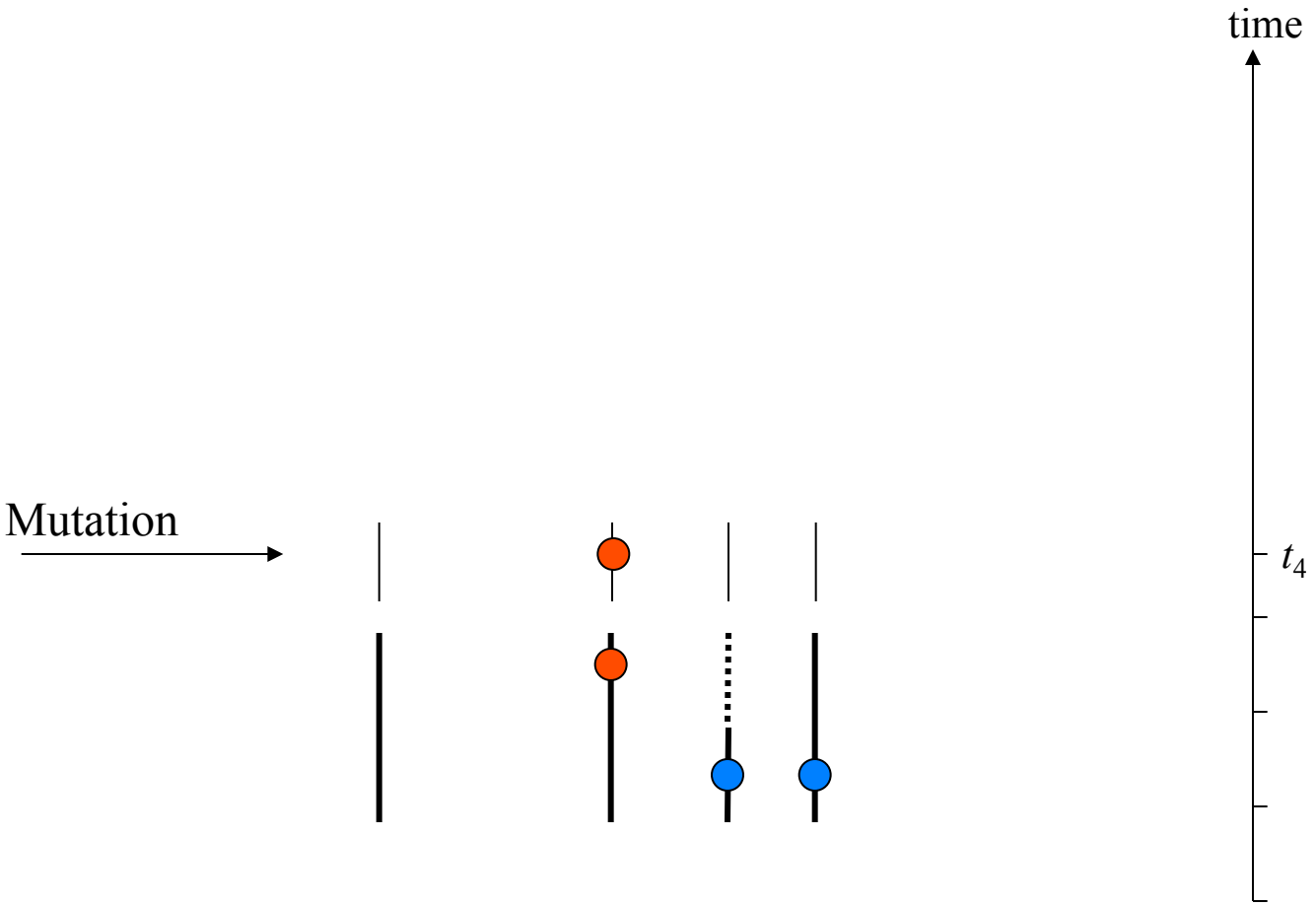
- A likelihood is proportional to the probability of the data given a model (and parameter values)
- We can estimate likelihoods by simulation (Monte Carlo). Broadly, we simulate lots of possible ancestral recombination graphs and calculate the probability of observing the data given the graph
- Advanced statistical techniques (MCMC and/or importance sampling) have to be used to make these approaches computationally efficient

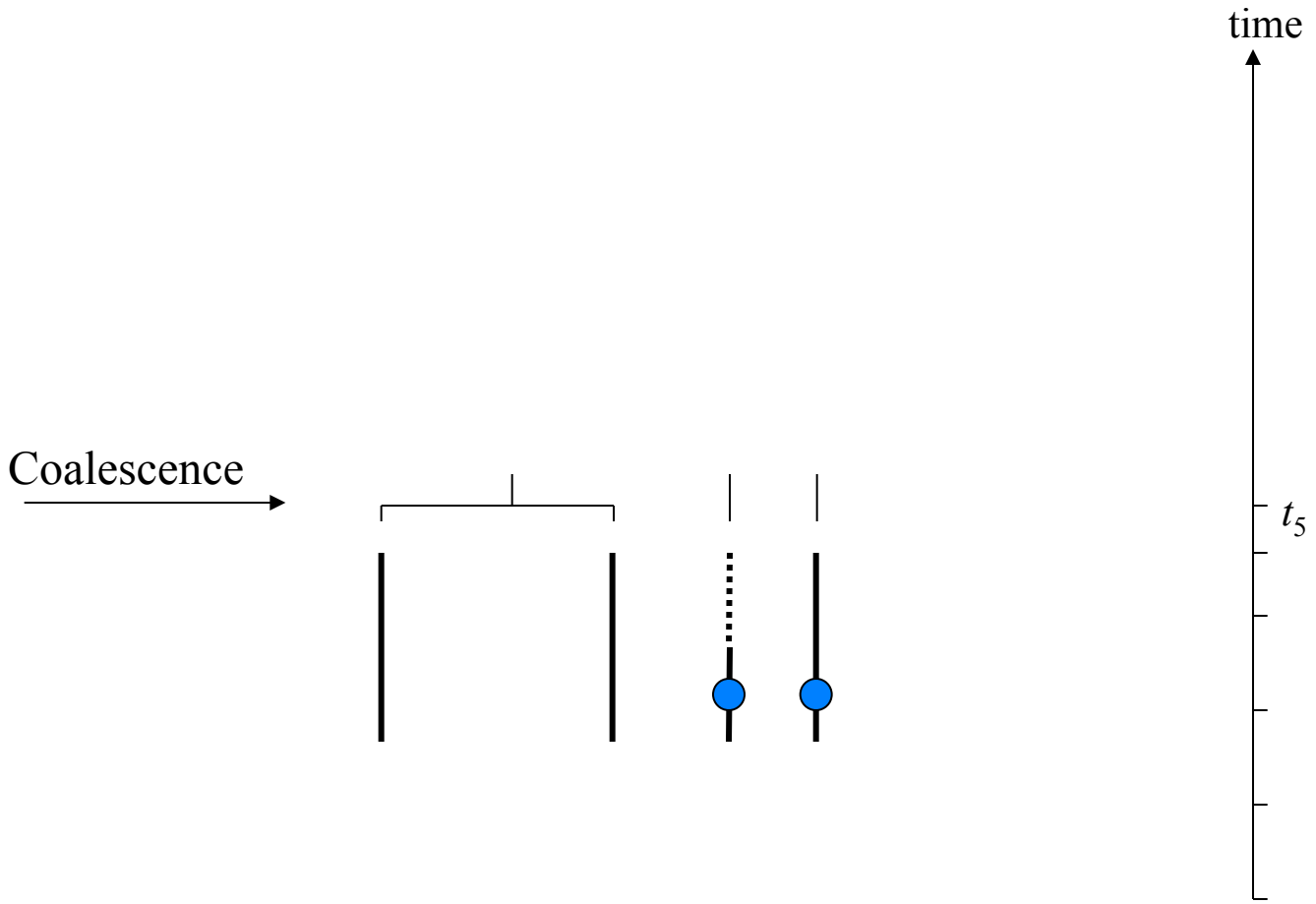

time
$t = 0$

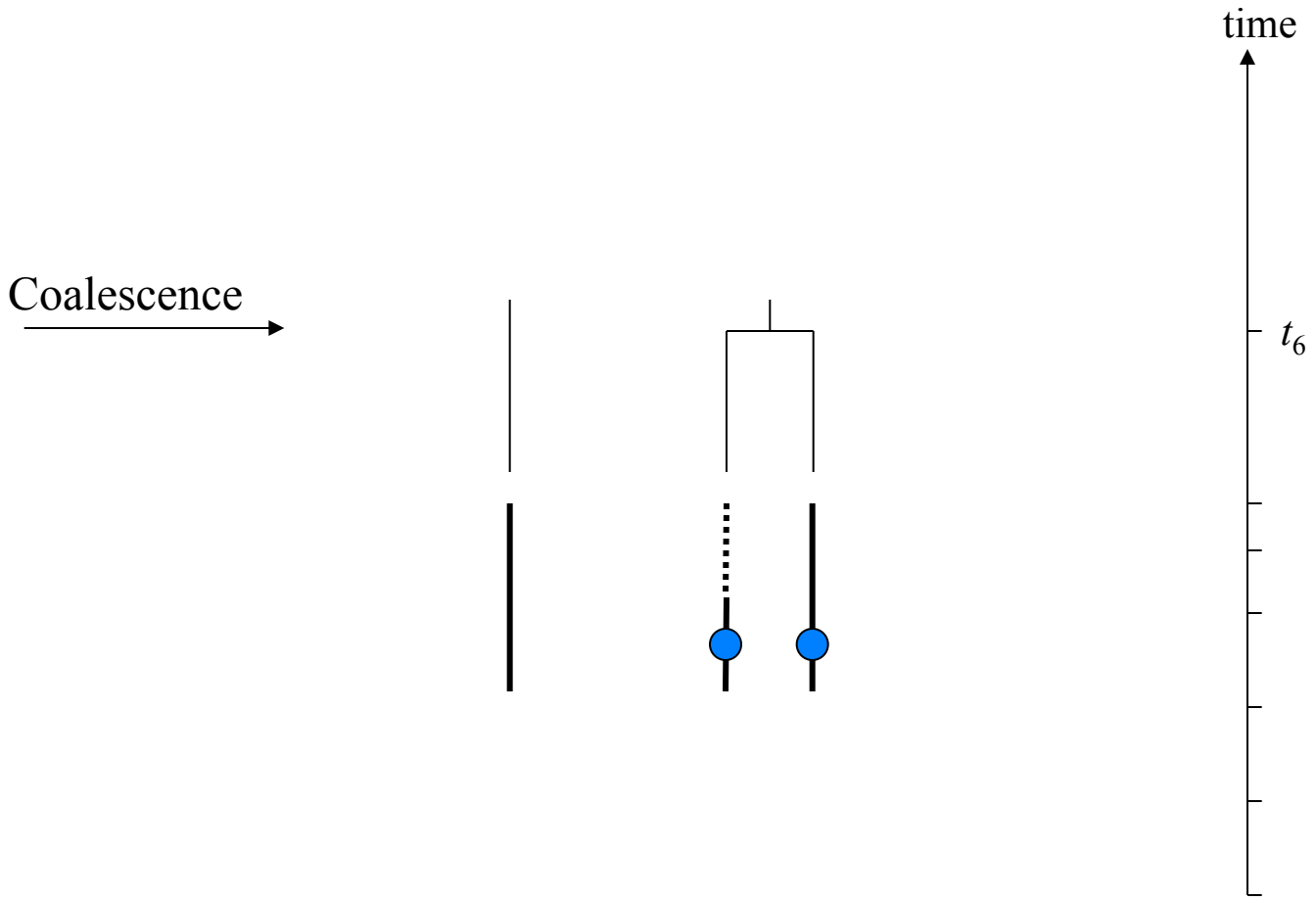

time
Coalescence
t_1

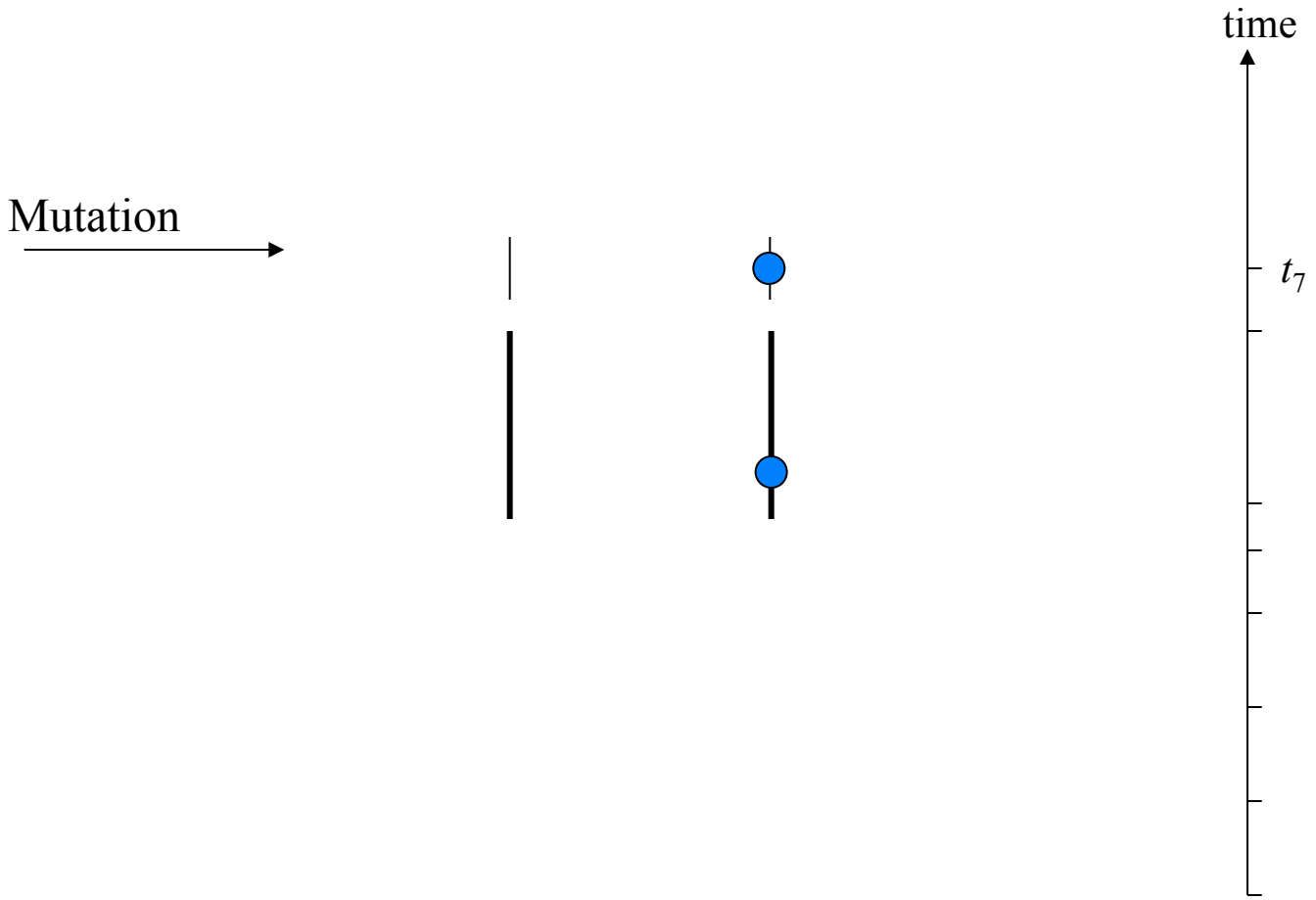

time
Recombination
t_2

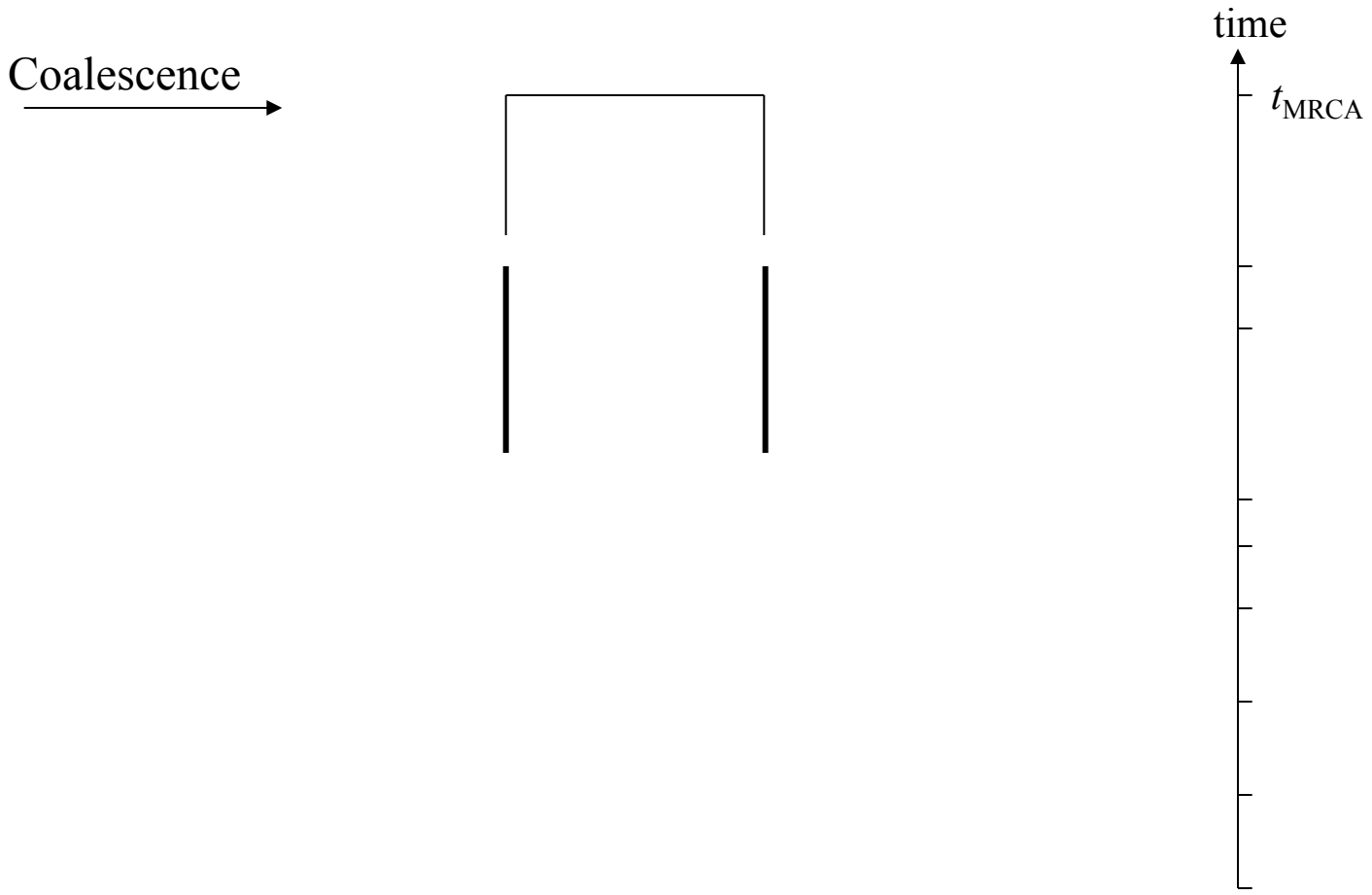

time
Coalescence
t_3

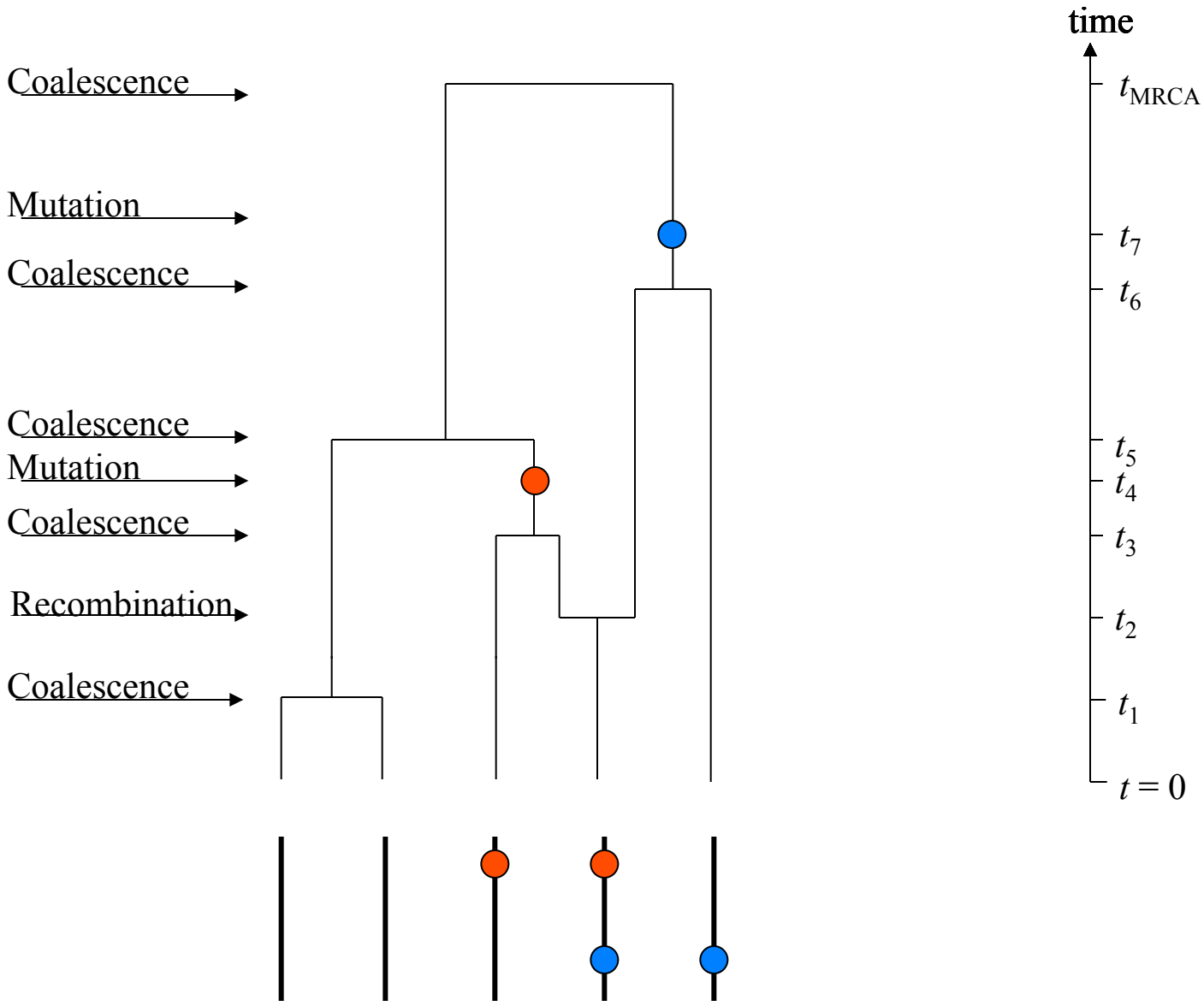

time
Mutation
t_4


time
Coalescence
t_5


Coalescence
time
t_6


time
Mutation
t_7


Coalescence
time
t_{MRCA}

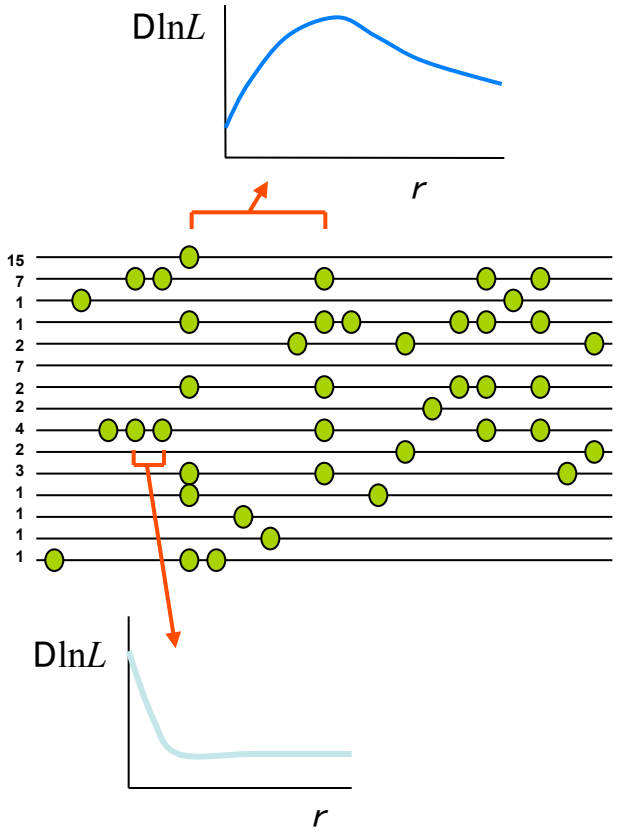


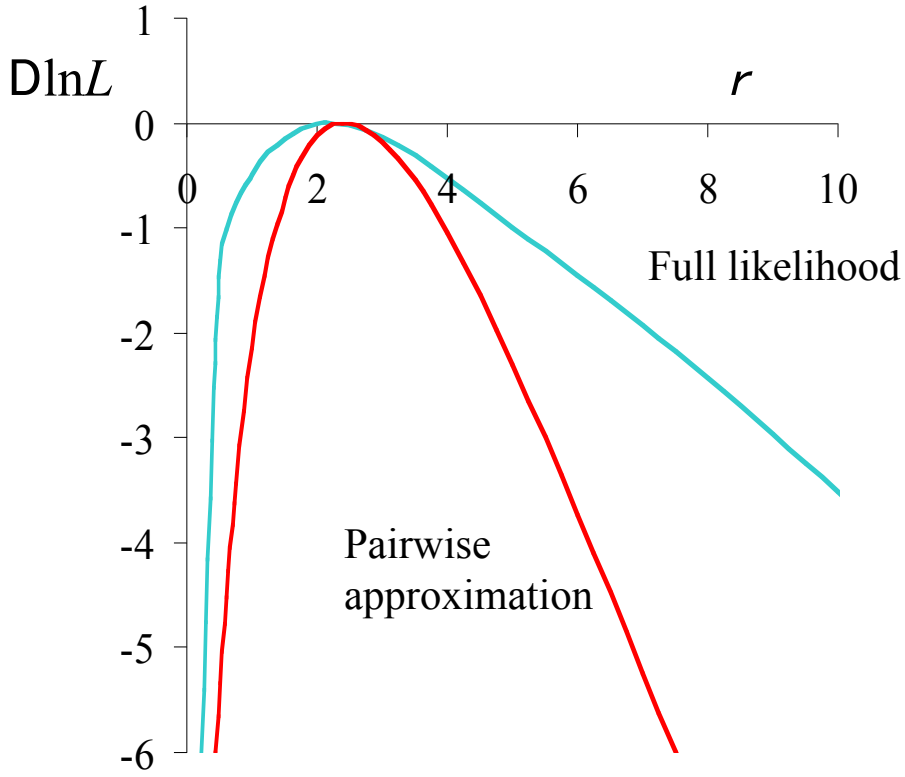
$$L(History) = \prod_{events} \Pr(event)$$

$$L(Data) \approx \frac{1}{N} \sum_{i} L(Guess_i)$$

(this approach is computationally unfeasible for any realistic data set)

Composite likelihood estimation of $4N_e r$: Hudson (2001)

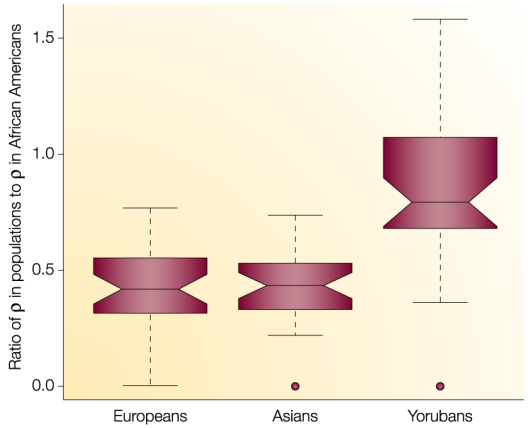




What can we learn from estimating the population recombination rate?

- Does one genomic region experience more or less recombination than another?
  - Does the region experience any recombination?

- Does one population have a greater effective population size than another?



Stumpf and McVean (2003)

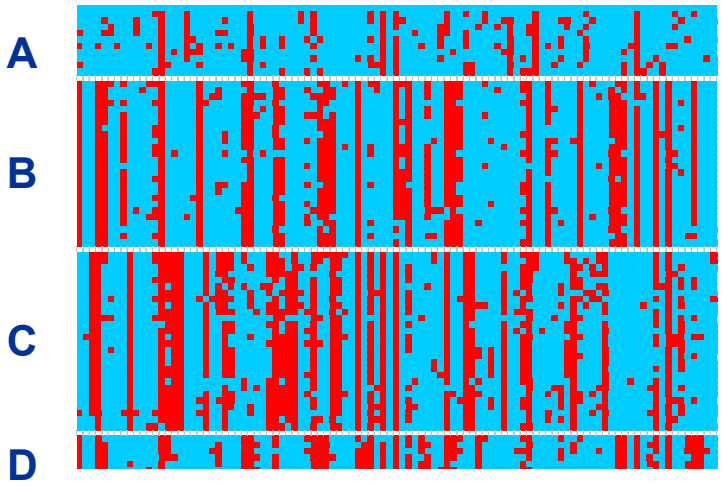
- Is a model of constant population size and constant recombination rate a good description of the data?
  - Evidence for selection, complex demography or recombination rate variation

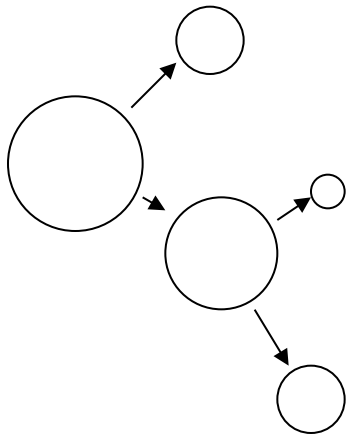
- Is there recombination rate variation within the region studied?

Example: HIV subtypes

- Mosaic genomes indicate recombination between different genomes can happen
  - But how common is recombination

Genome	Gene	L	n	q_W	P_{Lk}	r
HIV1A	env_{12}	1412	17	0.061	**0.012**	**34**
HIV1B	env_{12}	1316	93	0.102	**0.020**	**>100**
HIV1C	env_{12}	2073	28	0.062	**0.002**	**>100**
HIV1D	env_{12}	1394	14	0.092	**0.009**	**>100**
HIV2	env_{12}	1364	21	0.102	**0.000**	**>100**





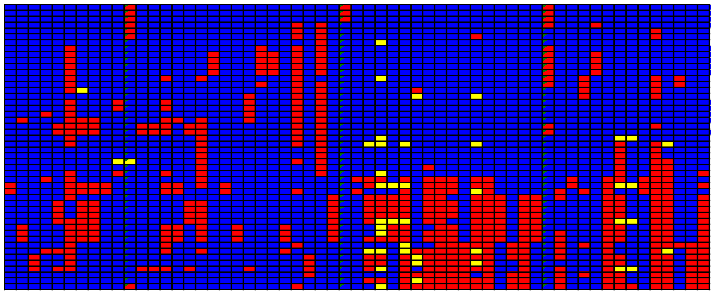
Example: Recombination rate variation in human LPL

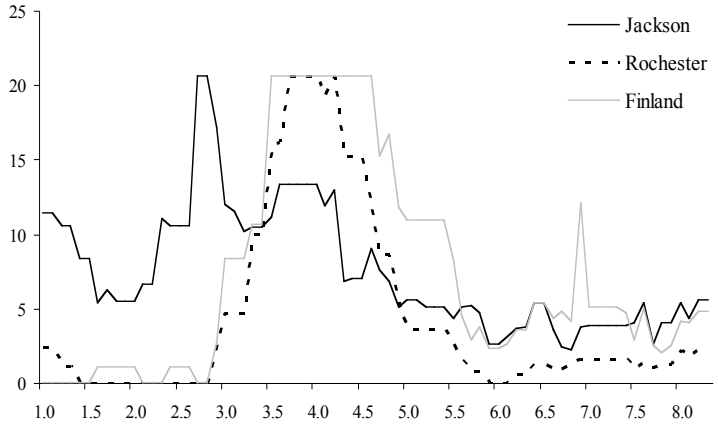
- 9.7 kb region sequenced in 48 African American chromosomes (Nickerson *et al.* 1998)
- Using N_e = 10,000 r = 1cM/Mb: $r = 4$
- Using composite likelihood approach:

$$\hat{\rho} = 29$$

- BUT
  - Evidence for a region of elevated recombination within the gene indicating constant recombination rate model may be inadequate

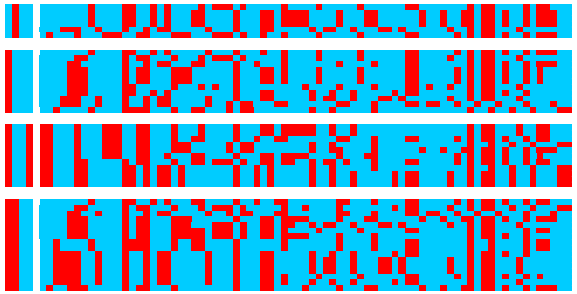
Haplotypes



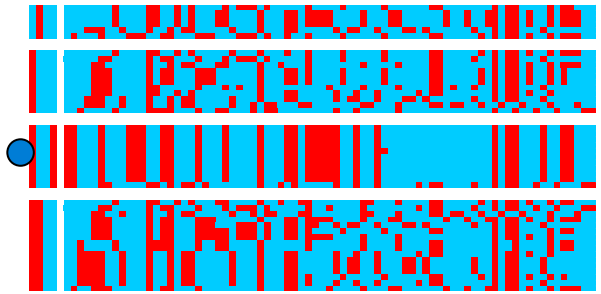


Using recombination to learn about natural selection

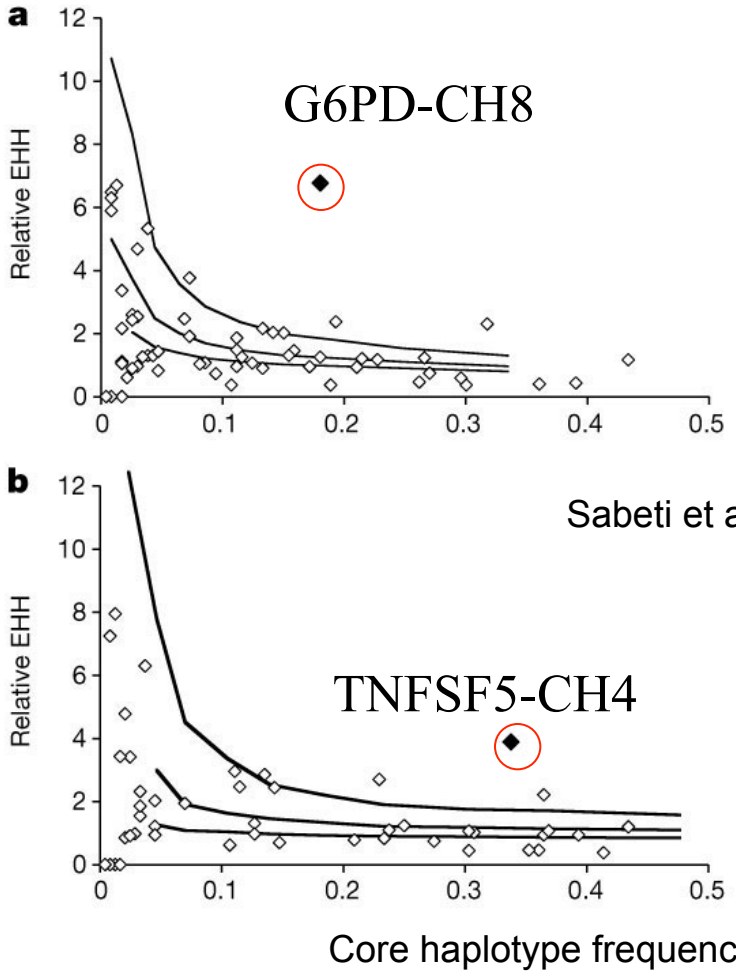
- Various tests have been proposed to look for the signature of recent adaptive evolution in patterns of linkage disequilibrium and haplotype structure



Neutral – decay of haplotype structure even across core haplotypes



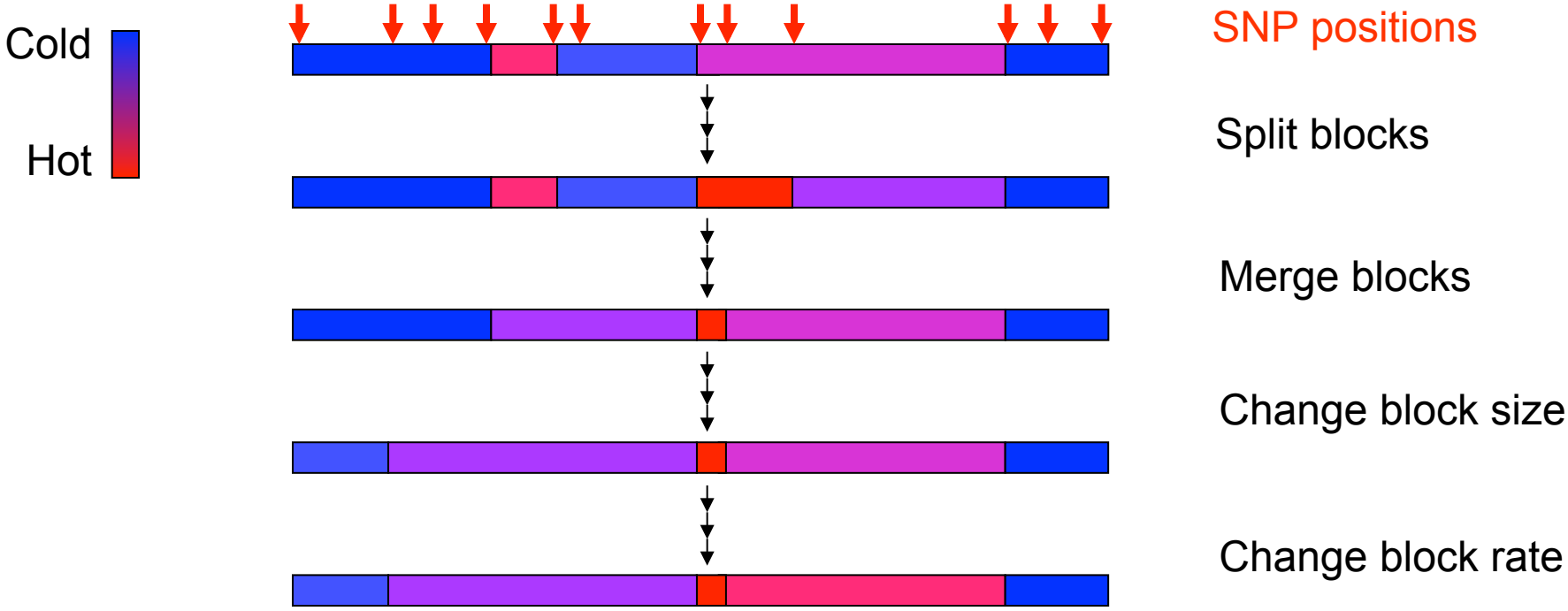
Selected – one (or more) core haplotypes show unusually extensive haplotype structure



Sabeti et al (2002)

Fitting a variable recombination rate

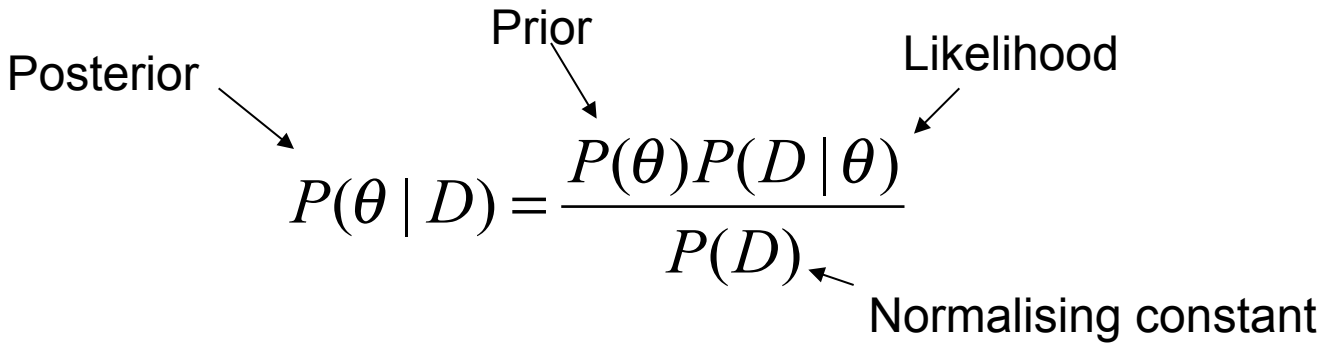
- Use a reversible-jump MCMC approach (Green 1995)



What is MCMC?

- MCMC stands for Markov Chain Monte Carlo. It is a widely used technique in modern Bayesian statistics

- In Bayesian statistics we want to learn about the probability distribution of the parameters of interest given the data = the **posterior**



- The prior represents our uncertainty before the experiment, the posterior represents our uncertainty after the experiment

What is the Monte Carlo bit?

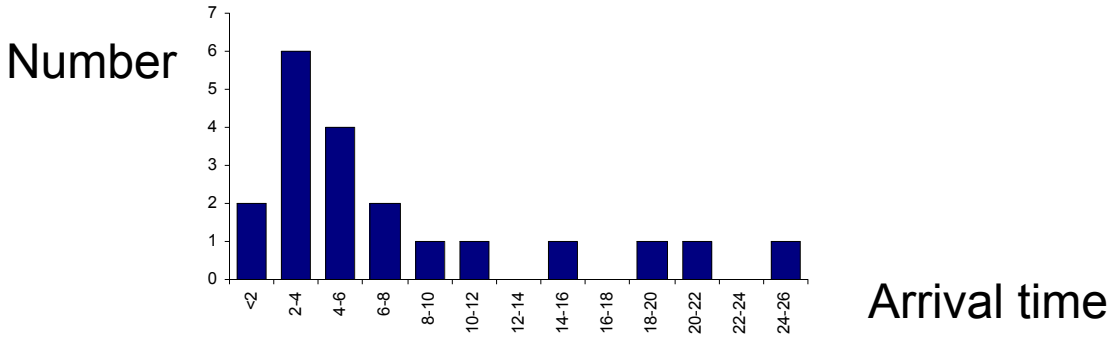
- In our case, there is no simple analytical expression for the likelihood, so to explore the posterior distribution of recombination rates we use stochastic simulation

- In effect, this means making an initial guess at the recombination rate, then making small changes to the rate profile. Improvements (i.e. increases in likelihood) we always accept, but we only reject 'worse' changes by the probability that they are indeed worse.

- There is some very elegant theory (called Metropolis-Hastings) that says that simply making small changes allows us to fully (and appropriately) explore the posterior

An example

- There are two types of bus: frequent (average arrival time = 5 mins) and rare (average arrival time = 10 mins). I see the following times

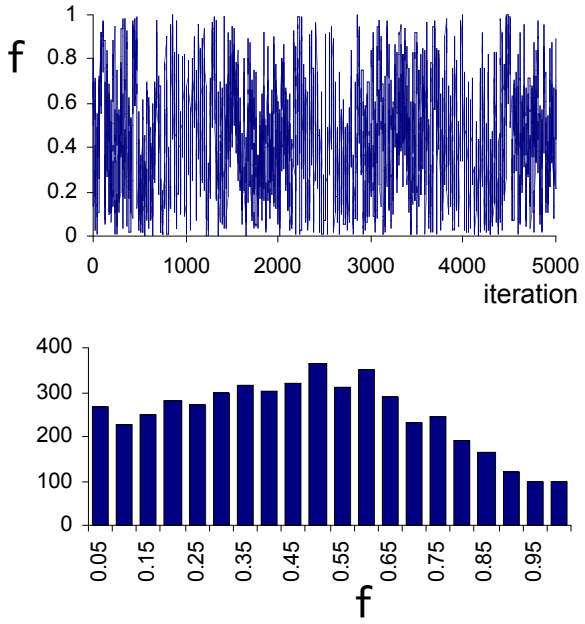


- If I model the inter-arrival times as a mixture of two exponential distributions, I want to perform inference on the proportion of frequent and rare buses

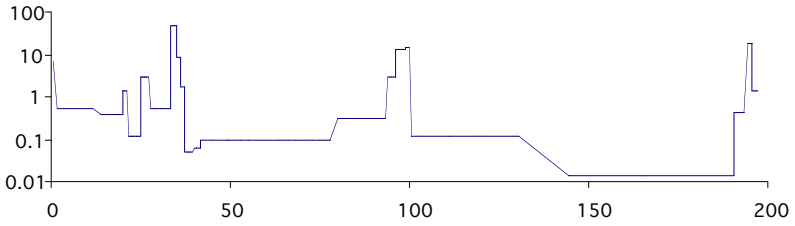
$$P(X_i = x) = \frac{\phi}{5} e^{-x/5} + \frac{1-\phi}{10} e^{-x/10}$$

- There is no analytical expression for the result

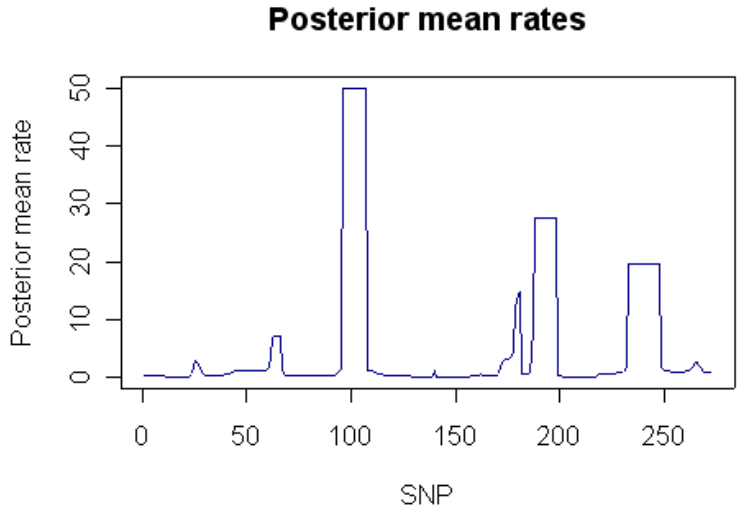
- I will put a uniform prior on the mixture proportion f
- Starting with an initial guess of f = 0.2, I will propose changes uniform on $f-e, f+e$
- (Note that if I propose a value < 0 or >1 I must reject it)
- 5000 samples from the chain give me the following Markov Chain

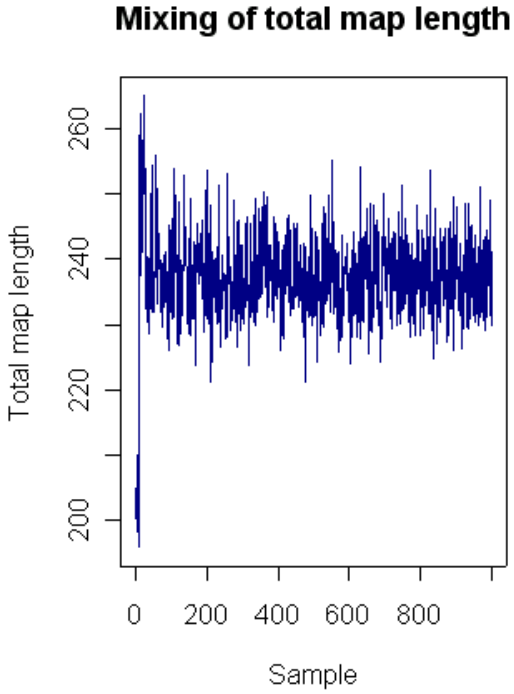


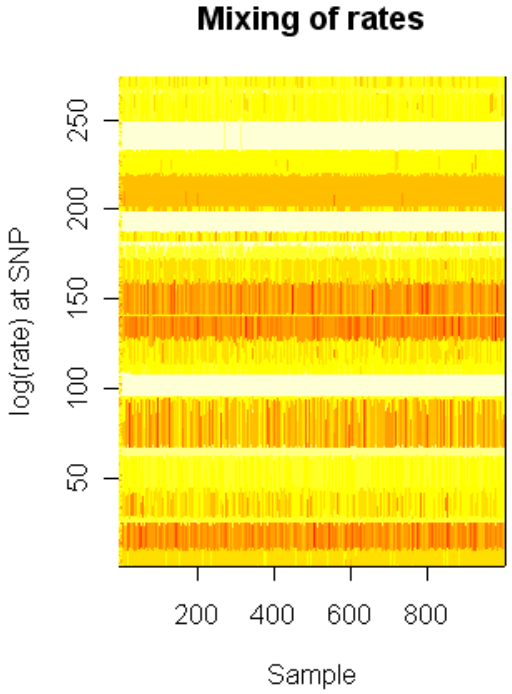
RJMCMC in action for recombination rates



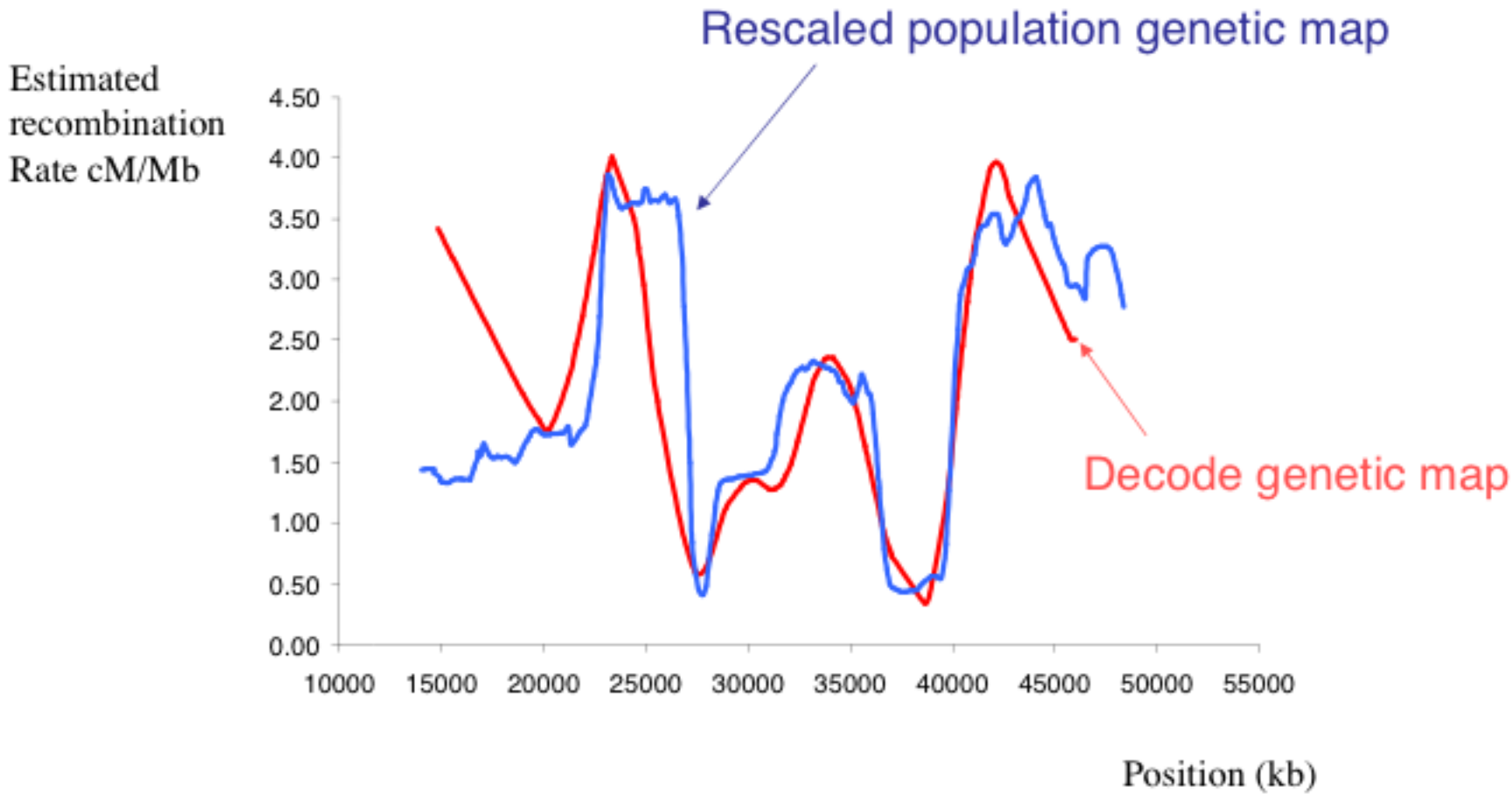
Summarising the output







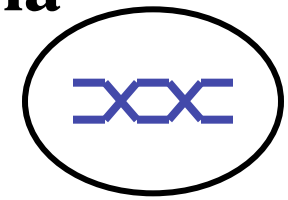
Mapping recombination hotspots … (human chromosome 22)



Composite Likelihood - RJMCMC - LdHat v.2.0

Population genetic inference and recombination

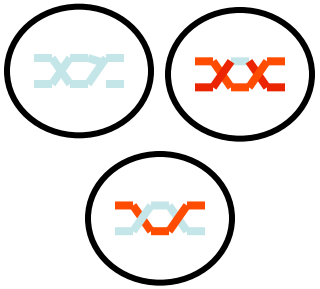
Malaria



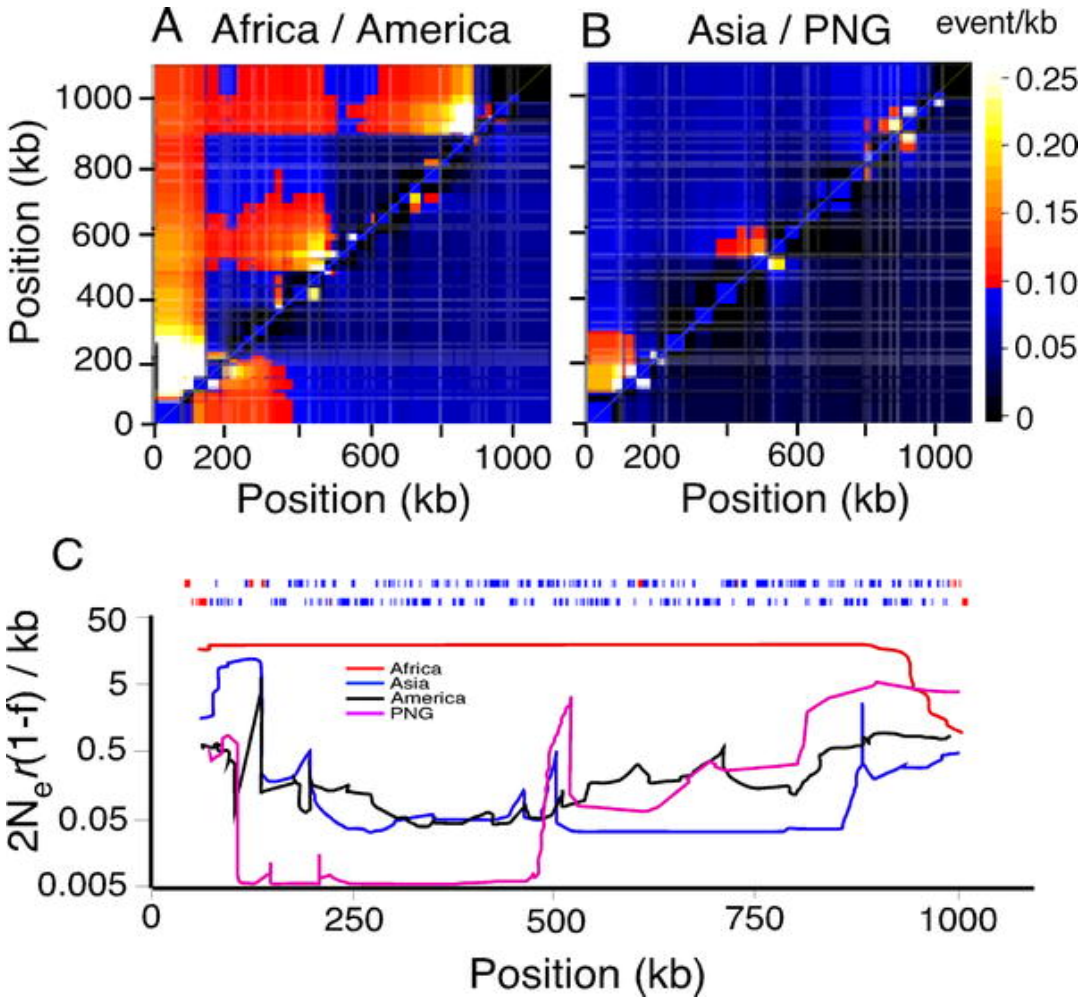
Recombination essential

BUT

Need multiple infection
for evolutionary effect



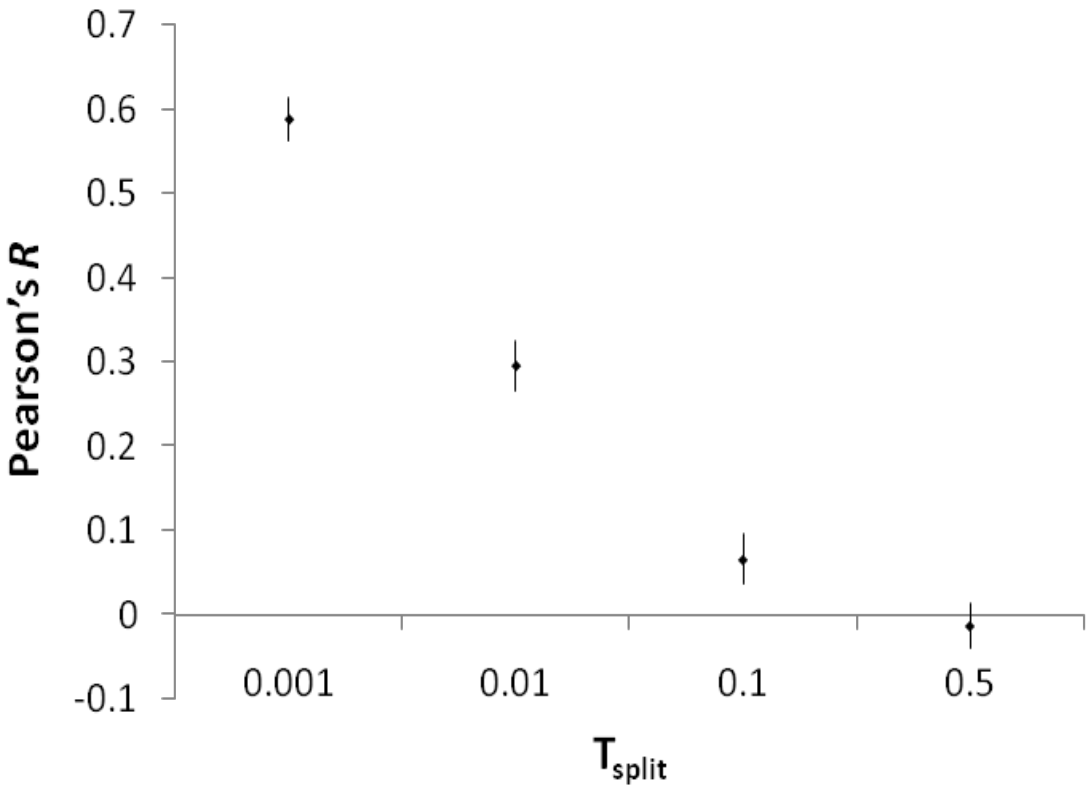
Recombination variation along Chromosome 3 in *Plasmodium falciparum*



Rmin
(Myers and Griffiths)

LDhat

Correlation in recombination landscapes between pairs of populations that diverge at different time intervals



Summary

- Recombination is a fundamental biological process that influences patterns of genetic variation
- Ignoring the effects of recombination can lead to serious problems in inference, but its presence can also increase power
- Recombination generates splitting in genealogical histories and results in ancestral recombination graphs
- We can learn much about recombination from nonparametric approaches
- Coalescent-based approaches allow us to estimate recombination rates and learn about variation in the recombination rate

Plan for Module 21

Wednesday 6/23	1:30-3:00	Introduction	Philip
	3:30-4:00	Introduction (continued)	Philip
	4:00-5:00	Introduction	Mary
Thursday 6/24	8:30-10:00	Recombination	Philip
	10:30-12:00	Recombination practical	Philip
	1:30-3:00	Population size and structure	Mary
	3:30-5:00	Gene flow practical	Mary
	5:00-6:00	Tutorial	Mary/Philip
Friday 6/25	8:30-10:00	Selection	Philip
	10:30-12:00	Selection practical	Philip
	1:30-3:00	Applications and study design	Mary
	3:30-5:00	Coalescent practical	Mary

Details–Thursday 6/24

- Thursday morning: Recombination
  - Genetic recombination
  - Linkage disequilibrium
  - LDhat, RJMCMC, Phase
  - Hands-on recombination exercise
- Thursday afternoon: Growth and Gene Flow
  - Population growth and shrinkage
  - Population subdivision and gene flow
  - Population divergence
  - Genealogy samplers: Migrate-N, Lamarc, Beast, IM
  - Hands-on gene flow exercise

Variants and extension of the coalescent

- Population growth/shrinkage over time
- Migration between populations
- Population divergence
- Mutation rate variation
- Times of significant events
- Recombination (Philip Thursday)
- Selection (Philip Friday)

Outline

- What kind of information is available about this evolutionary force or process?
- How is it usually parameterized?
- What algorithms or programs deal with it?
- What special issues arise with this force?

Variable population size

- During times when a population is large, lineages coalesce slowly
- During times when a population is small, lineages coalesce quickly

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

Parameterization of growth

- No growth – Θ is constant
- Exponential growth model – growth rate g
- Logistic growth model – growth rate r, carrying capacity K
- Stepwise growth model – step time, Θ before and after
- Free growth model – a series of steps

Not easy to tell the models apart!

A cautionary tale

- Mismatch distribution–distribution of number of differences between haploptypes
- Theoretical distribution is exponential when no growth
- With growth, it has a peak
- Score all pairwise mismatches from human data–peak is seen
- This makes sense as human population has grown....

BUT!

A cautionary tale

- Simulate non-growing populations
- Distribution NEVER looks exponential
- (next slide from Slatkin and Hudson 1991)
- Why?

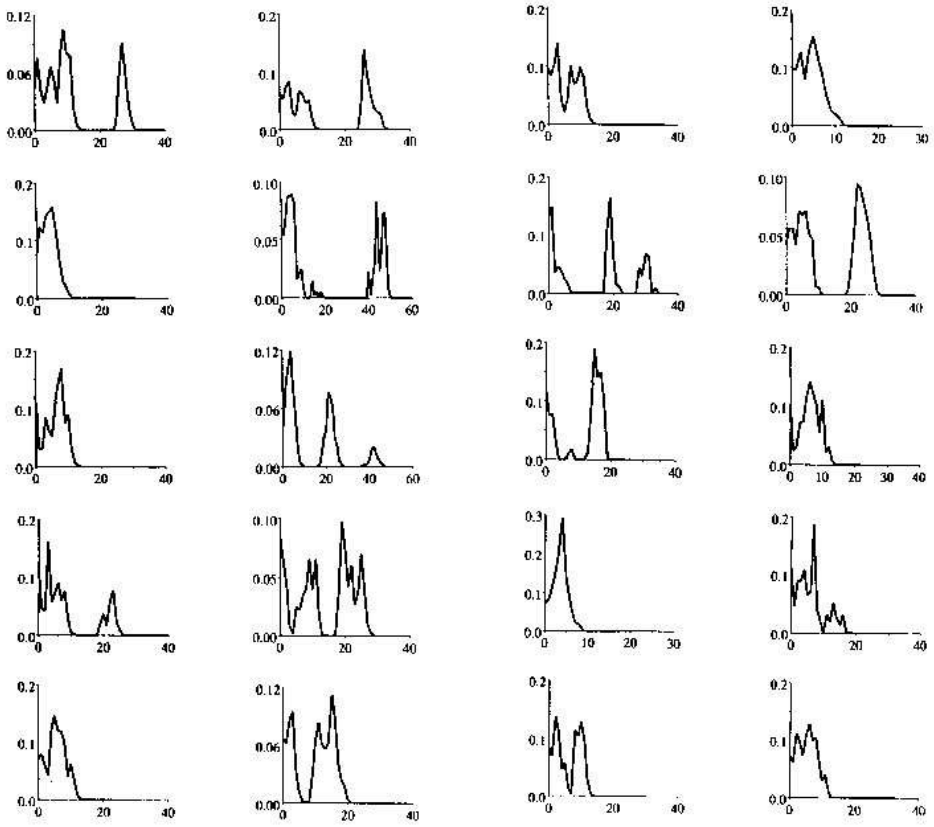


FIGURE 3. Frequency distributions of pairwise differences for 20 replicate simulations. The distributions of all 1225 pairs in samples of 50 genes from a panmictic population are plotted. The data were generated using a simulation program described in the text. In each graph, the abscissa is the number of sites at which two samples differ and the ordinate is the fraction of pairs that differ.

A cautionary tale

- Expected distribution of ONE pairwise mismatch is exponential
- Multiple draws from the same population are not independent
- The peaks come from deep coalescences in the tree

A cautionary tale

Are mismatch distributions still useful?

- One pair per locus would give "expected" distribution
- Raggedness of distribution can suggest lack of growth
- Not an efficient way to use the data

Bottom line: locus history is a TREE

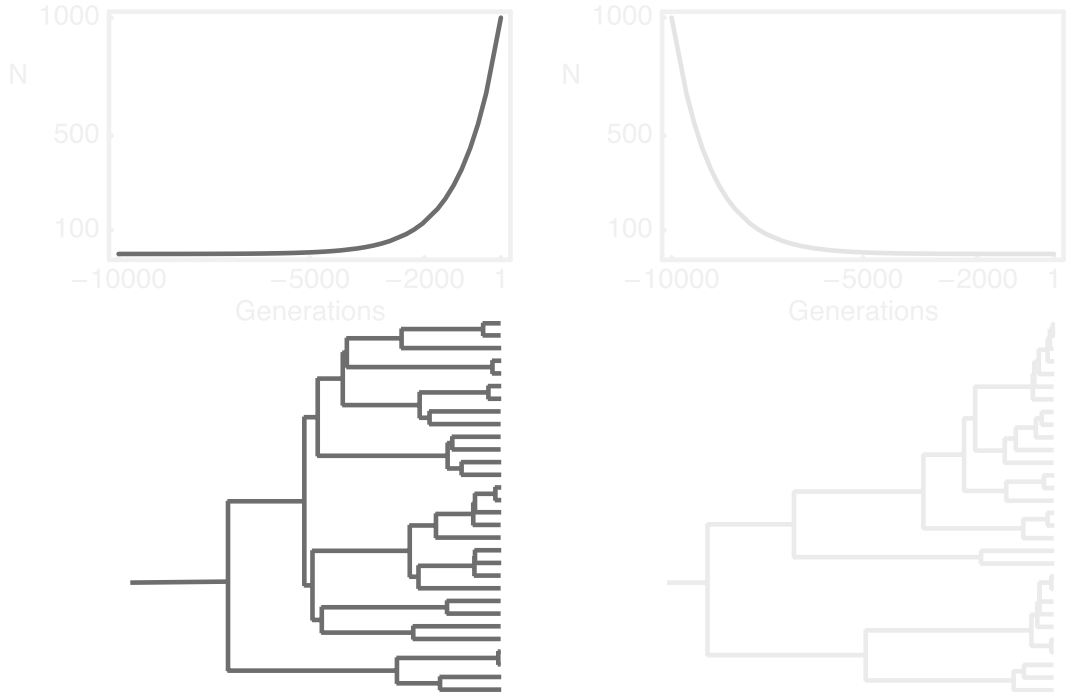
Population growth: non-sampler approaches

- Summary statistics based on:
  - Mismatch distribution
  - Between-locus variability
  - Allele size distribution (microsatellites)
  - Imbalance between variance and heterozygosity
- Nested clade analysis
- Skyline plots
- Approximate Bayesian Computation bottleneck detection

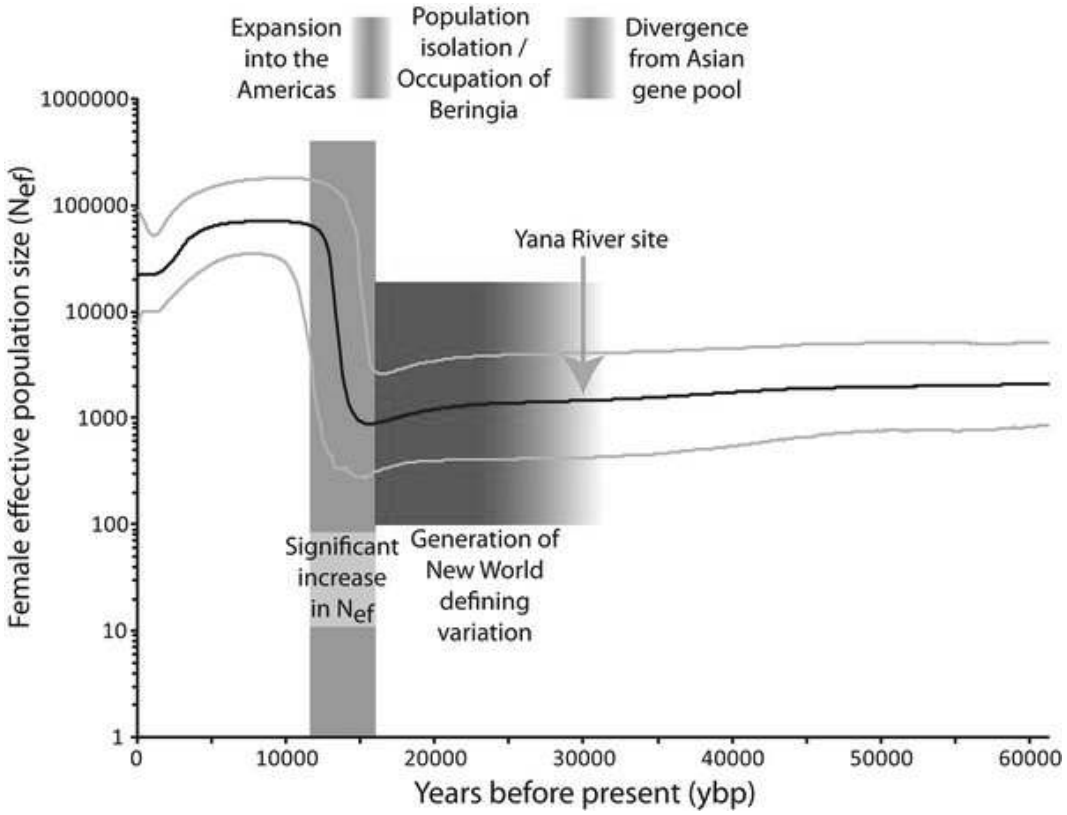
Population growth: sampler approaches

- Exponential growth: LAMARC, IM, BEAST, GENETREE
- Growth estimation is biased upwards with one locus
  - Confidence intervals are more reliable than maxima
  - Multiple loci help a lot
  - Multiple time points are even better
- BEAST offers Bayesian skyline plots for more detail on growth

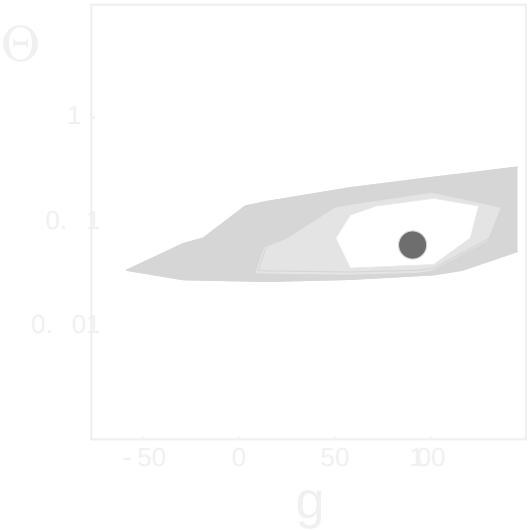
Exponential population size expansion or shrinkage



Bayesian skyline plot: Heled and Drummond 2008



Genealogy-sampler inference of exponential growth



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

What does g mean?

The parameter g often causes confusion.

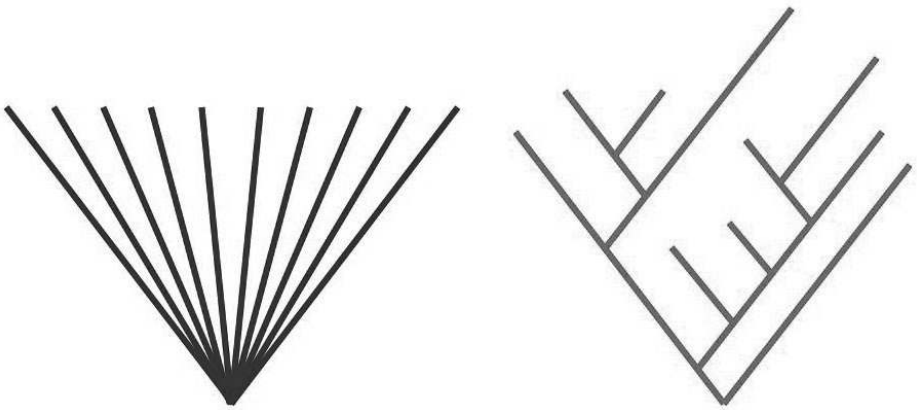
$$\Theta_{t\mu} = \Theta_{now} e^{gt\mu}$$

To interpret g we need an external estimate of the mutation rate μ. Given that, we can ask how large the population was a given number of generations or years (depending on the units of our mutation rate estimate) in the past, using this equation.

Positive g is a growing population, negative g is a shrinking one.

General issues with growth estimation

- Low statistical power–multiple loci needed
- Too-fast growth turns the tree into a star
  - Star shows fast growth but can't pin down rate
  - Star-like data have little power for inferring other parameters



General issues with growth estimation

- If growth is fast, all growth models look similar
- Very recent growth has little effect on the coalescent, even if extreme
- Only most recent episode of growth likely to be visible
- Ancient DNA gives MUCH more power for growth inference than single time point samples

Specific issues with genealogy-sampler growth estimation

- Multiple unlinked loci not always available
  - LAMARC can compensate by using partially linked loci
  - BEAST can compensate by using multiple time points
- Growth rates so high that co-estimation of Θ and g not possible
  - If one parameter is held constant, the other can be estimated
  - Multiple time point samples in BEAST can resolve this problem
- Growth estimate contains unknown μ
  - Multiple time point samples in BEAST can resolve this problem
  - Mutation rate can be measured experimentally or inferred from phylogenetic data plus fossil dates

Migration among stable populations

Parameterization of migration rates

- m – chance that a lineage migrates each generation
- $M = \frac{m}{\mu}$ – scaled by mutation rate
- $4N_e m$ – scaled by population size (migrants per generation)

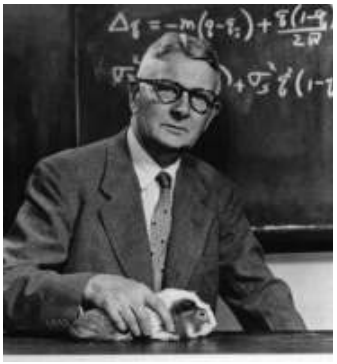
Non-sampler approaches

- F_{ST} summary statistics (contrast within-population and between-population variation)
- Haplotype sharing

Arlequin is a widely used program which carries out these (and many other) tests:

http://cmpg.unibe.ch/software/arlequin35/

FST in practice



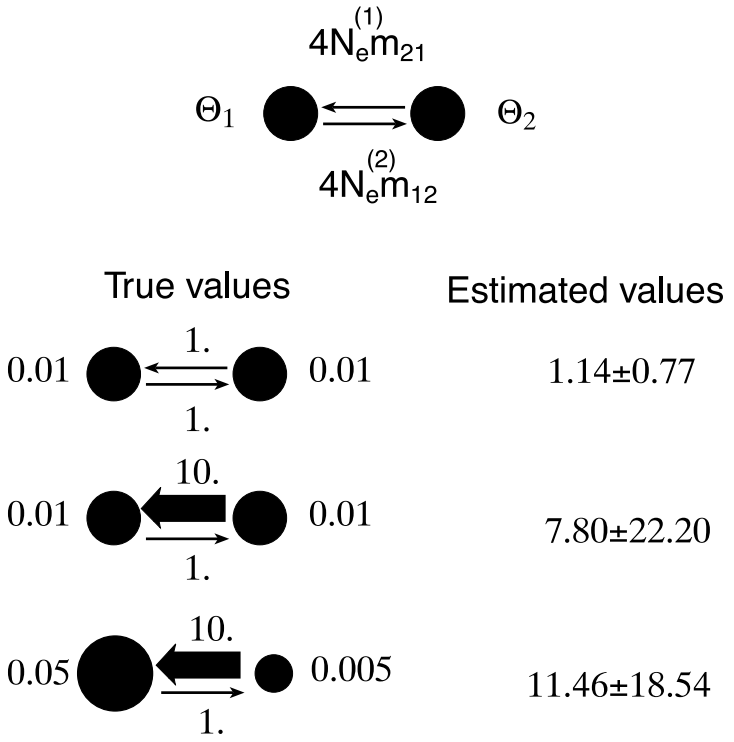
Sewall Wright showed that

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

and that it assumes

- All migration rates are the same
- All subpopulation sizes are the same

FST in practice



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.

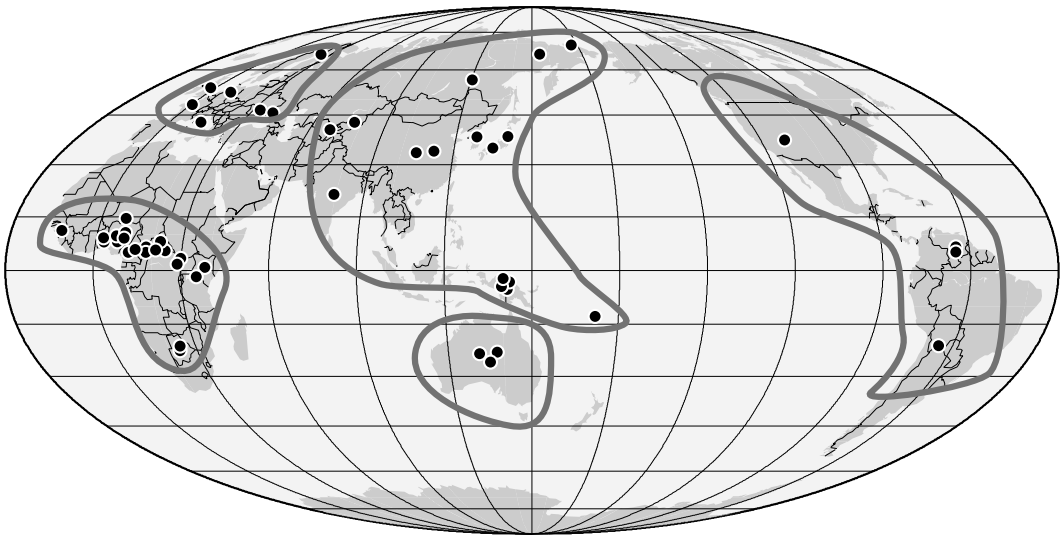
	Population 1		Population 2	
	Θ	$4N_e^{(1)}m_1$	Θ	$4N_e^{(2)}m_2$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

AMOVA

Analysis of Molecular Variance

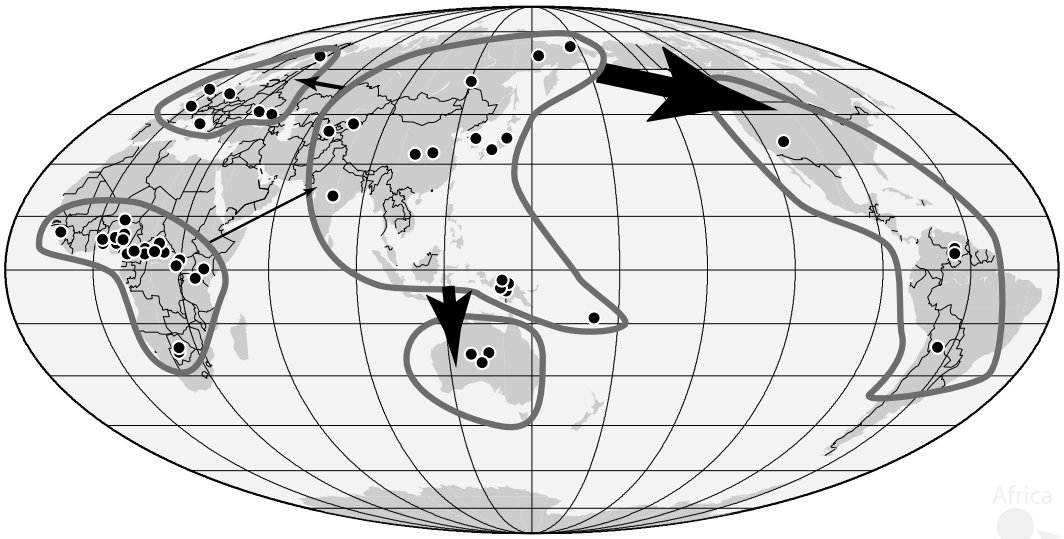
- Count differences among haplotypes within and between populations
- Compare to a null expectation from permuted data
- Infers degree of population subdivision; can be mined for estimates of specific migration rates
- More flexible than F_{ST}
- Commonly done with Arlequin

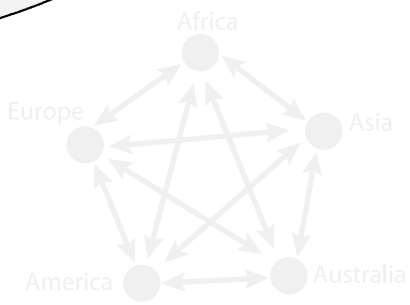
Complete mtDNA from 5 human"populations"



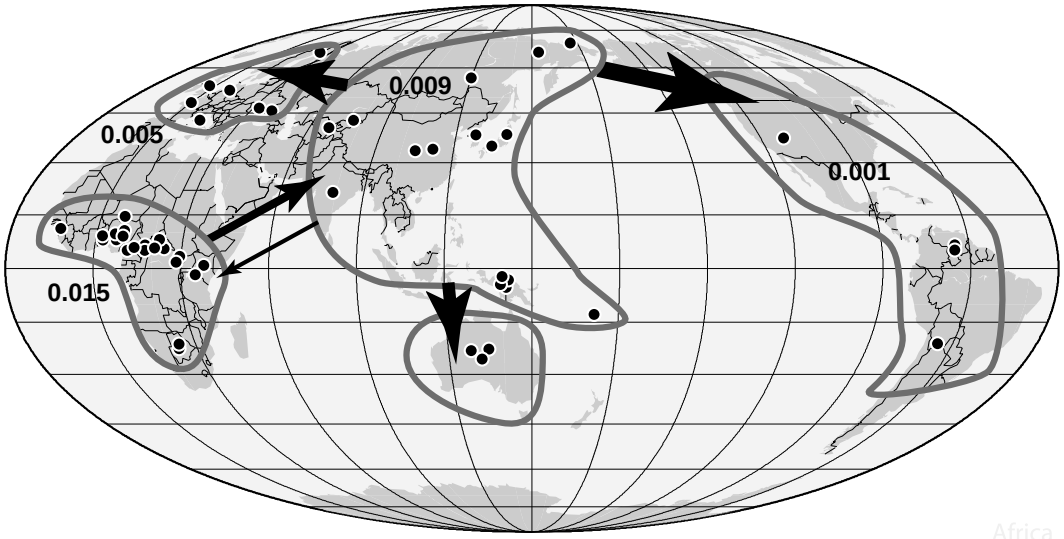
A total of 53 complete mtDNA sequences ($\sim$ 16 kb):
Africa: 22, Asia: 17, Australia: 3, America: 4, Europe: 7.
Assumed mutation model: F84+Γ

Full model: 5 population sizes + 20 migration rates





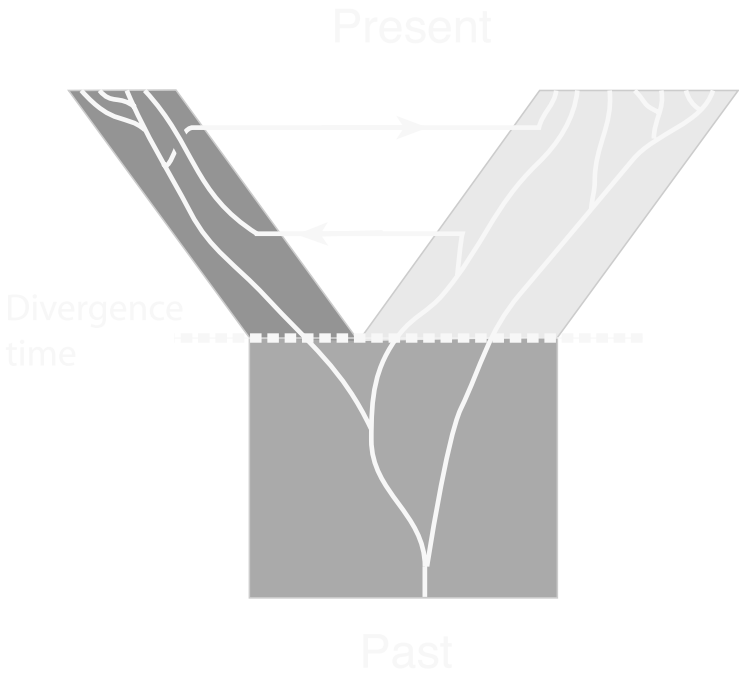
Restricted model: only migration into neighbors allowed



Migration in stable populations

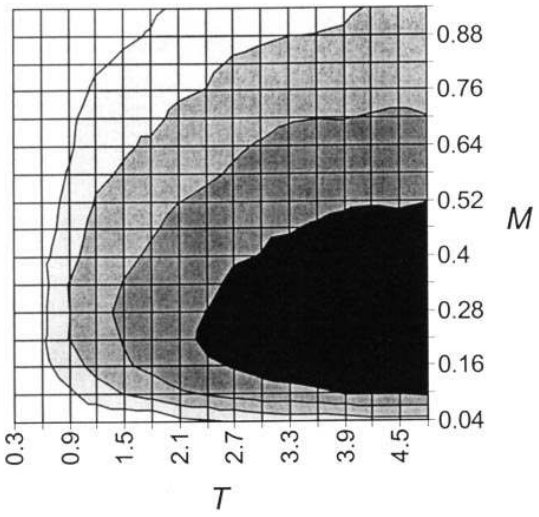
- MIGRATE and LAMARC have similar capabilities
- These programs estimate:
  - Θ per subpopulation
  - Immigration from each subpopulation into each of the others
  - Can use a restricted migration matrix
- Assumptions: no selection, stable population structure
- Unlimited populations in theory but huge data sets needed for more than 2-3 populations
- MIGRATE offers migration skyline plots for additional detail

Migration with divergence



Migration with divergence

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.



Migration with divergence

- IM/IMa/IMa2 and LAMARC estimate:
  - Θ per subpopulation
  - Θ of ancestral populations
  - Immigration in each direction
  - Times of divergence
- IM and LAMARC can also estimate growth or shrinkage of daughter populations
- LAMARC model is similar to IMa2 with recombination, but slower
- Assumptions: no selection, single time point; IM programs no recombination; IM/IMa 2 populations, IMA2/LAMARC up to 10

General issues with migration estimation

- Migration too high–can't infer divergence time
- Divergence too recent–can't infer migration rate
- All methods assume constant migration rate, not bursts (skyline plots in MIGRATE can help here)
- Multiple loci needed for good power

Specific issues with genealogy sampler migration estimation

- Multiple unlinked loci not available
  - LAMARC may help here by using partially linked loci
- Population structure changing too fast
  - For MIGRATE population structure must be stable
  - Population divergence masquerades as excess migration
  - IM programs and LAMARC can handle divergence
- Think of "migration" broadly:
  - Movement between geographic regions
  - Movement between host types
  - Movement between partitions within a patient

Migration skyline plots

- MIGRATE can produce skyline plot of migration events over time
- Non-quantitative way to detect divergence
- Can also show other violations of homogeneity

When to use which model?

- Static migration model (MIGRATE, LAMARC)
  - Populations stable for approximately $4N_e$ generations
  - As many populations as your data can stand
- Divergence model with migration (IM programs, LAMARC)
  - Populations arose less than $4N_e$ generations ago
  - Up to 10 populations
  - Relationships among populations known (if more than 2)

Nested clade analysis

NPA or NPCA (Templeton et al. 1995)

- Infers assorted forces based on shape of haplotype network
- Limited support for recombination based on multiple alternative networks
- In simulation studies with simple scenarios, fairly good recovery of real effects
- However, up to 75% false positives (claims of effects that are not there)

Nested clade analysis

An opinion statement:

- It's often true that if force A is in effect, symptom B results
- NPCA is fairly good at identifying these
- But–how often does symptom B occur WITHOUT force A?
- Massive false positives arise if this is ignored

Nested clade analysis

An opinion statement:

- NPCA promises more than **any** method can now deliver
- Biologists want what it offers, so they use it
- 88% of NCPA results based on one locus; NOT ENOUGH
- Any result from NPCA needs validation by another method
- That other method **may not exist;** this is a weakness in the field (and grounds for more research)

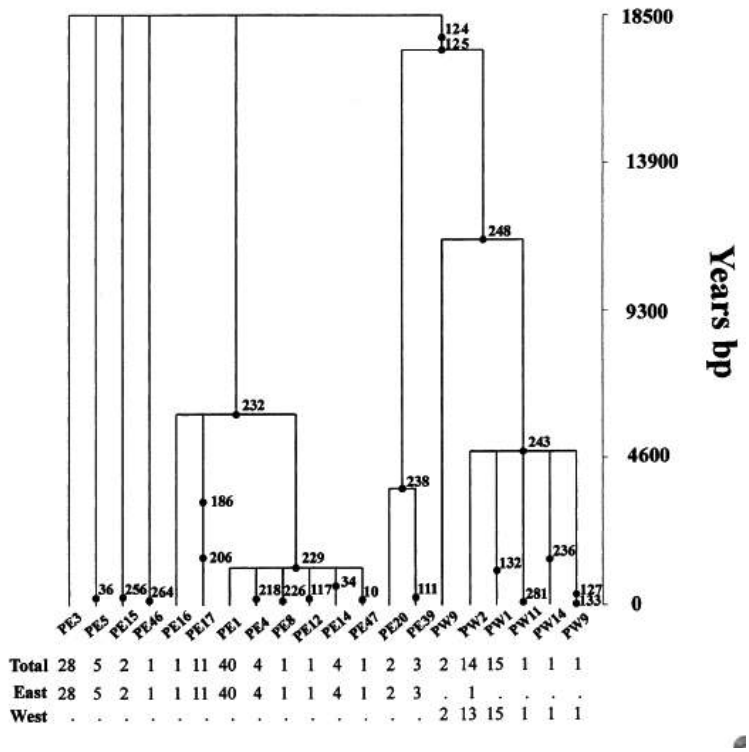
Nested clade analysis

"As far as using a single locus for inferring historical demography, it's fine to argue that it's unacceptable (and anyone with any knowledge of coalescent theory would agree), but again, as a practical matter, unless you are working on a fairly well-studied group, you are often stuck. Mitochondrial loci are easy to amplify and variable enough for intraspecific questions (too bad they are all linked), while primers for sufficiently variable nuclear loci are often unavailable without a lot of pilot work (and even if you do attempt to develop your own primers, you could come up with nothing useful)." – blog post comment 2008

- The argument "It's all I've got so I'll use it" creates noise and self-deception in the scientific literature
- If there isn't enough data to answer a question, that question remains unanswered–and peer reviewers need to enforce this

Times of significant events

Milot et al. (2000)

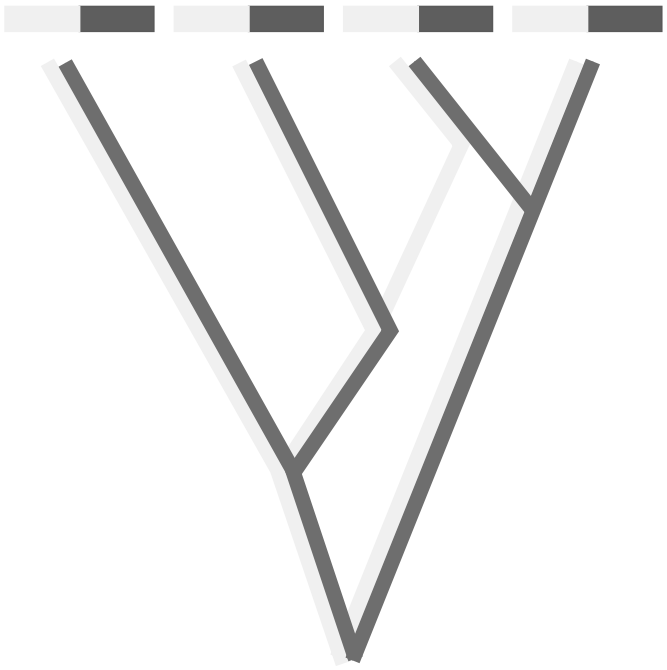


43

Specific issues for times of significant events

- GENETREE can estimate the times of specific mutations or coalescences
- Always pay attention to the error bars of these estimates
- Assumes infinite-sites model:
  - No multiple hits to the same site
  - No back mutation
  - Thus, Theta should be low (human SNPs); won't work on HIV
- BEAST is an alternative:
  - can infer times of coalescences, but not times of mutations
  - No problem with high Theta

Recombination rate estimation



Recombination rate estimation

- LAMARC and InferRho estimate per-site recombination rate
- Assumptions:
  - No changes in recombination rate over time
  - Single time point
  - LAMARC assumes equal rates everywhere; InferRho has hotspots and gene conversion
- Accurate (but slow) for high recombination rate
- Some difficulties with low recombination rate
- With LAMARC, even if recombination rate not of interest, including recombination can improve estimation of other forces

Specific issues for recombination estimation

- Genealogy samplers estimate a rate; they don't identify individual recombinants
- Complimentary to methods such as bootscanning
- If recombinations occur in bursts (during a viral co-infection, for example), rate will be overestimated

Gene tree/species tree analysis

- *BEAST ("star-beast") infers the species tree by coalescent sampling of gene trees for multiple unlinked loci
- Promising use of the coalescent on the population/species boundary
- This should be more reliable than straight species tree inference:
  - Genes may coalesce far earlier than species split ("ancestral polymorphism") and may have a different topology than the species tree
  - The amount of ancestral polymorphism depends on population size, which can vary a lot in the species tree
- This method is bleeding-edge currently

Getting ready for the practical

- In the practical we will run a genealogy sampler
- Questions to bear in mind:
  - What are we trying to find out? How is our answer parameterized?
  - What assumptions are we making?
  - How do we know if we've run long enough?
  - How could this result be validated?

Getting ready for the practical

- Running the genealogy sampler LAMARC
- Inference of: Θ, growth rate, migration rates
- Short description of sampler follows

Likelihood version: Driving value

- To sample trees, we need a distribution
- We don't know the true distribution, so we assume one
- The assumed parameter is called the *driving value*

Driving value

- This approach is only asymptotically correct
- For finite sample sizes, it has a bias toward its driving value
- We can greatly reduce this:
  - Start with an arbitrary Θ_0
  - Run the sampler a while and estimate the best Θ
  - It will be biased toward Θ_0, but...
  - Use it as the new Θ_0 and start over

How this works in practice

- Run multiple consecutive "initial chains"
- These allow the sampler to get good starting values
- When the values have settled down, run one "final chain"
- This will produce the final estimate
- Behavior of the initial chains is a clue:
  - Did we run long enough?
  - Do we need more chains or longer chains?
  - How long is this going to take, anyway?

Bayesian version: Priors

- Alternatively, we can use a Bayesian algorithm
- We place priors on each parameter
- New values are sampled from the prior
- We tabulate accepted values and form a curve
- Multiple cycles are not necessary

Bayes versus Likelihood

- Overall performance of likelihood and Bayesian samplers is similar
  - (Beerli 2006, Kuhner 2007)
  - Bayesian has edge when data are sparse
  - Likelihood gives more information about correlation of parameters

How this works in practice

- A Bayesian run is usually just one "final chain"
- (You could do an initial chain to get a time estimate)
- Starting values are not so important
- Good priors are essential!

Bayesian priors

- A prior should represent our pre-existing knowledge of a parameter
- Often biologists cannot quantify this
- Instead, we use "non-informative" priors and hope it won't matter
- No prior is really uninformative:
  - If it is too narrow, it may exclude the truth and will certainly overstate confidence
  - If it is too wide it will drastically slow the search and may understate confidence
  - If it is not roughly centered around the truth it may introduce bias
- To know the "right" prior we would need to know the answer....

Bayesian priors

Advice:

- Use all available information to set the priors:
  - Previous studies
  - Analogies with similar systems
- Give a little fudge room but do NOT be extremely conservative
- None of the samplers work well with extremely wide priors (they don't focus the search)
- If your results are piled up on one side of the prior, widen the prior next time
  - This is totally inappropriate in Bayesian theory
  - None the less, it improves results....

General advice

- Don't be afraid to try things
- Interrupt any run that will take too long for a 90 minute practical
- After completing one run, move or rename the results file or it may be overwritten
- Draw a picture of your results
- Ask questions
- Talk to your neighbors, TAs and instructors

Preparation for practicals

- Data files for demonstration can be downloaded from:
  - http://evolution.gs.washington.edu/lamarc/sisg-2014/demo/
- Please download these before the demonstration
- This will save time and pain during the demo. Thank you!

Lecture 5:
Selection, gene genealogies and medical genomics

Molecular population genetics and tests for selection

•Population genetics is concerned with estimating the amount of genetic variation in populations and explaining what evolutionary forces maintain the variation.

•In the last two decades, we have now managed to gather data on genetic variation within species at the sequence level, in some cases spanning entire genomes.

•Inferences of non-neutral processes can be made using parametric or non-parametric approaches.

These approaches can either use a model to estimate parameters (e.g. the population selection coefficient) or use the genome as ‘our null’ to test for departures from neutraliy.

In this lecture we will…

- Describe the various deterministic forces shaping variation within populations
- Show how these forces affect the shape of genealogies within and among populations
- Calculate expectations and variances of key genealogical properties
- Use these expectations to make inferences of selection using parametric and non-parametric approaches

The fitness effects of variation in populations…

- The **neutral mutation** hypothesis assumes that most if not all mutations have no fitness consequence.

- Those mutations that are deleterious will largely be kept at low frequency in the population, or are lost.
  - regardless, the most common polymorphisms found in the population are selectively neutral.

- Therefore, the dynamics of neutral, or nearly-neutral, mutations are governed largely by random genetic drift.

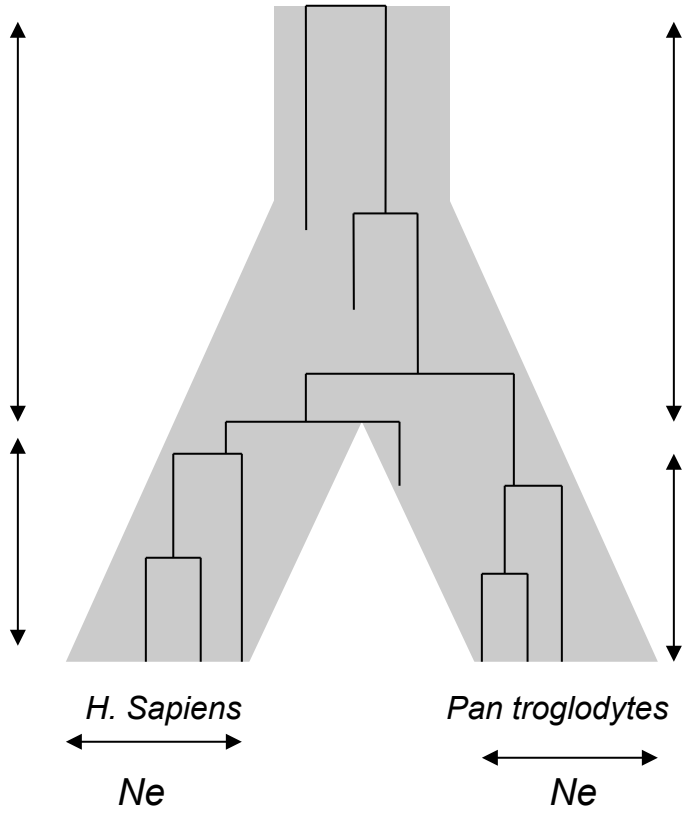
The fitness effects of variation in populations (cont.) …

According to the neutral mutation hypothesis, both the variation *within* and differences *between* populations (or even species) are mainly due to neutral mutations.

Variation is a transient phase of molecular evolution.

Rate of molecular change between populations (or species) should be positively correlated with levels of within-population (or species) variation.

Several tests have been developed to explore whether this prediction of within-species variation is indeed positively correlated with between-species evolution.



Rates of Fixation (under neutrality)

Rate at which a new neutral mutation is fixed in population is equal to neutral mutation rate.

K = rate of fixation of alleles

= (rate of allele formation)(fixation probability)

= (2N gametes)(mutation rate/gamete)(1/2N)

$K = \mu$

Rates of fixation under selection

$$K = (2N)(\mu)(2s)$$

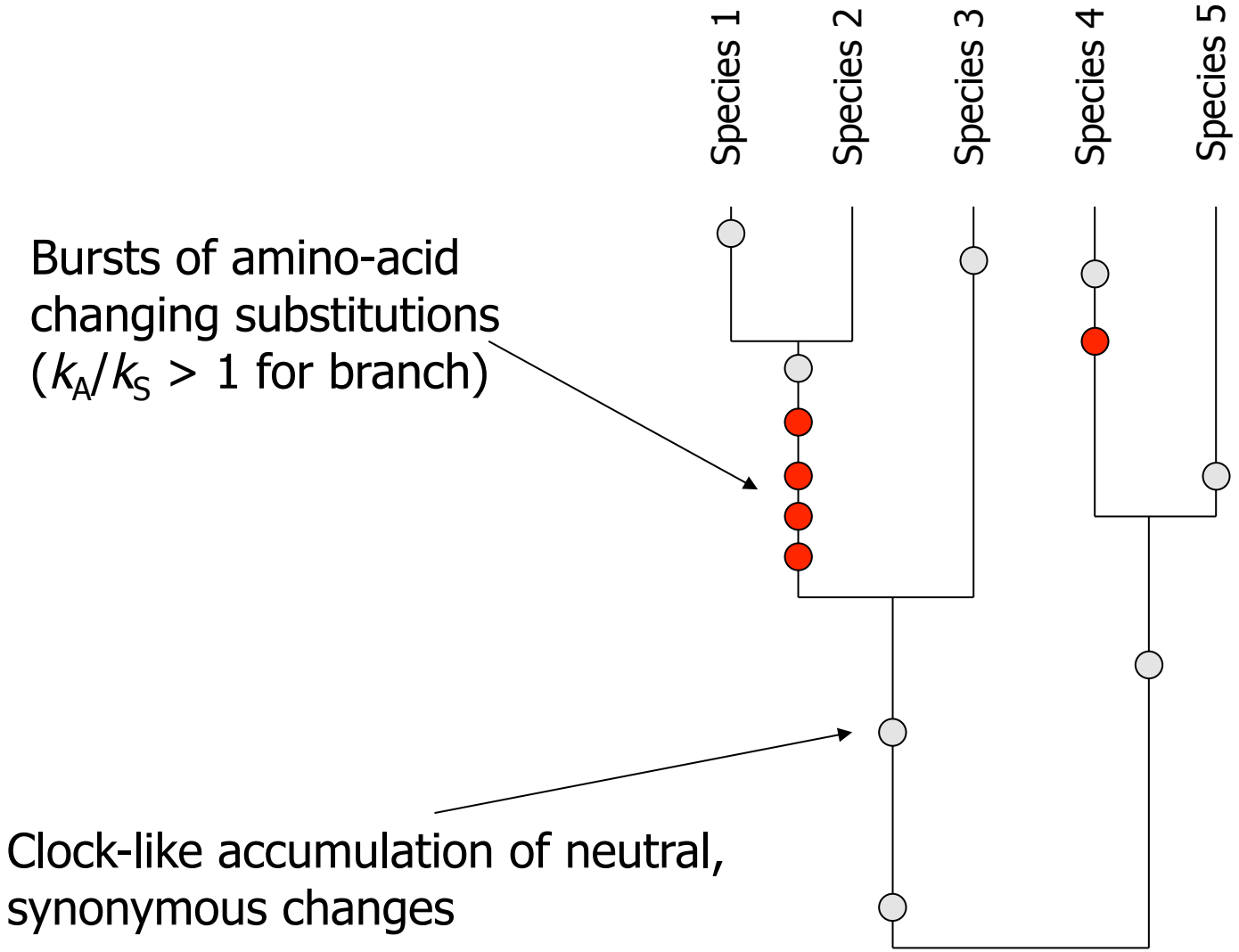
$$= 4N\mu s$$

Thus, under selection ($|Ns| > 1$ or < 1) the rate of gene substitution is either greater, or less, than under neutrality.

- under neutrality, since rate of gene substitution is equal to μ, the average time between consecutive fixations is $1/\mu$.

-higher the mutation rate, the smaller the time between fixations

Phylogenetic methods use this information to infer selection among species



Inferring selection from inter-species data …

Consider first coding regions. When comparing different species, you may have divergence in non-synonymous sites (K_A) or in synonymous sites (K_S).

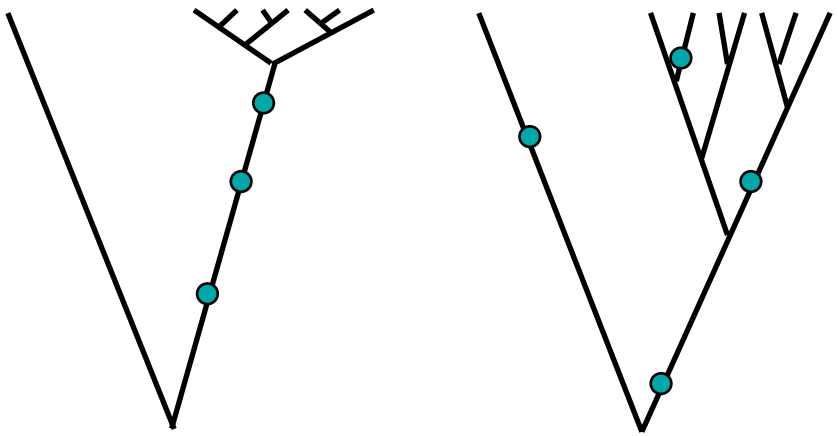
A ratio of these measures can be interpreted as

$K_A/K_S << 1$	Purifying selection
$K_A/K_S = 1$	neutrality
$K_A/K_S > 1$	Positive selection

Phylogenetic methods vs population genetics

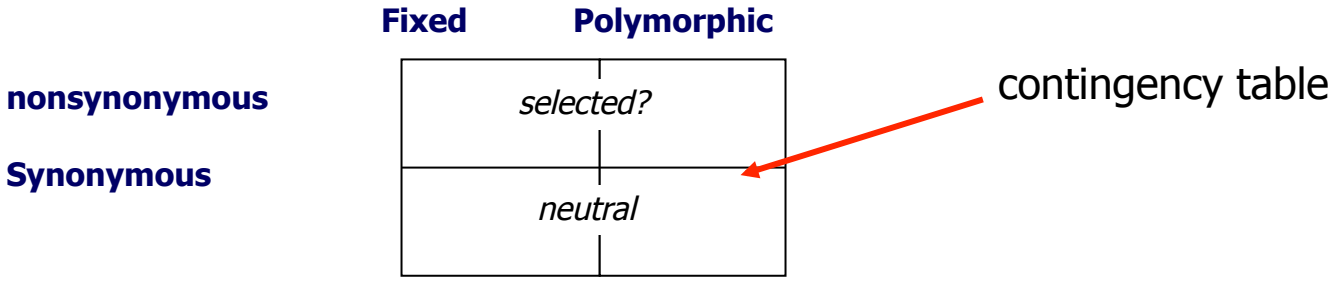
- Most adaptive evolution likely consists of a single change at a particular location – recurrent advantageous mutations only likely in situations of coevolutionary antagonism.

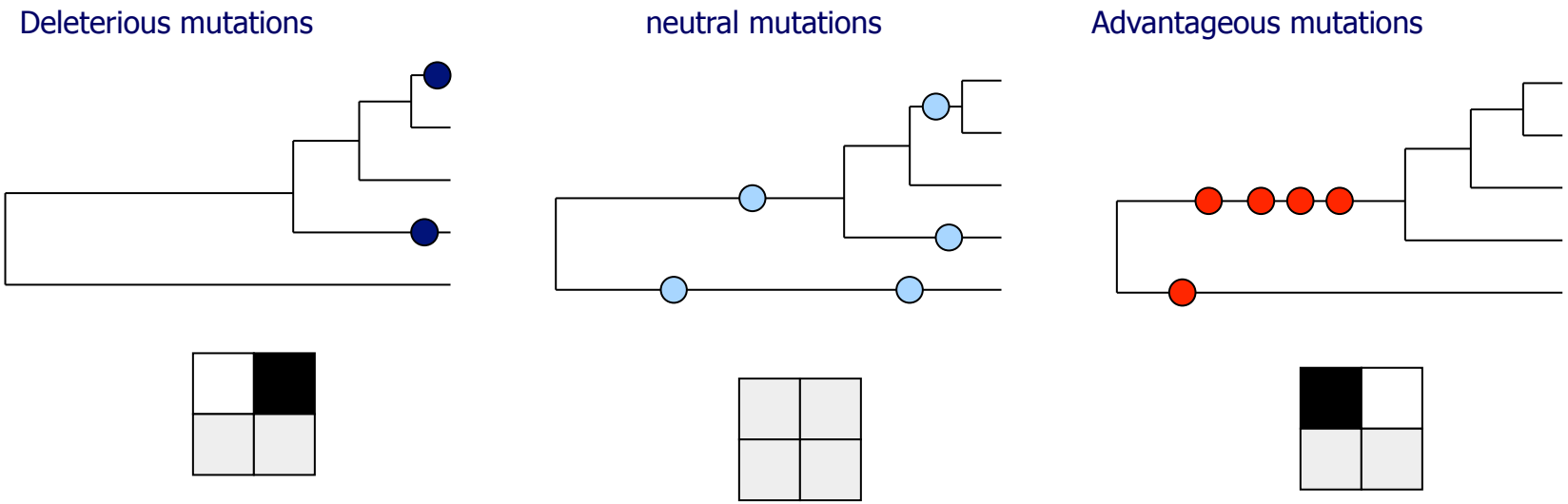
- Patterns of genetic diversity *within populations* reveal signature of more **recent** selection events.



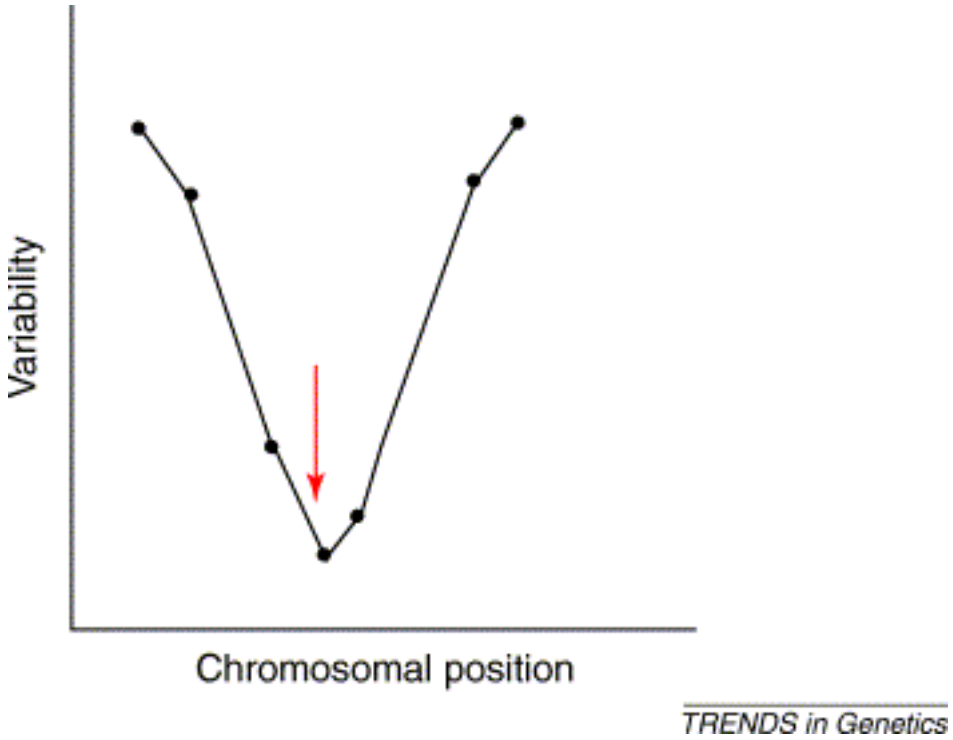
The McDonald-Kreitman test (1991)

- Compare patterns of polymorphism and divergence at different classes of interspersed mutations





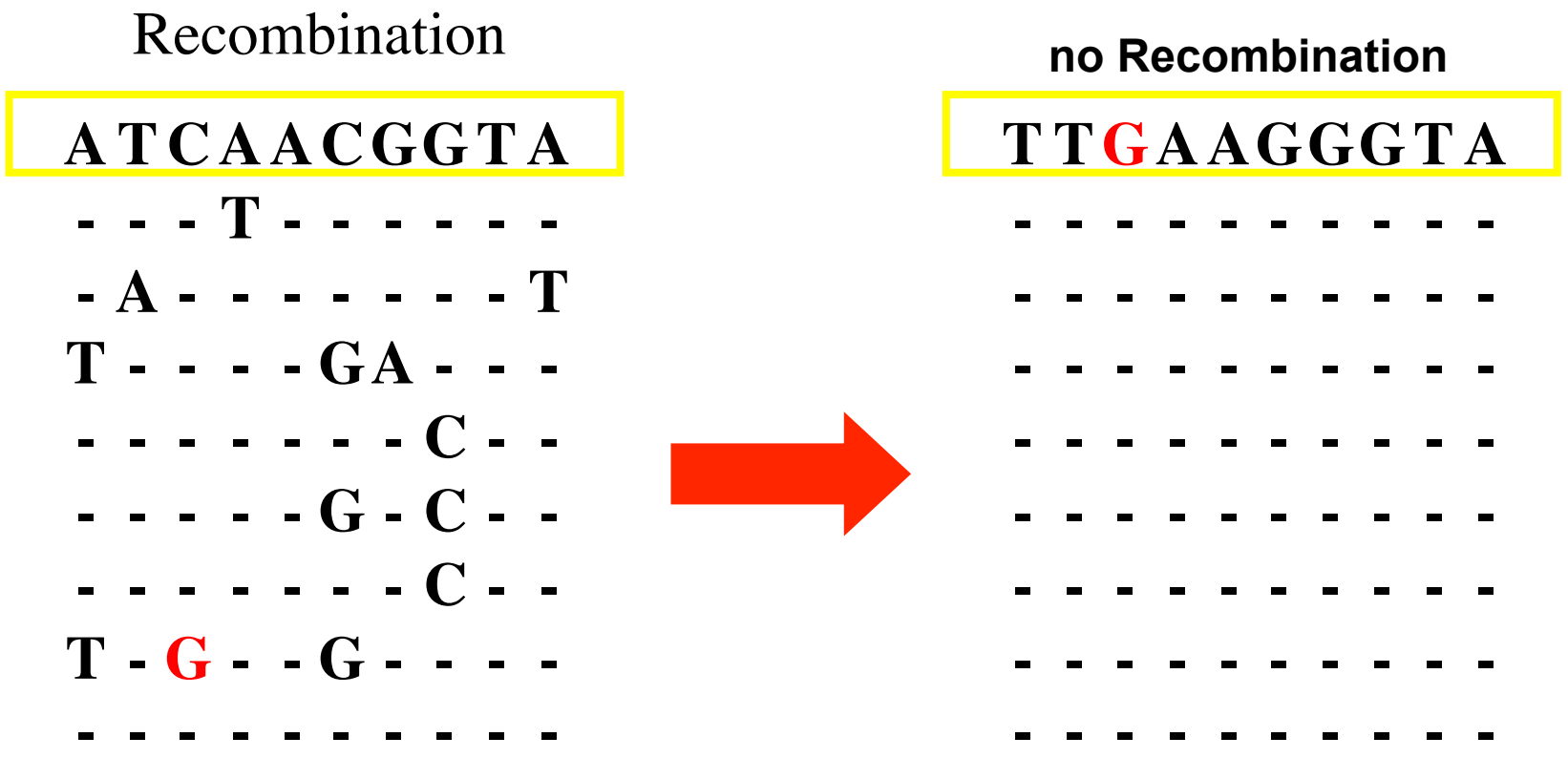
Selective sweeps and patterns of genetic diversity



Expected allele frequency distortion due to a selective sweep.

The distortion in allele frequency is measured by a reduction in variability.

Positive selection produces “Selective Sweeps”

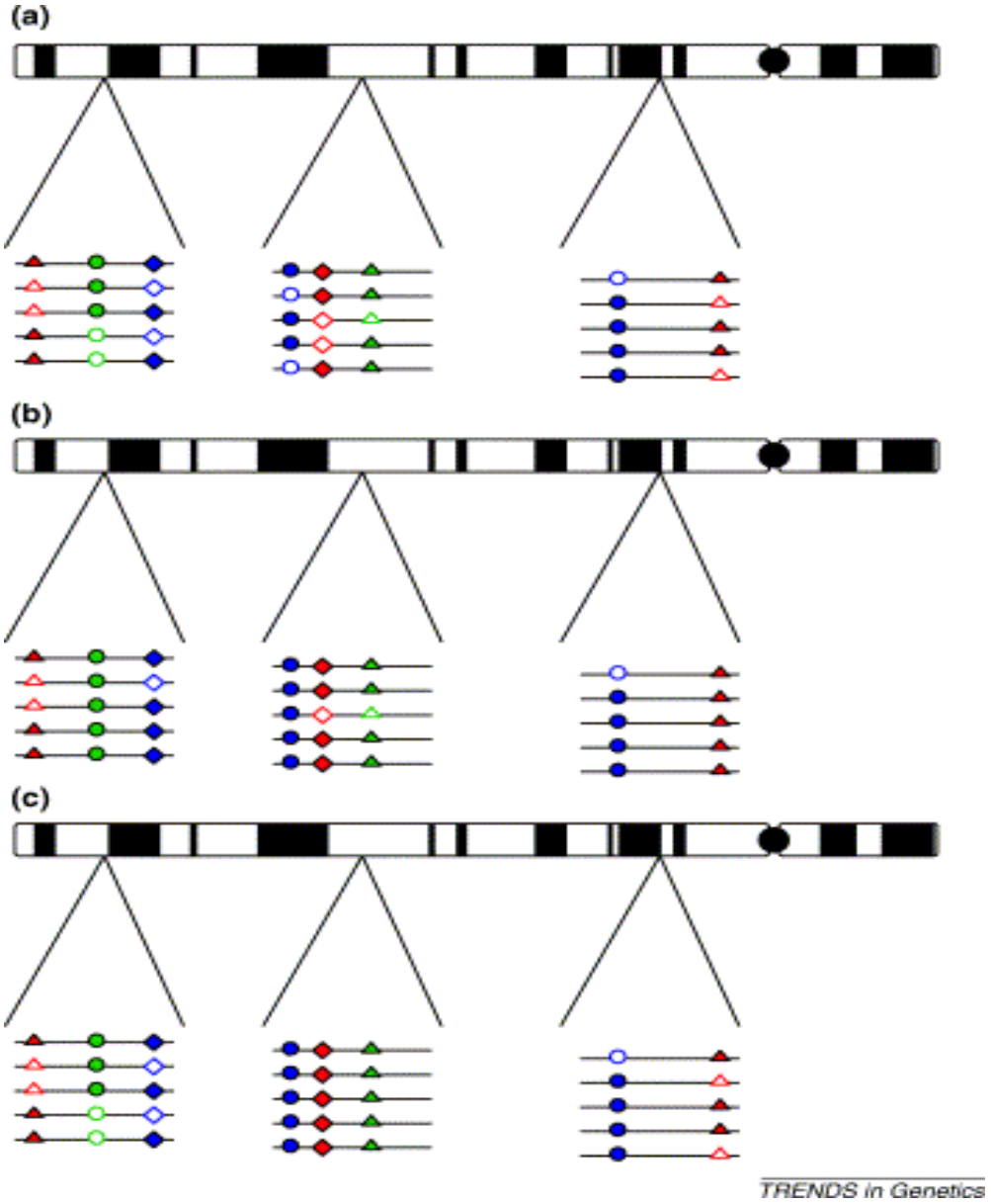


‘Local’ changes in diversity vs. ‘genome-wide’…

neutral

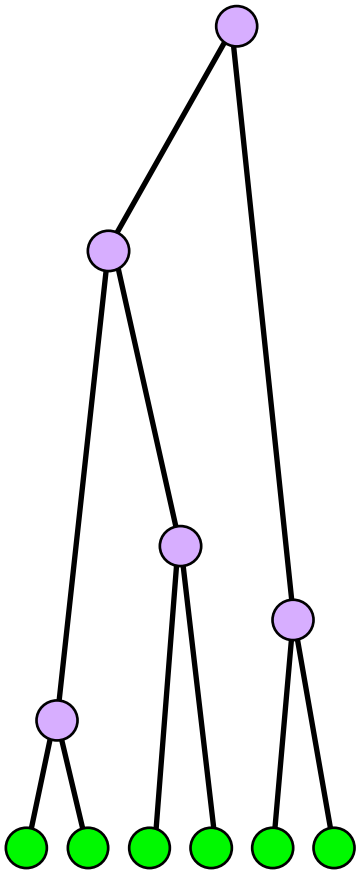
Population bottleneck

Selective sweep

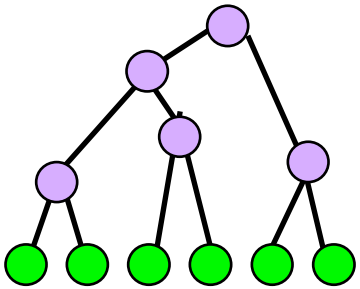


Positive Selection changes the topology of the tree

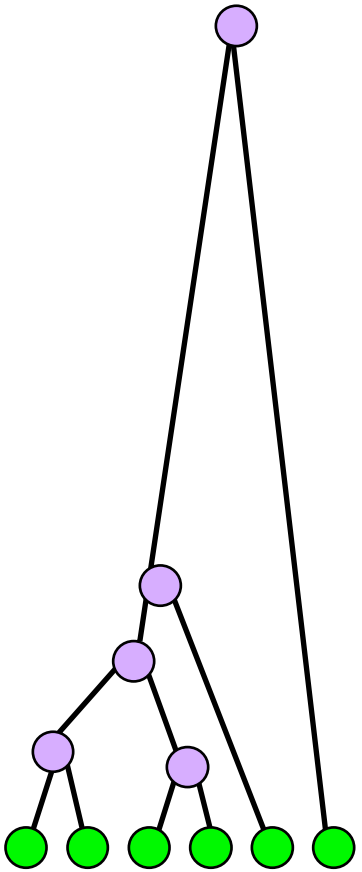
neutral Tree



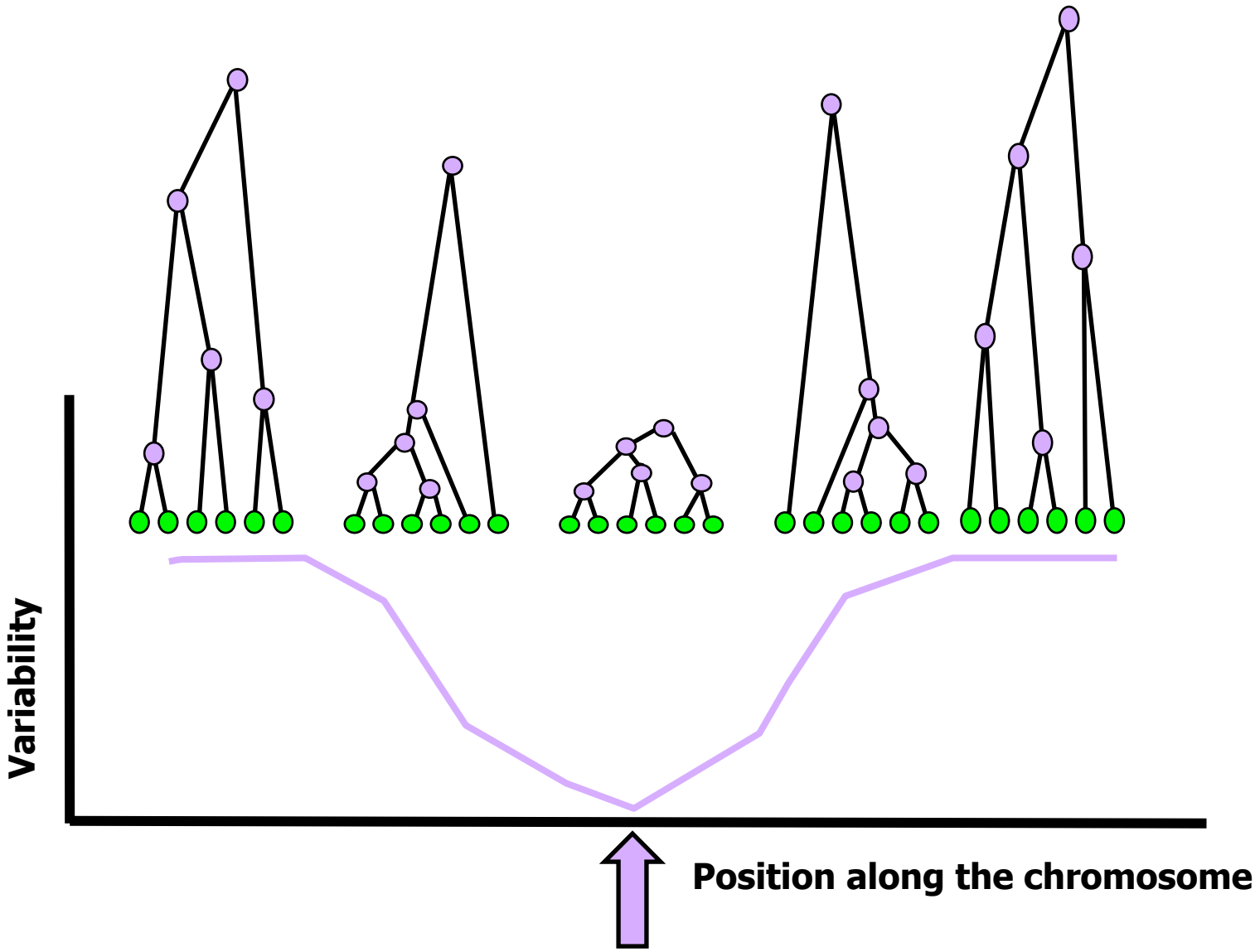
Selective sweep
(no recombination)

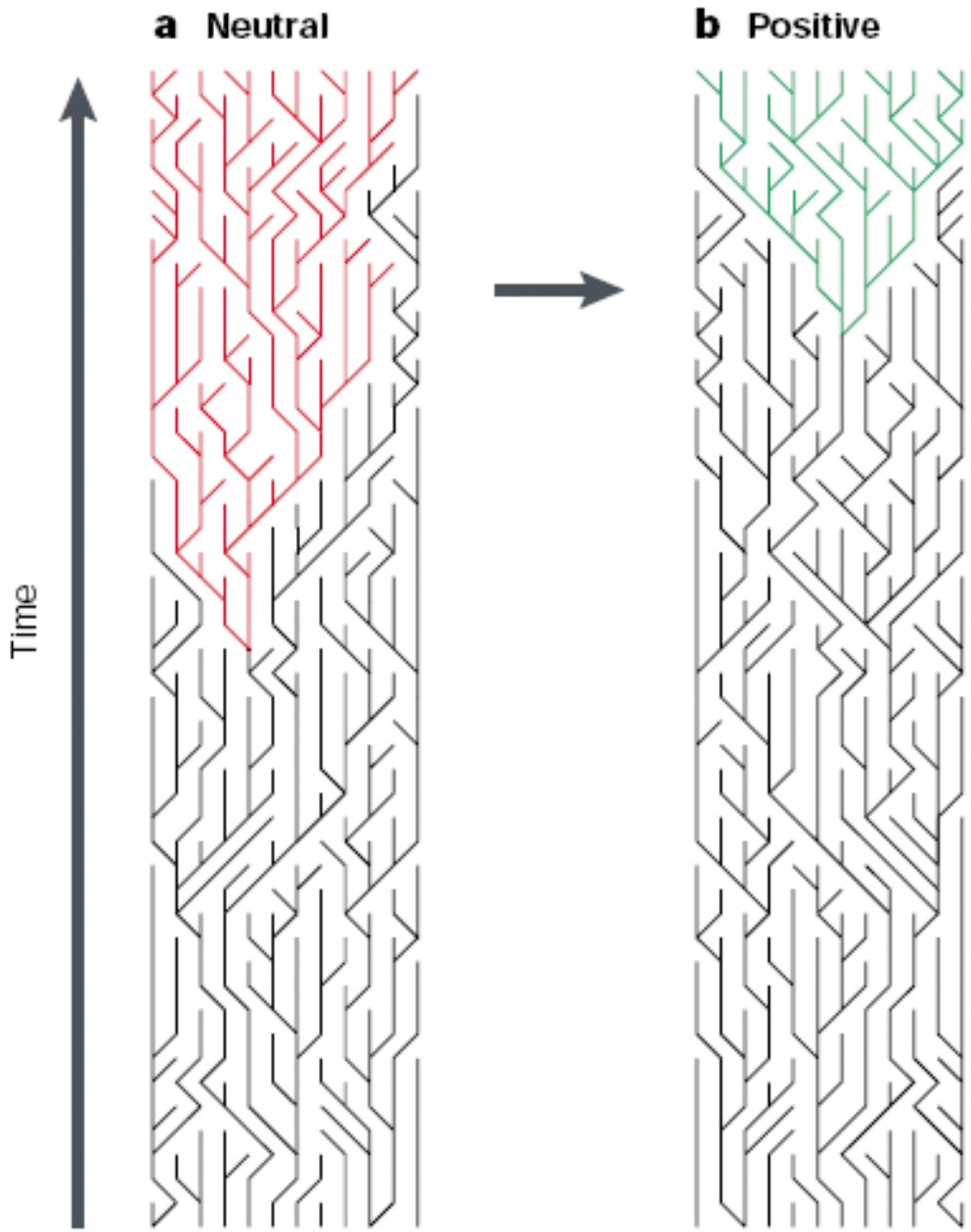


Selective sweep
(some recombination)

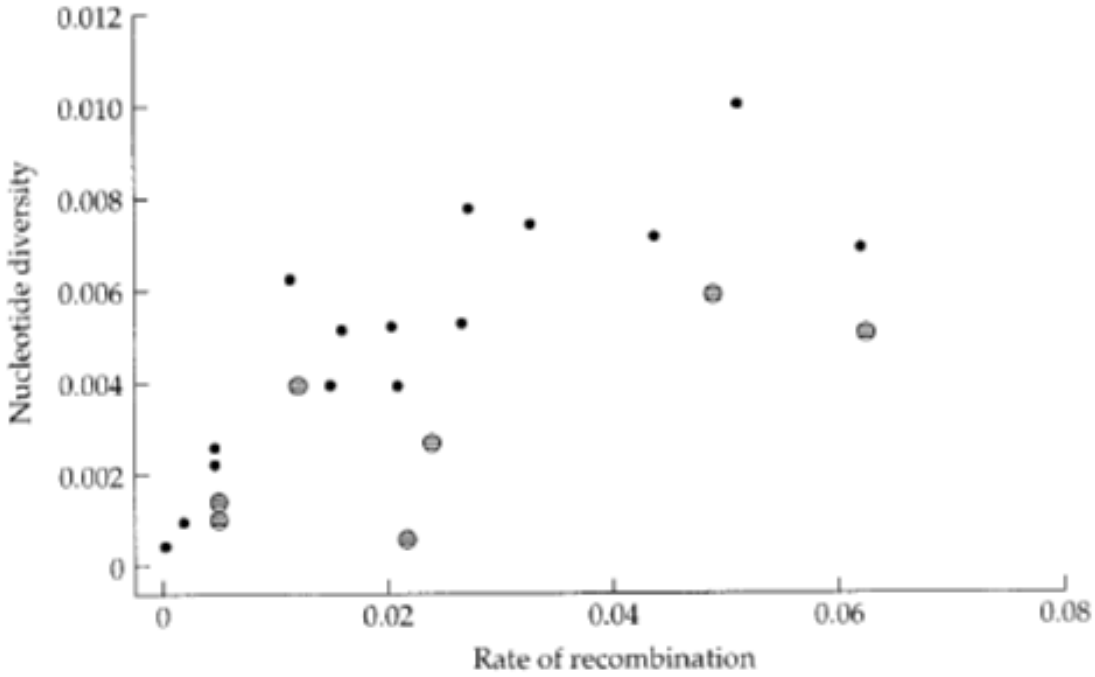


Positive Selection changes the topology of the tree




a Neutral
b Positive
Time

Pervasiveness of ‘sweeps’ across genomes…



Begun et al. 1991

The presence of selective sweeps may explain a well-documented generalization in molecular population genetics that nucleotide diversity is positively correlated with recombination rate in the genome.

Effects of a selective sweep…

Effect	Description
Reduced variability	Because of the replacement of other alleles by the selected one
Allele frequency distribution	Immediately after the sweep a surplus of higher frequency dervied alleles is generated. With increasing time since the selective sweep, new mutations are generated, which results in a surplus of low frequency alleles
Linkage disequilibrium	Selection increases linkage disequilibrium around the selected site.

Detecting recent selection events - Tajima's *D*

$\prod$ = Pairwise nucleotide diversity S = number of segregating sites

$$D = \frac{\Pi - \theta_W}{\sqrt{Var(\Pi - \theta_W)}} \Rightarrow D = \frac{\hat{\theta}_\Pi - \hat{\theta}_S}{\sqrt{Var(\hat{\theta}_\Pi - \hat{\theta}_S)}}$$

Under neutrality, *D* = 0 because

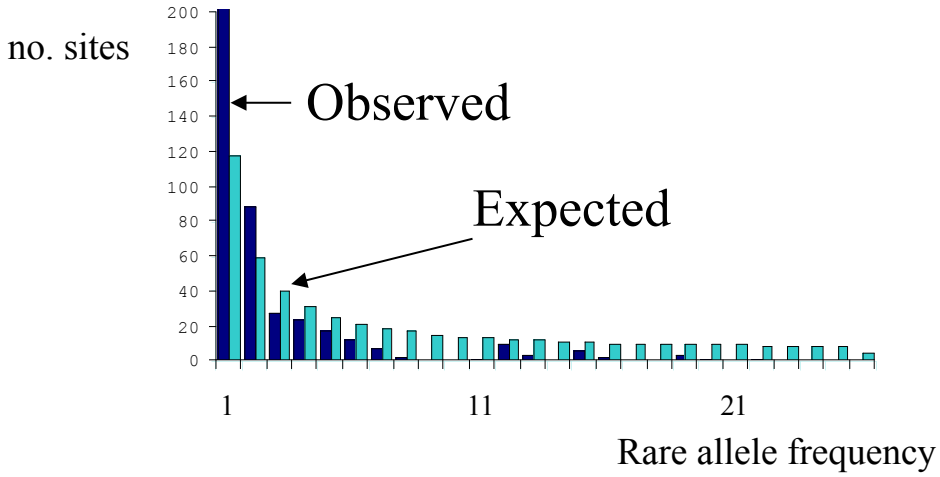
$$\theta_W = \Pi = \theta = 4N_e\mu$$

In case of selection or demographic changes, $\prod$and S are affected in different ways and D can change.

With positive selection or demographic expansion, D is expected to be negative.

Tajima and Fu & Li Tests

The neutral theory predicts what the frequency spectrum of mutations should be under neutrality. Deviation from this expected frequency spectrum signals non-neutral evolution.



1. GAGGTGCAACAG...
2. GAGGACCAACAG...
3. GAGGTGCATCAA...

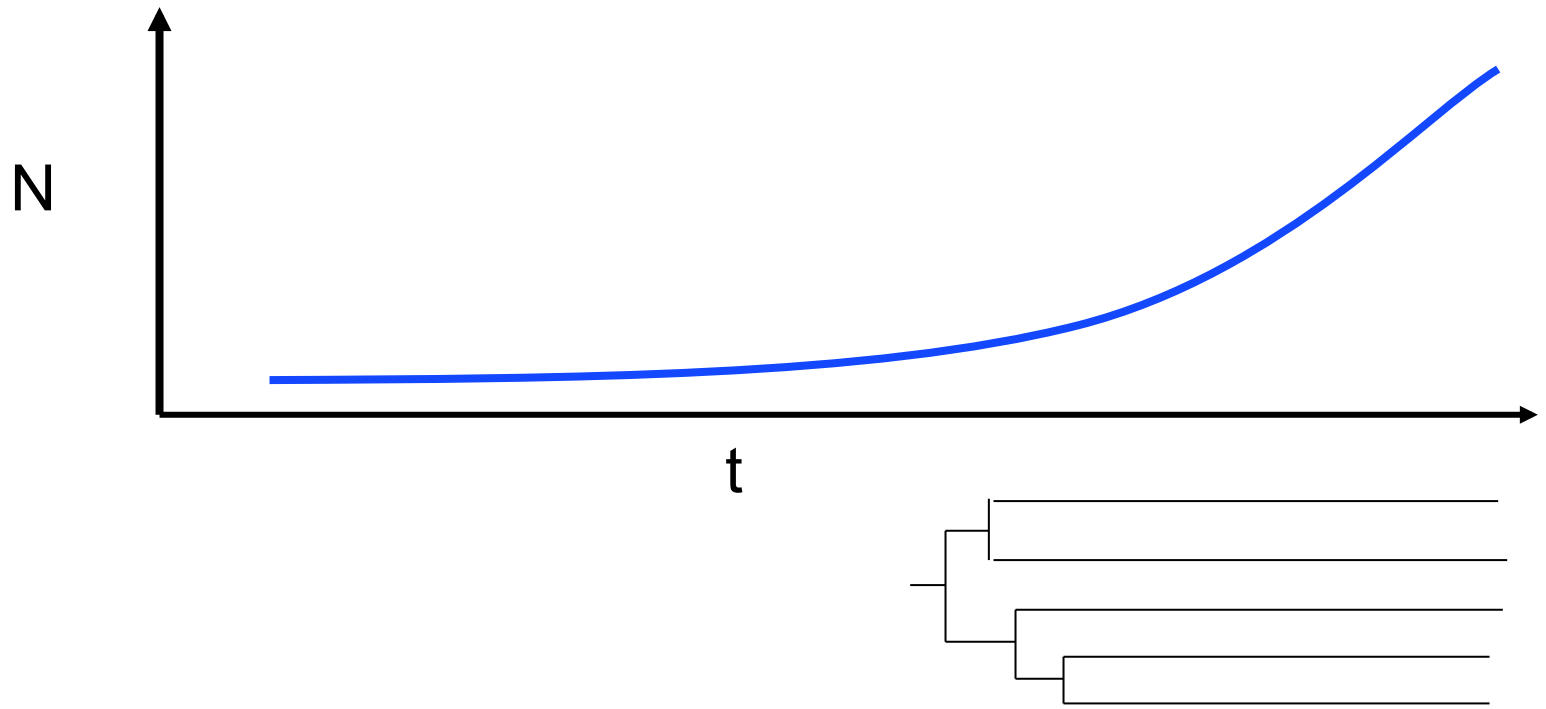
.

.

.

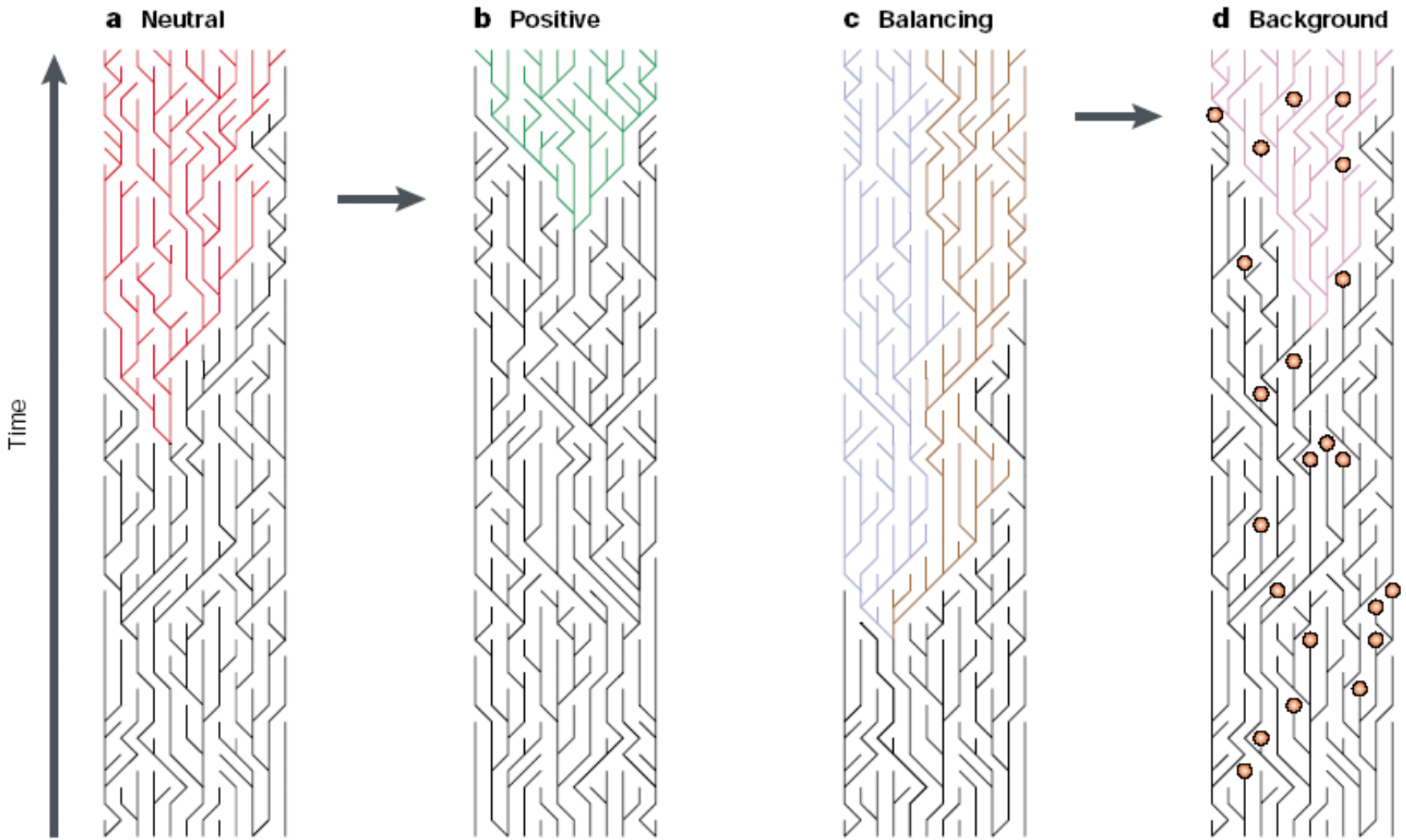
n. GGGGTGGAACAG...

e.g. Population growth and positive selection generate very similar trees

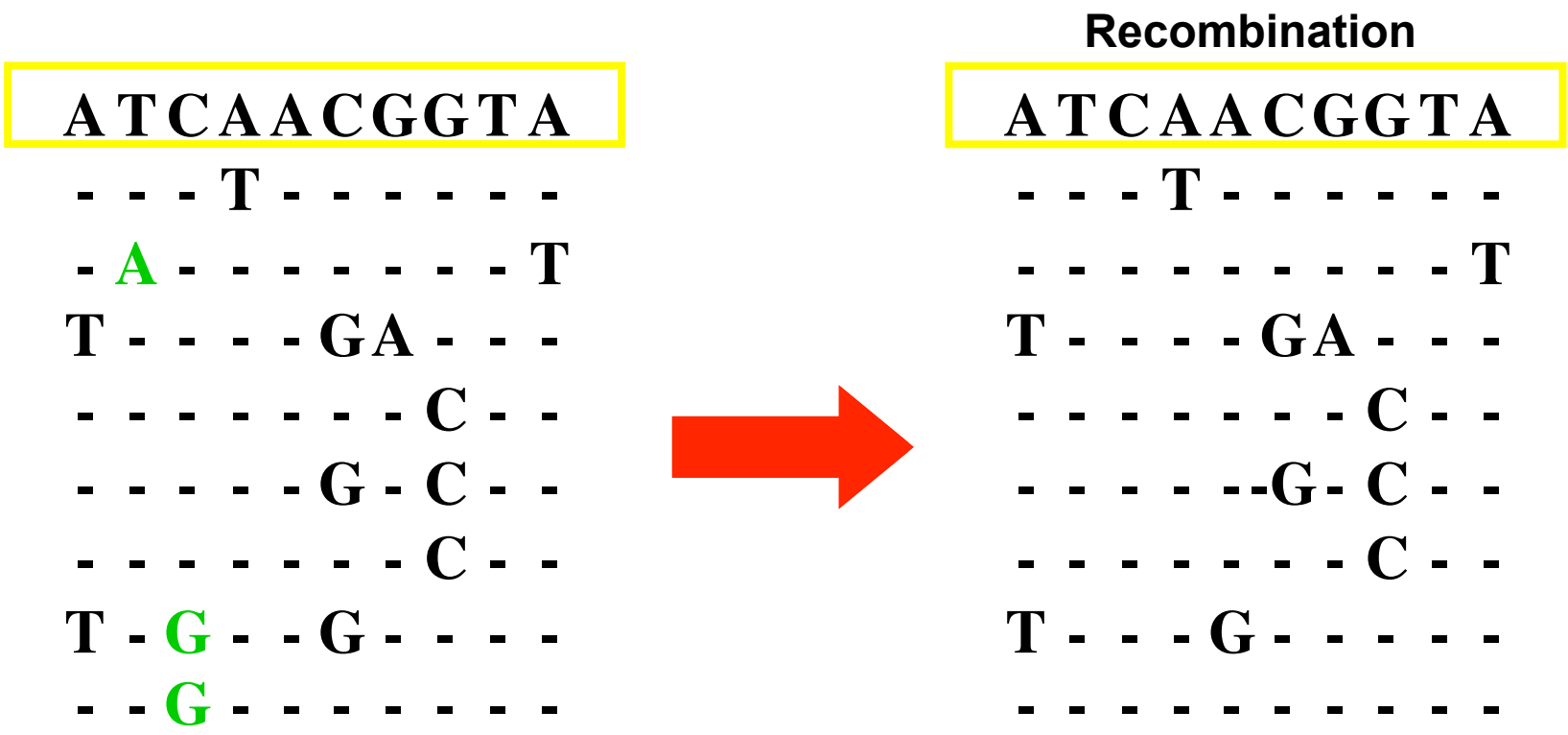


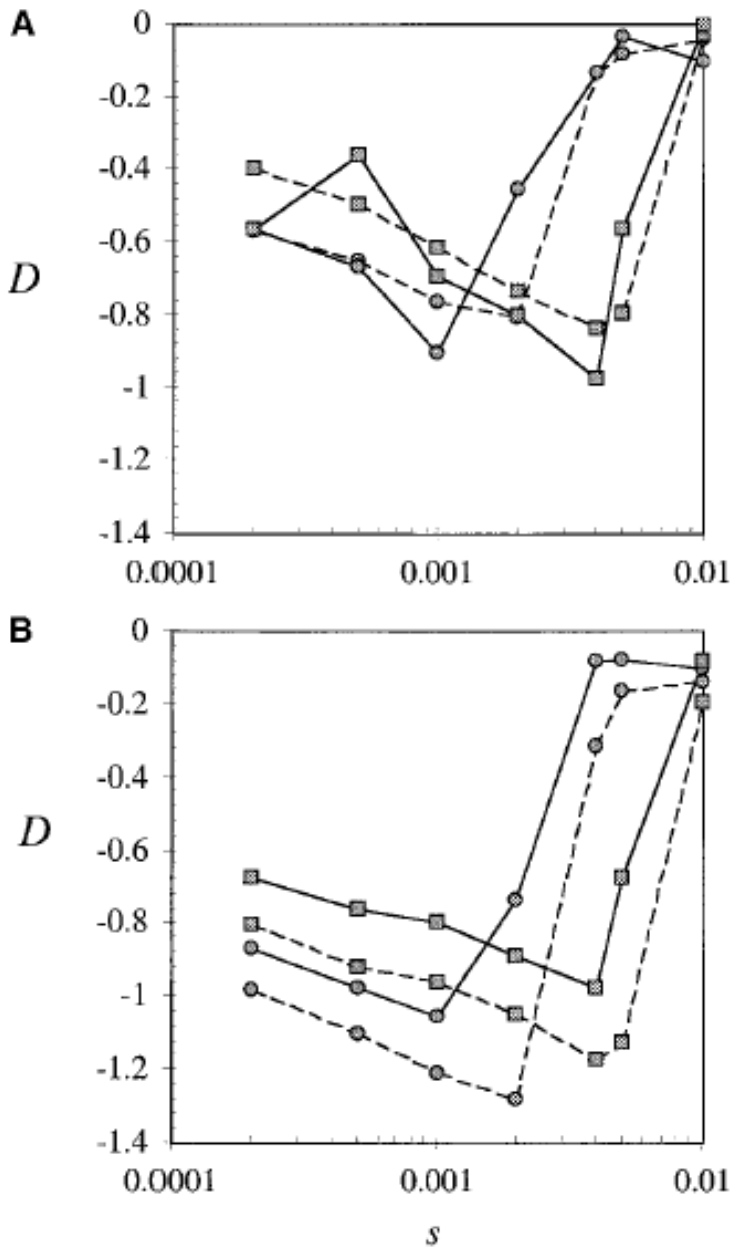
many low-frequency mutations…

Other kinds of selection…



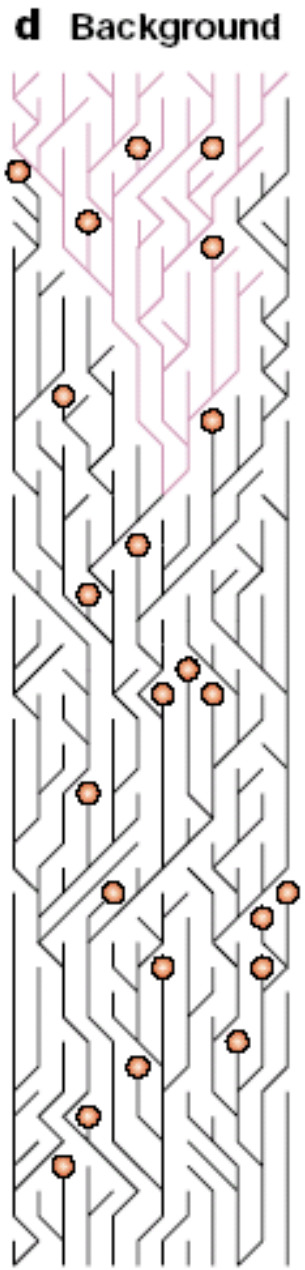
Background (purifying) selection eliminates some linked variability



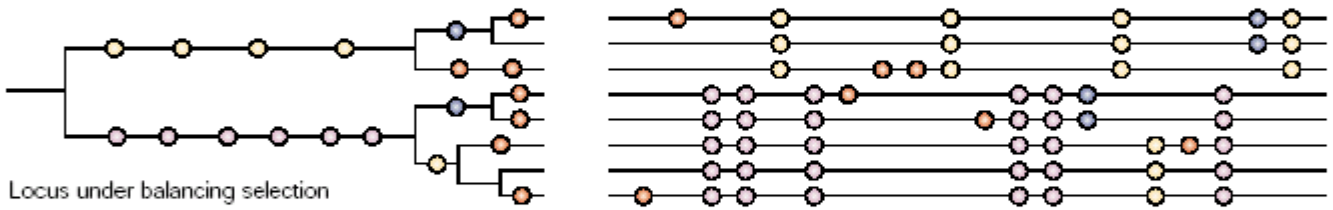


With weak purifying (background) selection and low recombination, Tajima's D is also negative.

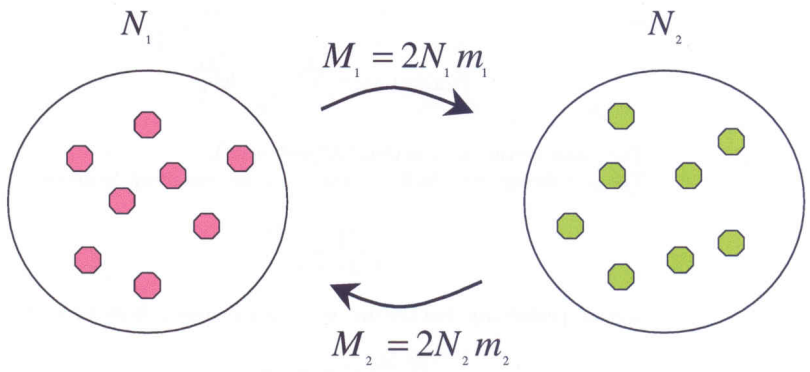
Strong purifying selection has no effect on Tajima's D



Genealogically, balancing selection is similar to demographic subdivision. Tajima's D is positive because there are lots of intermediate frequency variants.



$$D = \frac{\Pi - \theta_W}{\sqrt{Var(\Pi - \theta_W)}}$$

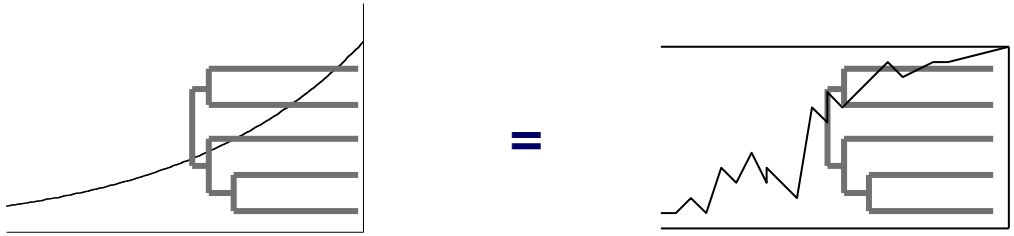


Effects of a balancing selection…

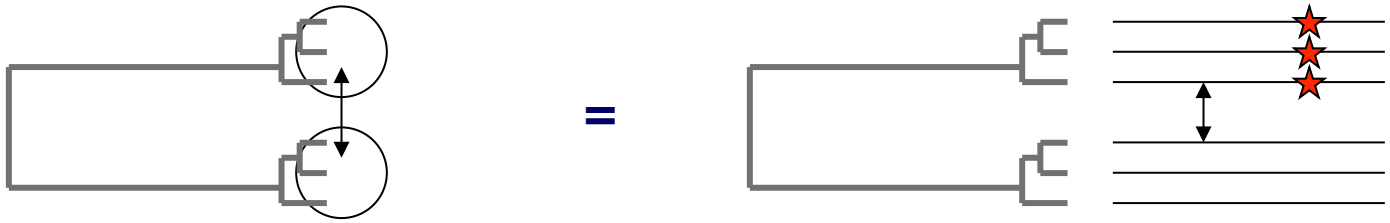
Effect	Description
Increased variability	Per site/marker frequencies are maintained in a “balance”
Allele frequency distribution	More intermediate allele frequencies
Linkage disequilibrium	Selection increases linkage disequilibrium around the selected site.

Demographic processes can mimic selection…

Population growth can look like selective sweeps



Population subdivision can look like balanced selection



Differences between loci can distinguish between genome-wide effects and the local effect of natural selection

BUT need to know about variance of demographic processes……

How do we assess if Tajima's D is significant?

We use a goodness-of-fit approach

We run coalescent simulations under neutrality, using the same n and S than in our data. We measure Tajima's D in there and get the distribution.

Is our observed D an extreme deviant?

Test statistics and hypothesis testing

- Let ***H*** be a hypothesis (or statement) about a population parameter
  - E.g. q = 1, or the human population started expanding 10,000 years ago
- Let ***T*** be a statistic of the data
  - Can be any function, but ideally low dimension informative summary (e.g. number of segregating sites, difference between two estimators of q)
- Define a rejection region ***R*** such that the probability of observing a value of ***T*** that lies in ***R*** given that ***H*** is true is equal to the desired rejection probability **a**
  - e.g. given the hypothesis that q = 5 and a sample size of 20 (with no recombination) 95% of observations would have between 6 and 48 segregating sites.
  - In population genetics, rejection regions are often estimated by simulation
- In goodness-of-fit tests ***H*** = *The assumed model is correct*
  - May include statements about parameter values

An example: Tajima's D and human mtDNA

- Ingman *et al*. (2000) 52 complete mtDNA molecules from a worldwide sample (linguistic groups)
- 521 segregating sites excluding D-loop

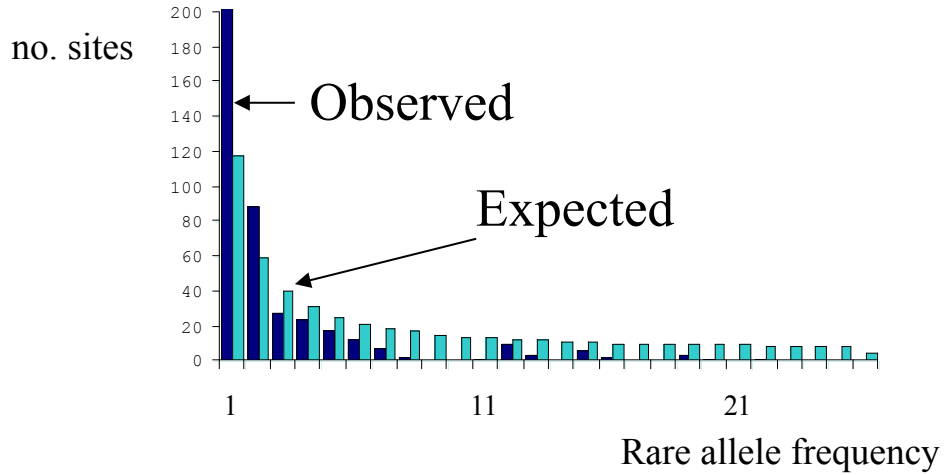
$$\pi = 44.2$$

$$a_{52} = 4.52$$

$$S / a_{52} = 115.3$$

$$\sqrt{\hat{V}(d)} = 31.8$$

$$D = \frac{44.2 - 115.3}{31.8} = -2.23$$



Probability of observing such an extreme value under neutrality = 0.01

Human mtDNA have an excess of low-frequency variants

⟹ Population growth, selection, or sampling?

Coalescent-based (frequentist) tests

- Single-locus tests
  - Tests for differences between summary statistics at a single locus

    Watterson homozygosity test, Tajima *D* test, Fu and Li *D test, Fay and Wu *H* test, Haplotype-based tests**

- Multi-locus tests
  - Tests for heterogeneity between loci, or classes of mutations

    HKA test, variance tests, Lewontin-Krakauer test

The Hudson-Kreitman-Aguade (HKA) Test

This approach tests a prediction of the neutral theory - that levels of within-population variation is positively correlated with between-population divergence.

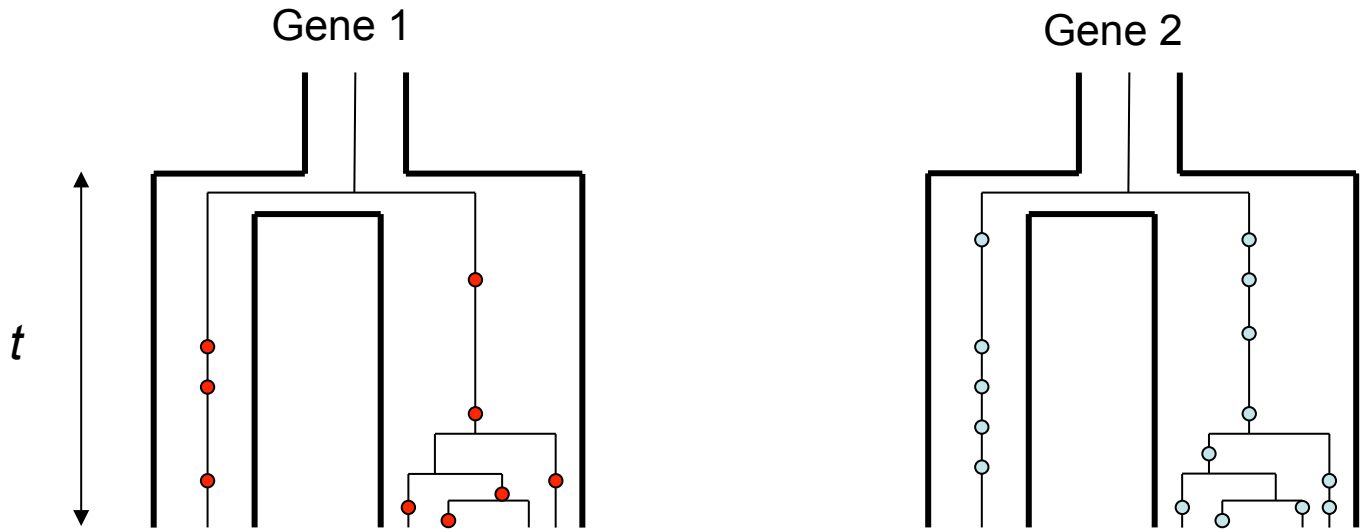
Let us say we have n_1 and n_2 random sequences from populations 1 and 2.

S_1 and S_2 are the number of segregating sites in populations 1 and 2, respectively.

D is the number of differences between a random sequence in population 1 and 2.

The HKA test…

- Hudson, Kreitman and Aguadé (1987)
  - Compare polymorphism and divergence at two or more loci within coalescent framework



$$E[S_1] = \theta_1 a_n \qquad E[d_1] = \tau\theta_1 \qquad E[S_2] = \theta_2 a_n \qquad E[d_2] = \tau\theta_2$$

Estimate parameters and calculate goodness-of-fit test statistic

More formally (HKA cont.)…

Let us say we have m number of genes. We can define these values for each gene *i* (*i* = 1 to m): S_{1i}, S_{2i} and D_i. Usually, we have two genes.

Let us say population 1 has size N_e, and population 2 has size $f N_e$ (where *f* is a factor that relates the two population sizes together).

What do we estimate? The test statistic is

$$\chi^2 = \Sigma_m[S_{1i}\text{-}E(S_{1i})]^2/V(S_{1i}) +$$

$$\Sigma_m[S_{2i}\text{-}E(S_{2i})]^2/V(S_{2i}) +$$

$$\Sigma_m[D_i\text{-}E(D_i)]^2/V(D_i)$$

(HKA cont.)

$$E(S_{1i}) = a_1\Theta_i \qquad V(S_{1i}) = a_1\Theta_i + b_1\Theta_i^2$$

$$E(S_{2i}) = a_2 f\Theta_i \qquad V(S_{2i}) = a_2 f\Theta_i + b_1(f\Theta_i)^2$$

We won't show it here, but the expectations and variances of d_i are

$$E(d_i) = [\tau + (1 + f)/2]\Theta_i$$

$$V(d_i) = E(d_i) + [\Theta_i(1 + f)/2]^2$$

We can use the data to calculate Θ's, f and t under the assumption of neutrality and thus come up with the expectations and variances.

The Goodness of fit test…

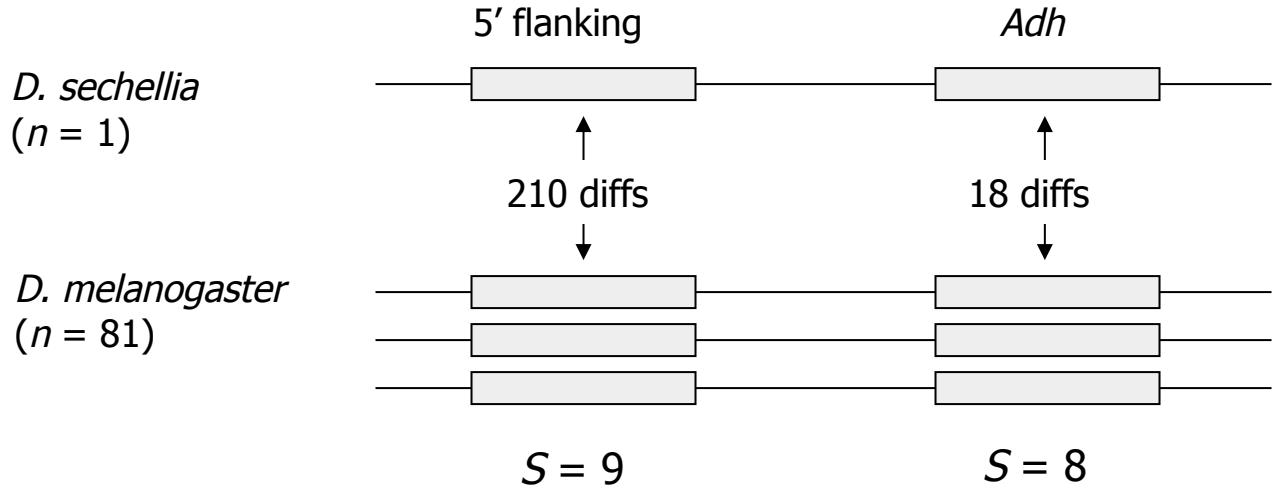
Hudson et al. (1987) showed that the test statistic χ^2 is approximately χ^2 distributed with 2m-2 degrees of freedom. The HKA test is thus a goodness-of-fitness test, and what we need is

- data from at least two genes
- within population data for at least one population
- divergence data for the two genes between two populations

From this, we

- measure S_{1i}, S_{2i} and D_i for the two genes
- estimate Θ's, f and τ under the assumption of neutrality.
- calculate χ^2.

Adh in *Drosophila*



Solving the simultaneous equations

$$\hat{\tau} = 13.4 N_e \text{ generations}$$

$$\hat{\theta}_1 = 2.7$$

$$\hat{\theta}_2 = 0.7$$

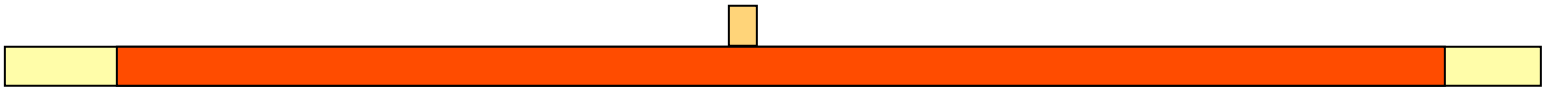
$$\chi^2 = 6.09 \quad P = 0.016$$

- Fast / slow polymorphism in exon 4 associated with a two-fold difference in enzyme activity
- Cline in polymorphism: Fast more common in northern America and at higher altitudes

A Haplotype based test (Pardis et al. 2002)…

no selection

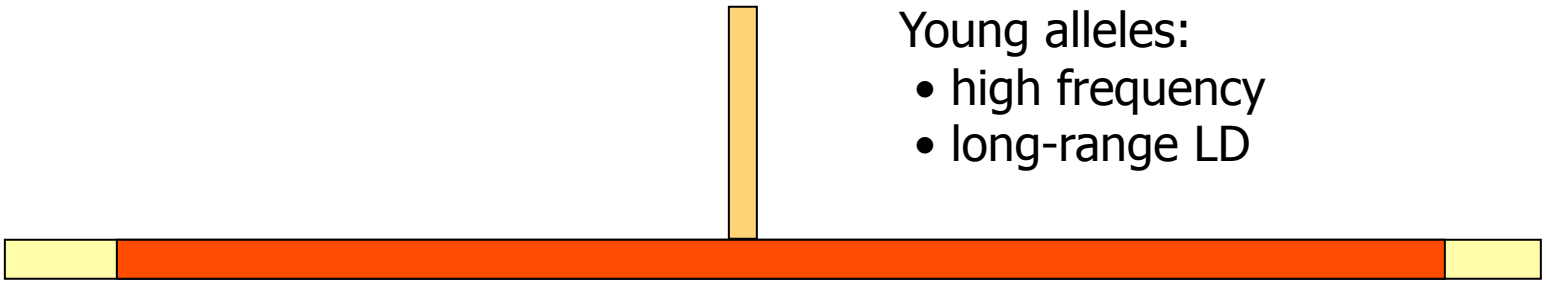
Young alleles:
- low frequency
- long-range LD

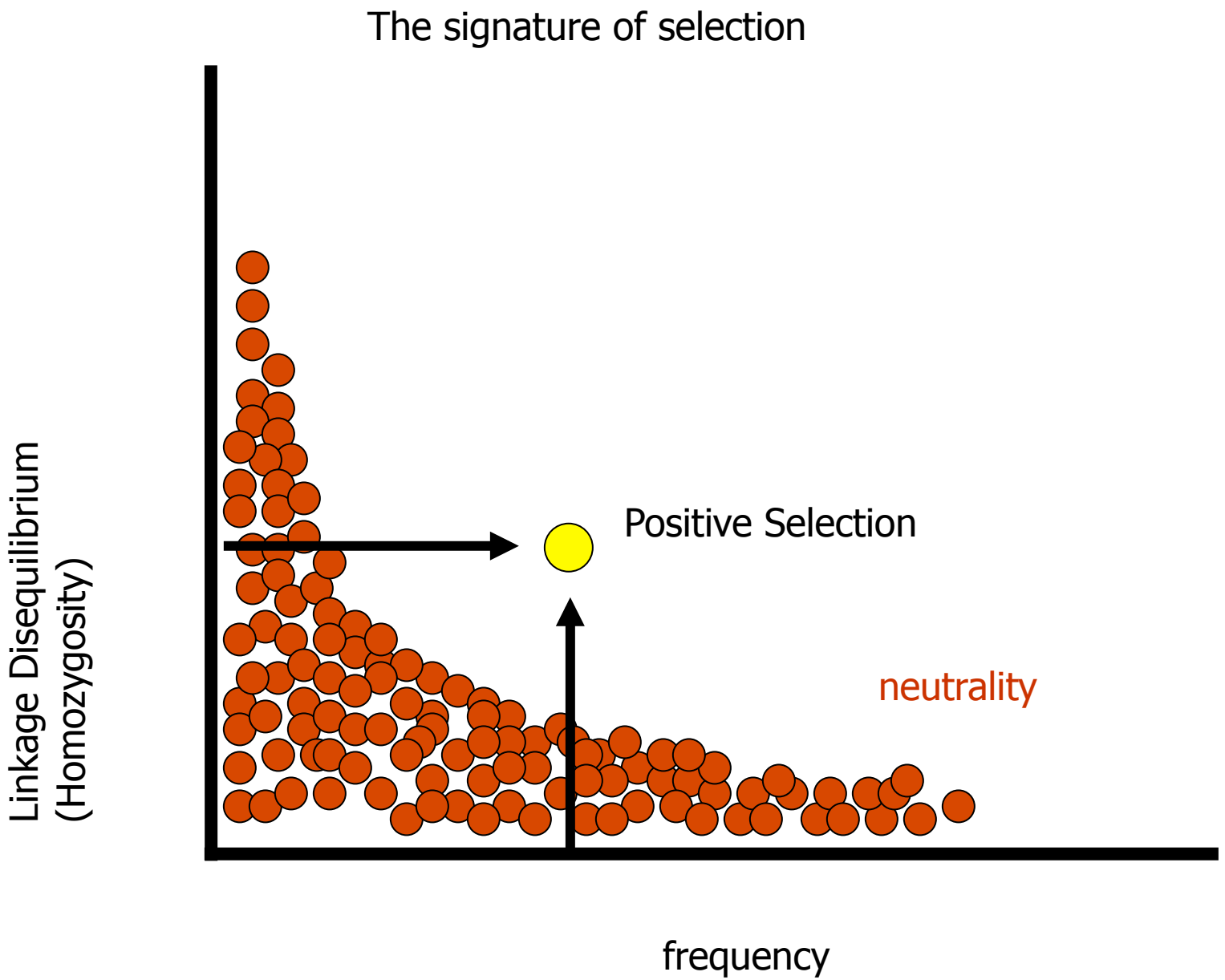


Old alleles:
- low or high frequency
- short-range LD

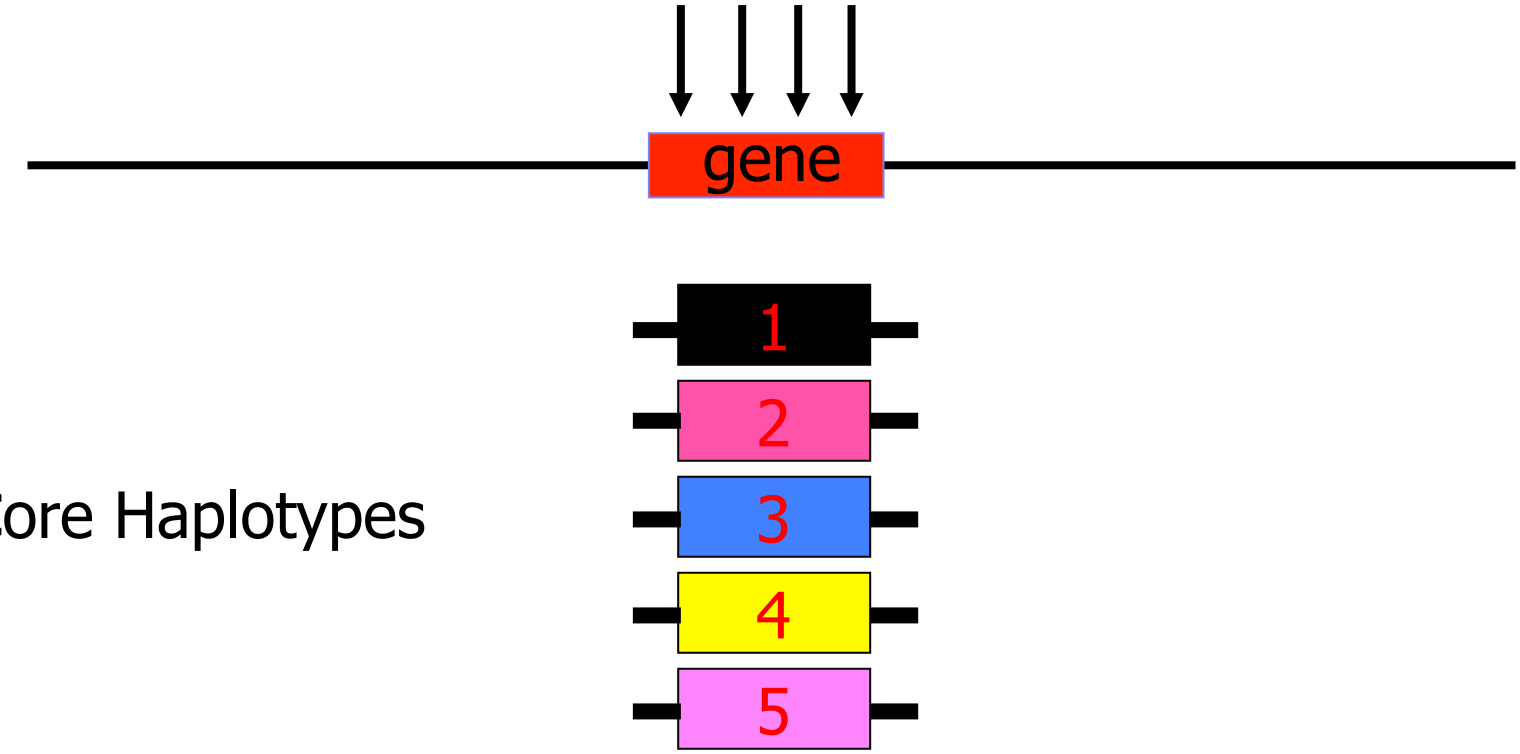
Positive Selection

Young alleles:
- high frequency
- long-range LD



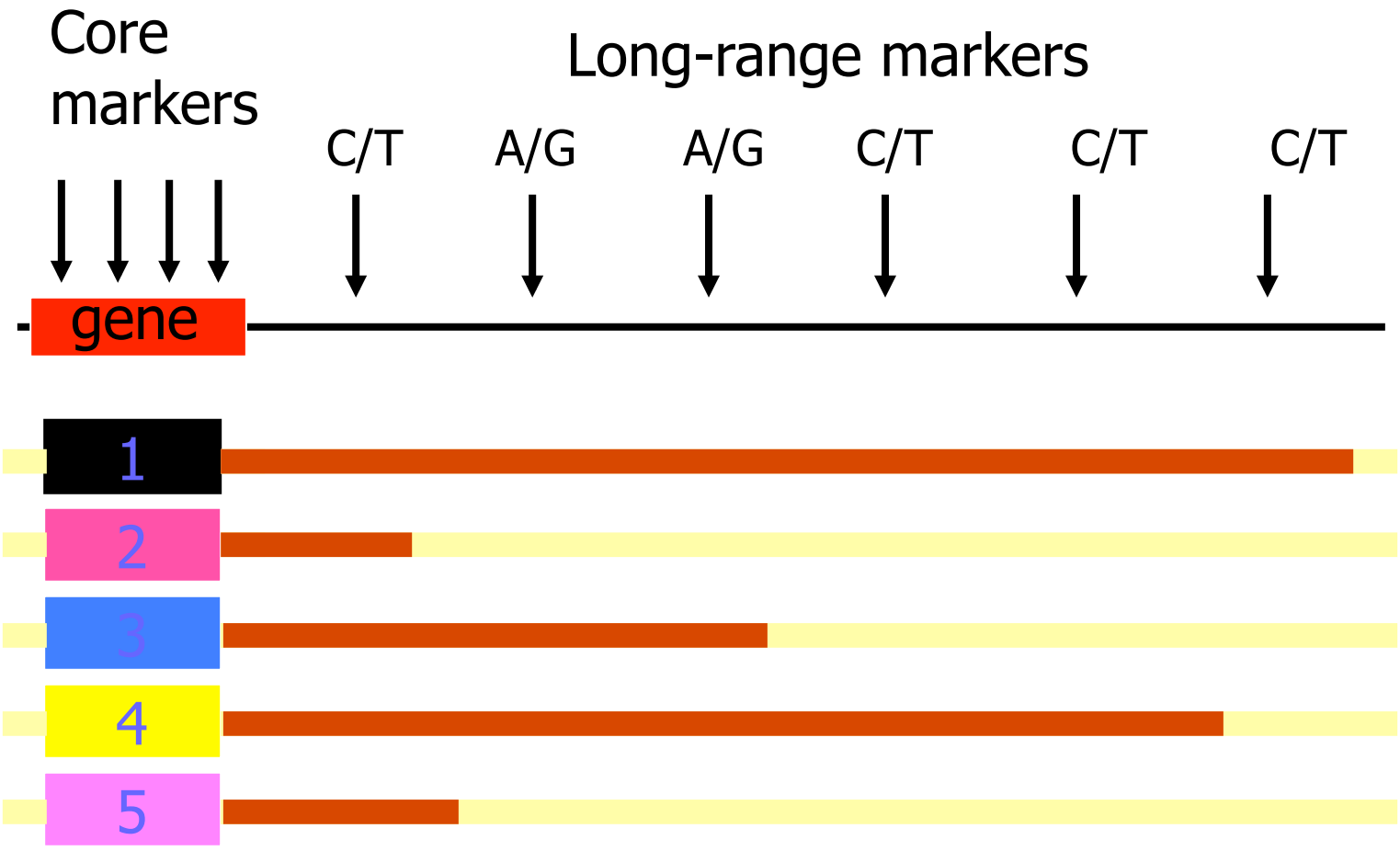

The signature of selection
Linkage Disequilibrium
(Homozygosity)
Positive Selection
neutrality
frequency

Consider a Core Region, putative target of selection

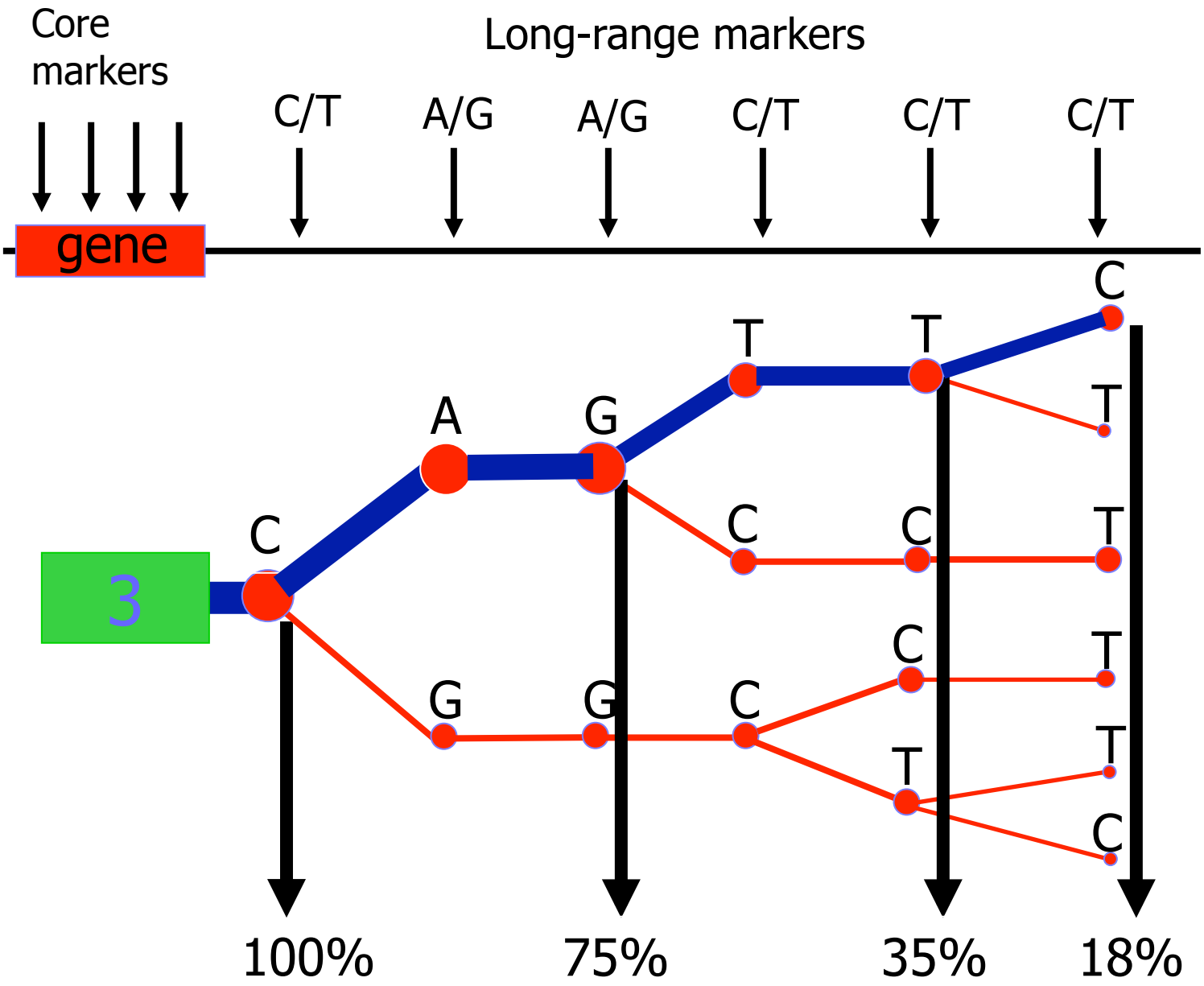


Core Haplotypes

Long-range multi-SNP haplotypes



Long-range multi-SNP haplotypes



An example: two genes associated with malaria resistance

G6PD (1960's)

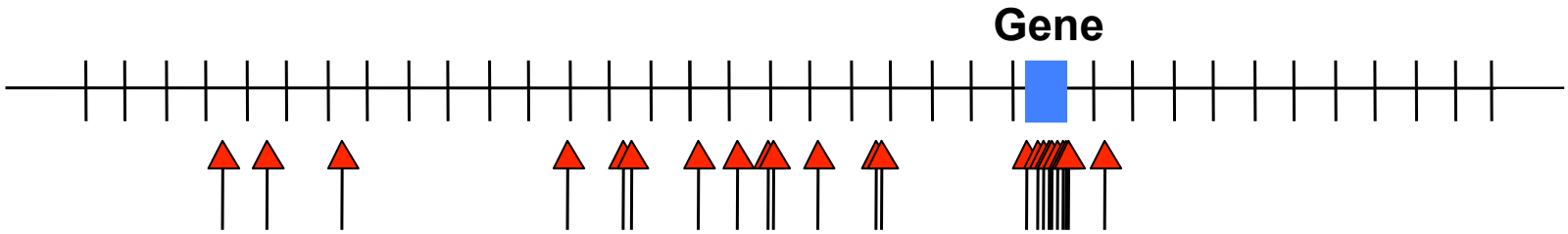
- well established association to malaria resistance
- selection demonstrated in 2001 by Tishkoff et al.

CD40 ligand (2002):

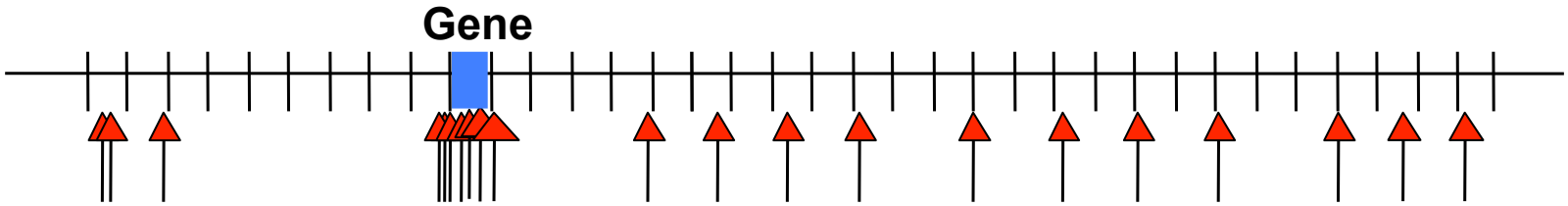
- Recent association by Sabeti et al.
- involved in immune regulation

Experimental Design…: SNPs

G6PD (11 SNPs in core, 14 at long distances)



CD40 ligand (7 SNPs in core, 14 at long distances)

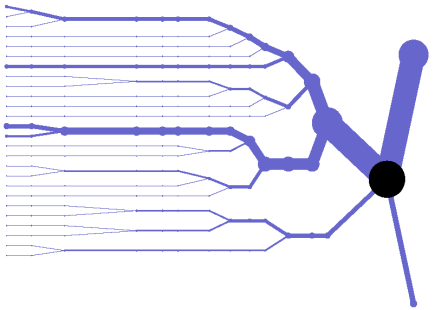


Experimental Design (cont.): samples

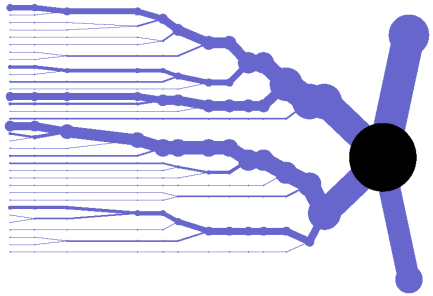
DNA samples from 231 African men

- Yoruba (Nigeria)
- Beni (Nigeria)
- Shona (Zimbabwe)

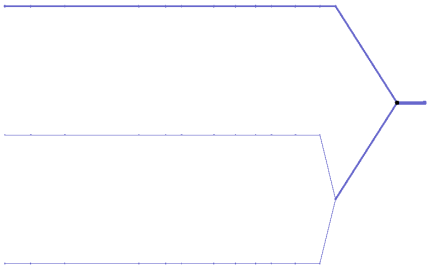
G6PD-corehap1



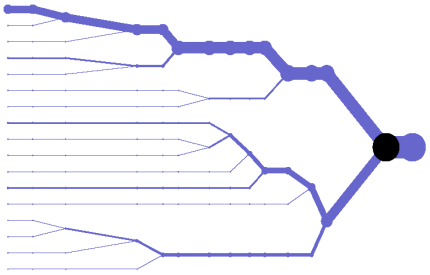
G6PD-corehap3



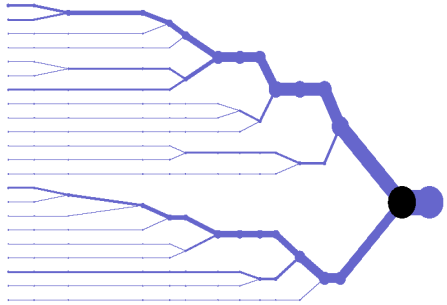
G6PD-corehap4



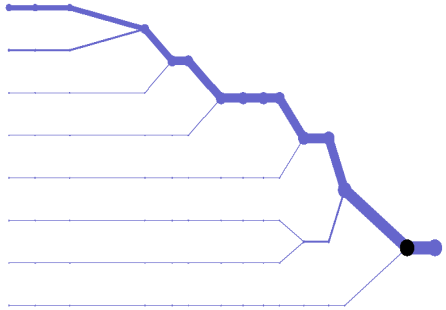
G6PD-corehap5



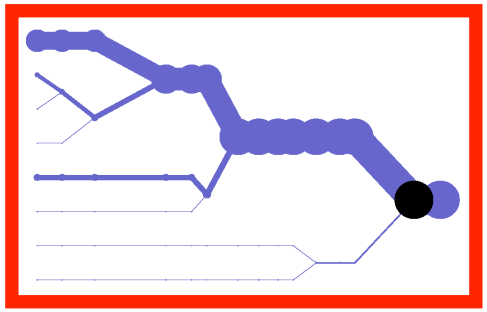
G6PD-corehap6



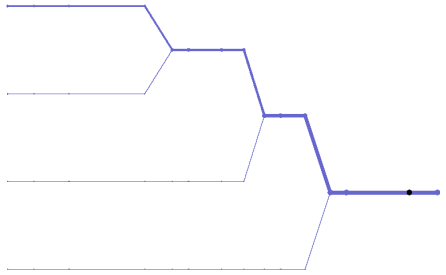
G6PD-corehap7



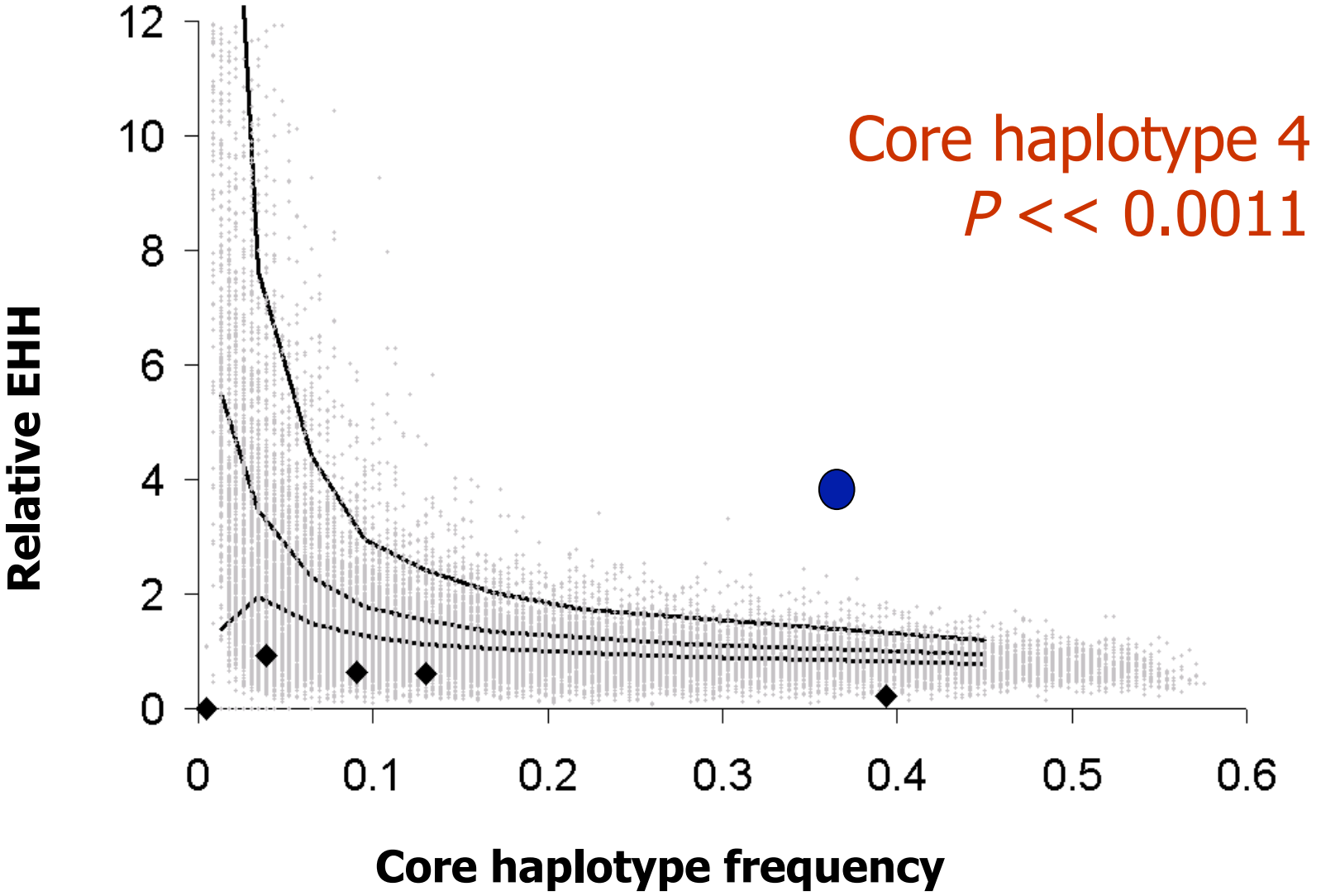
G6PD-corehap8
"A-" protective
haplotype



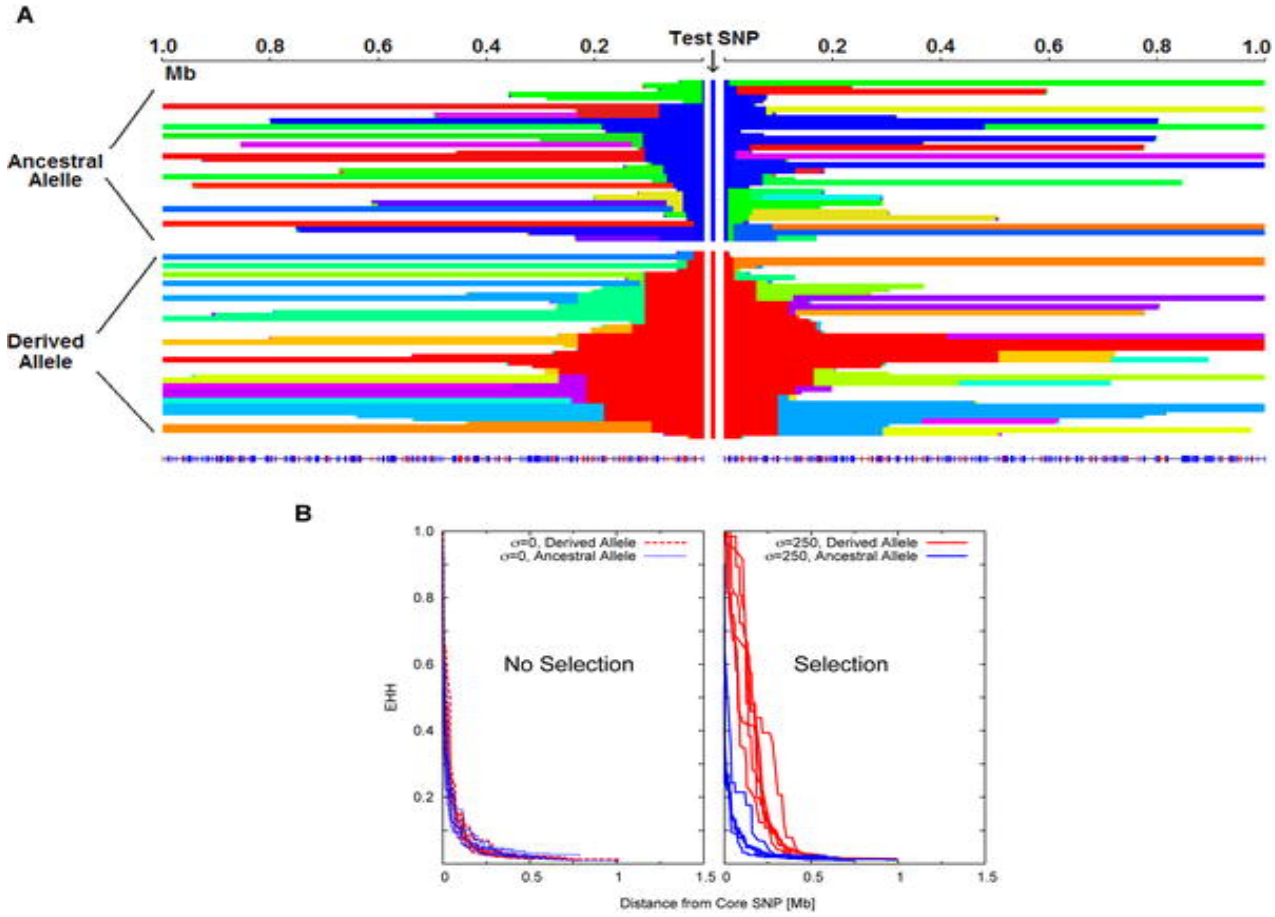
G6PD-corehap



CD40 ligand: computer simulation vs. data



A Genotype Relevant Test for Departure from Neutrality- Integrated Extended Haplotype Homozygostiy



Voight et al. 2005

Integrated EHH

- Area under EHH curve: iHH
- Calculated with respect to ancestral and derived alleles: iHH_A & iHH_D
- Unstandardized iHS = $\ln(iHH_A / iHH_D)$

~0 with equal decay of EHH

>0 with long haplotypes with ancestral allele

<0 with long haplotypes with derived allele

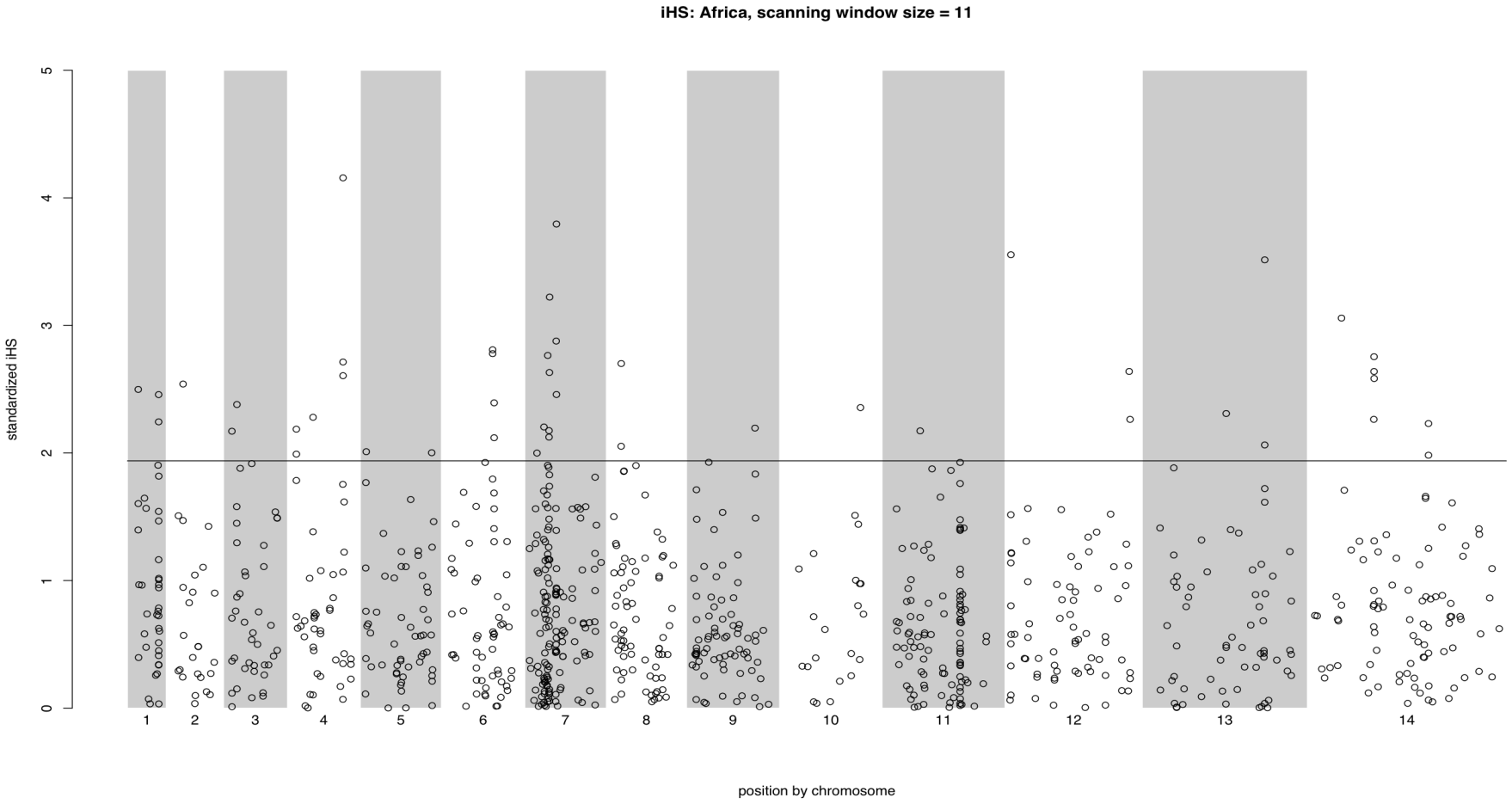
iHS

- Standardized iHS has mean 0, variance 1

$$\text{iHS} = \frac{\text{unst. iHS} - E_p[\text{unst. iHS}]}{SD_p[\text{unst. iHS}]}$$

- Expectation and standard deviation are estimated from empirical distribution for alleles with the same derived frequency as the core snp

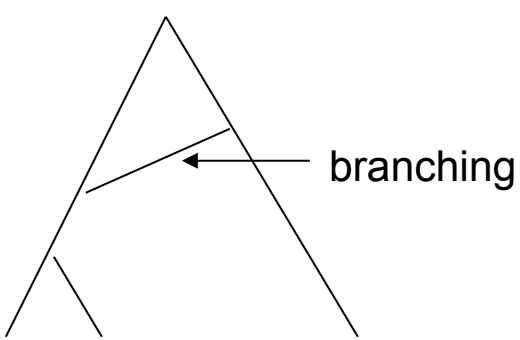
Genome Wide Scans for Selection in Malaria - Africa



Modelling or simulating selection with the coalescent…

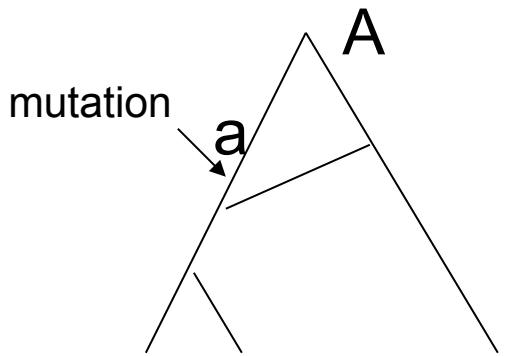
Ancestral Selection Graph (ASG) (Neuhauser and Krone)

For weak selection (k=3 lineages)…



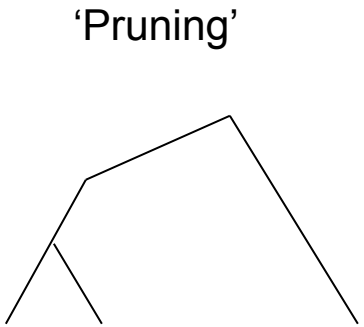
Step 1

Coalescent events and ‘branching’ events at rate $\theta k/2$



Step 2

Mutations thrown an the tree (θ)



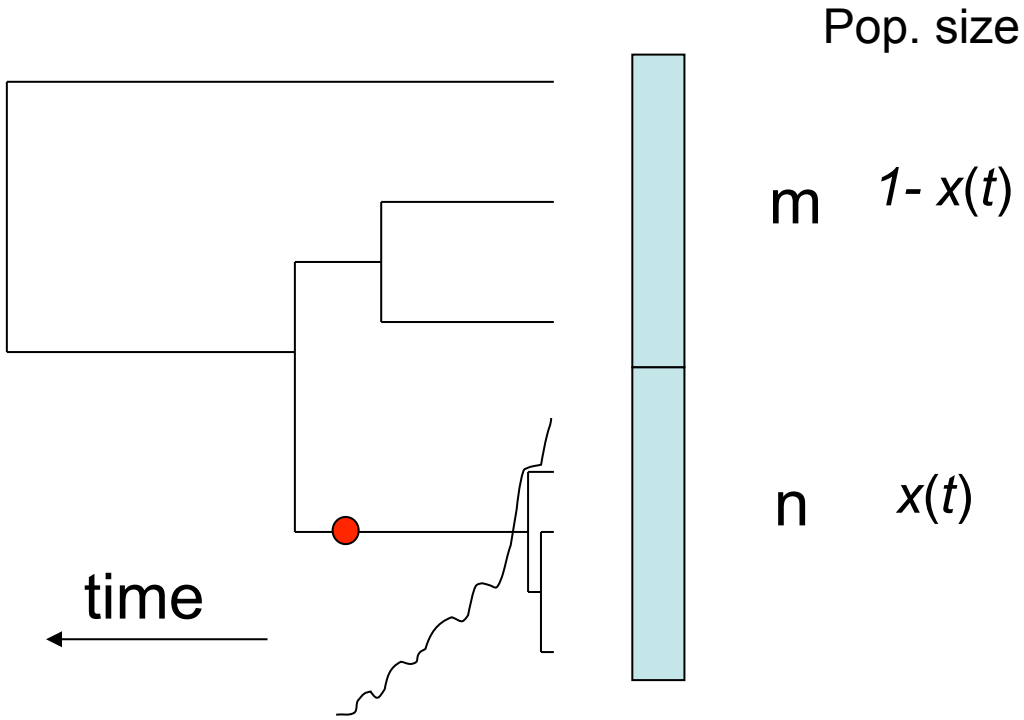
Step 3

Removal of the branch with the unpreferred ‘**a**’ mutation.

This forward and reverse process results in a shorter TMRCA

Simulating the coalescent with natural selection…

Generate a genealogy conditional on the reverse trajectory of a derived allele.
The reverse trajectory is generated through a (reverse) diffusion process.



The trajectory of the selected allele through time $x(t)$ is shown as the grey line. n samples have the derived selected allele.

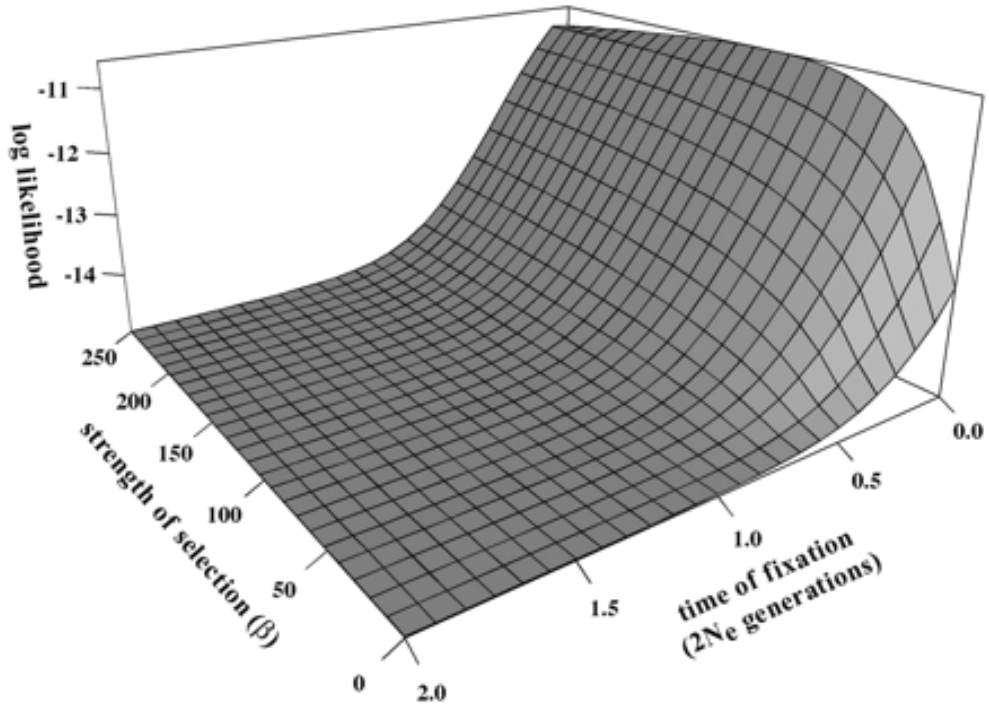
Griffiths, Coop and Griffiths, Coop and Spencer

Making inferences of selection with genealogies incorporating selection…

Importance sampling
scheme (Stephens and Donnelly)

Calculate the likelihood of the two subtrees (M and N) per trajectory Simulated.

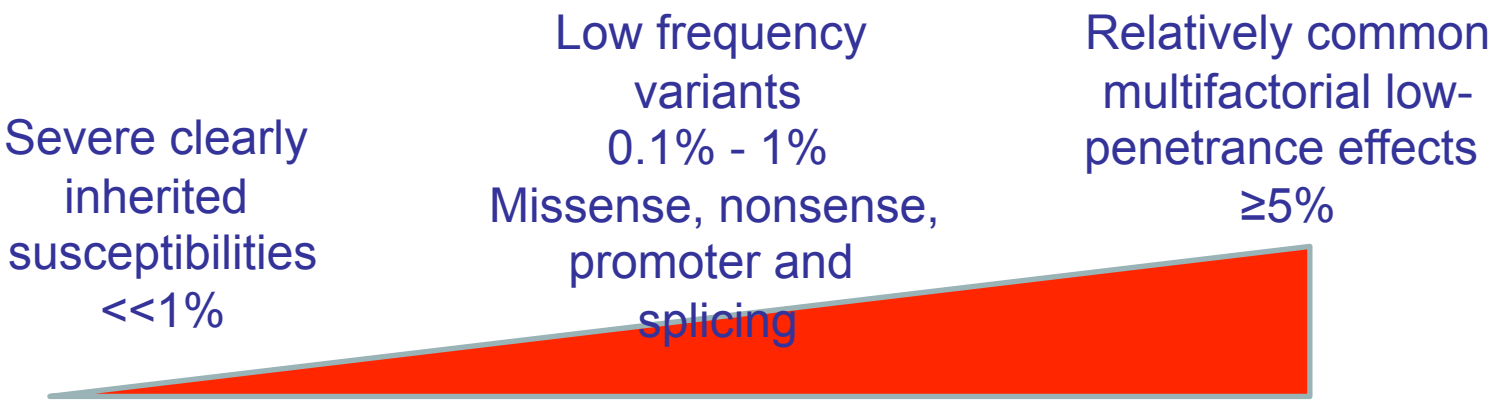
P(D|{X(t),t≥0})



The joint likelihood surface of the time since the end of the selective sweep ($T = 2N_e$) and the strength of selection ($\gamma = 4N_e s$).

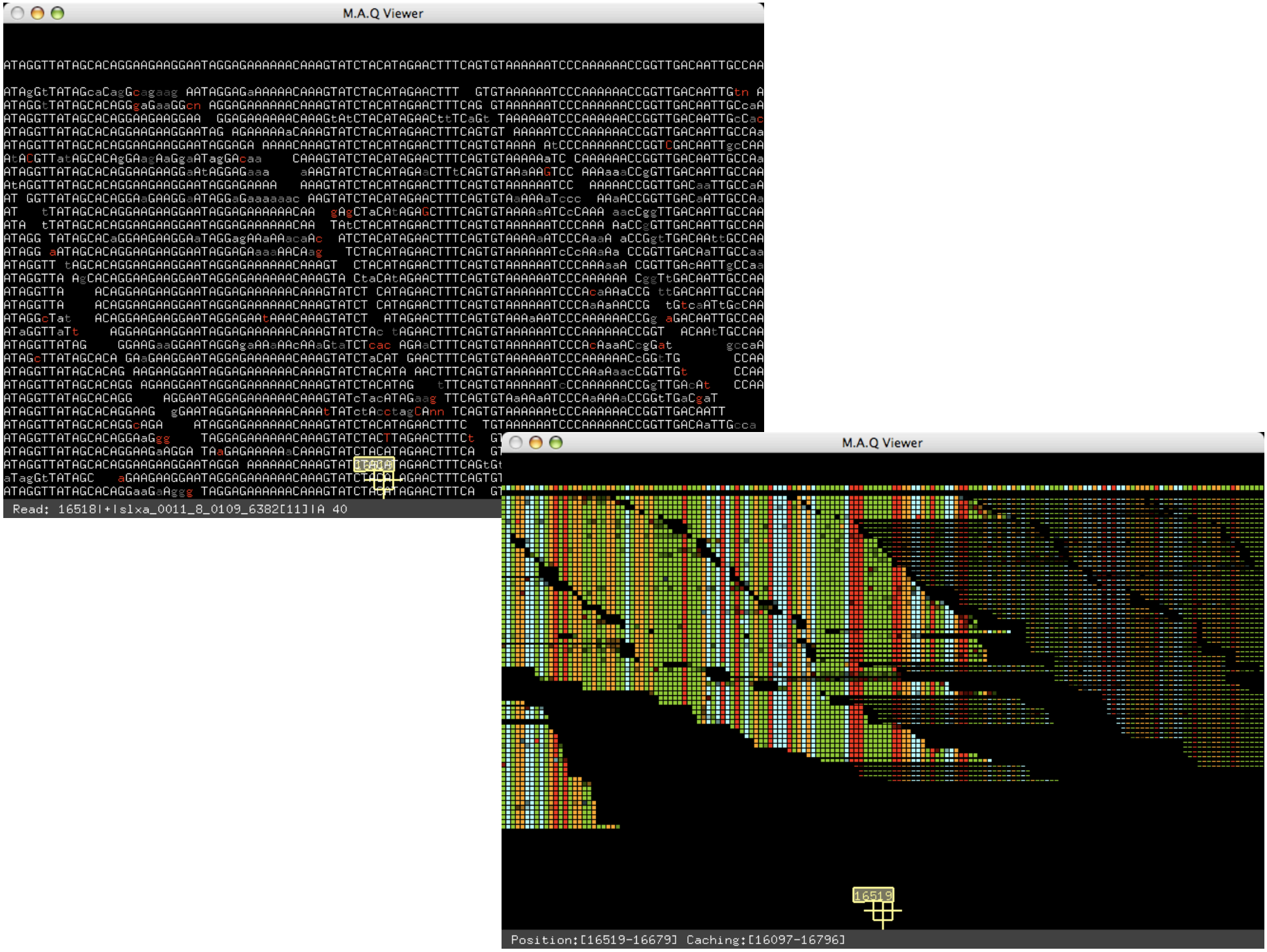
“Genomic Models” of Disease

The way we interrogate genomes to discover pathways, genes and mutations is going to be dependent on the kinds of variation we believe are responsible for disease, their frequencies and their relative risks.

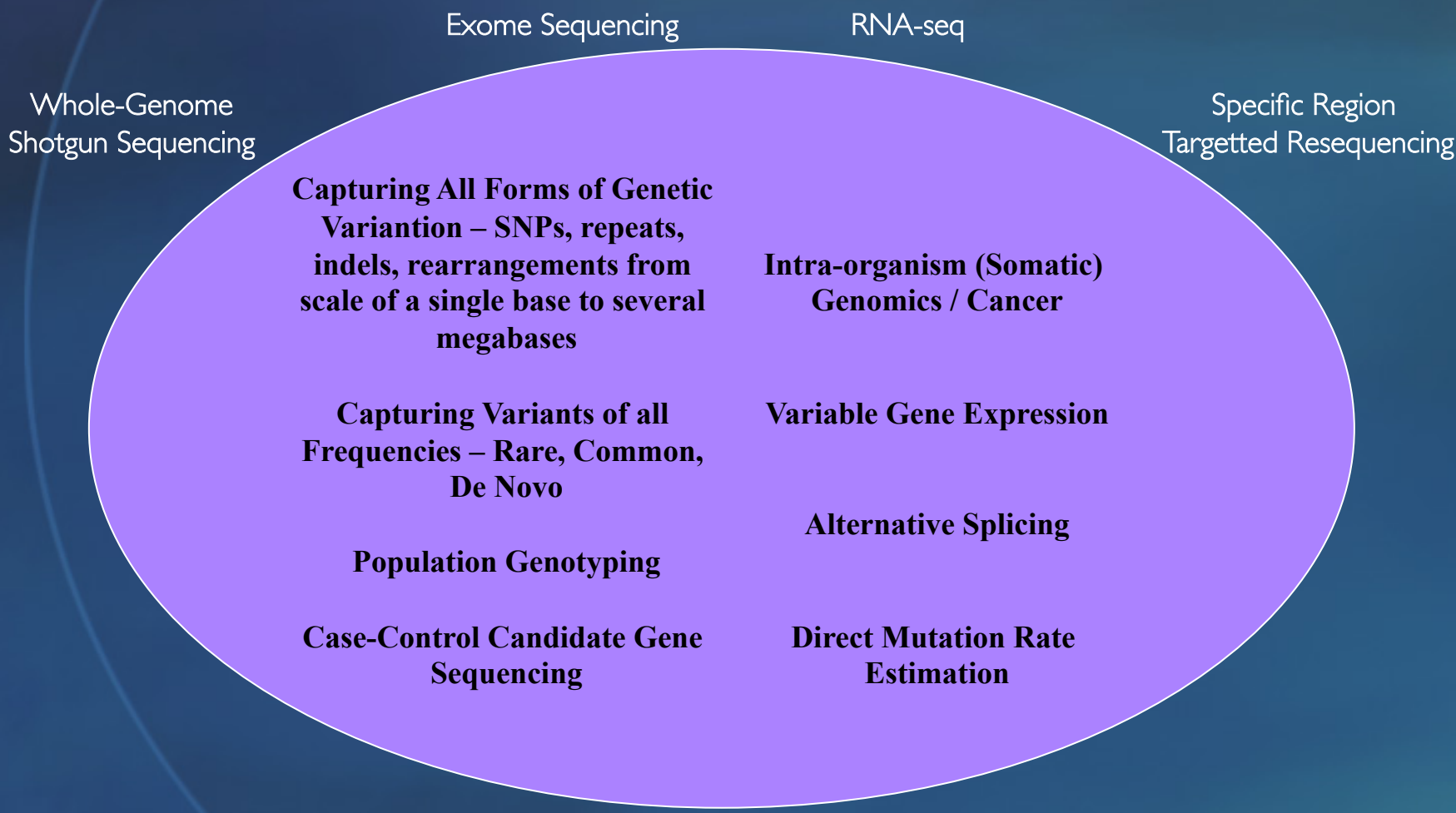


Discovery of Spontaneous (*de novo*) mutations using "Next Generation" Whole Genome Sequencing of Trio and Twin Family Cohorts (cont.)

- We propose a new probabilistic tool for directly estimating mutation rates within a resequencing study of nuclear families
  - We investigate the effect varying levels of sequence depth coverage have on the ability to estimate *de novo* mutation rates
    - We investigate the effect of family type (single offspring vs. mono- and di-zygotic twins.


M.A.Q Viewer
Read: 16518|+|slxa_0011_8_0109_6382[11]|A 40
M.A.Q Viewer
16519
Position:[16519-16679] Caching:[16097-16796]

Applications of Next-Generation Sequencing



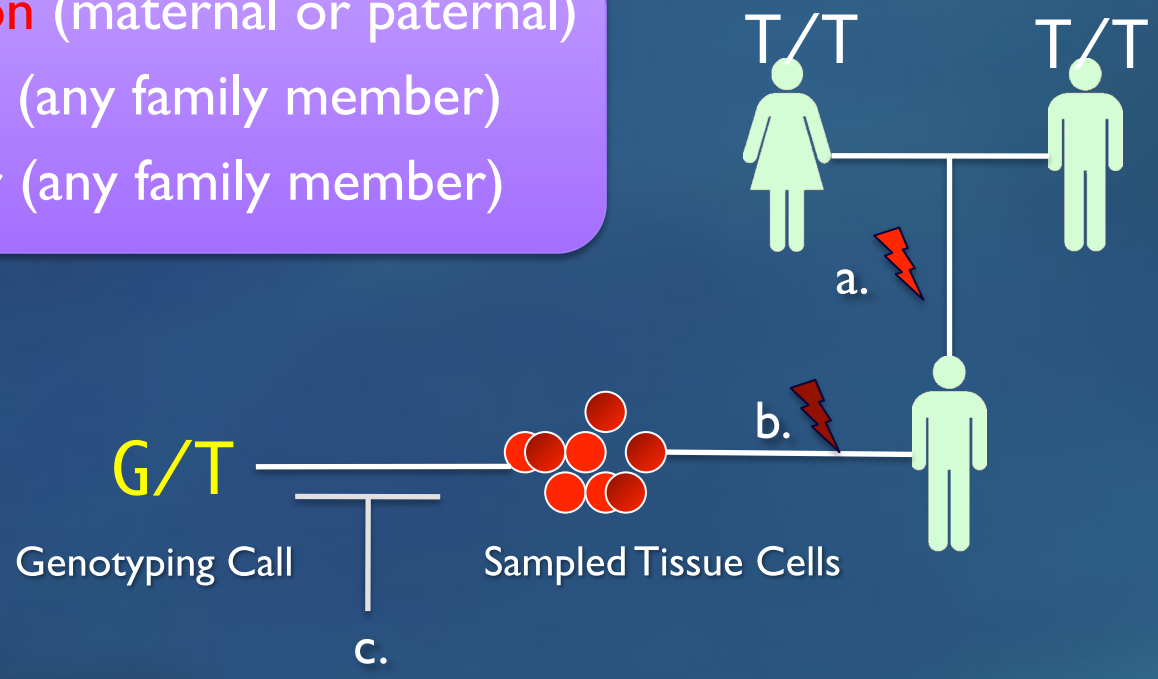
Summary

- Motivation & Introduction
- Pedigree-based Model
- 1000 Genomes Project Trios
- Somatic Mutation Discovery in Leukemia
- General Conclusions

Inferring a Mutation within a Pedigree

Mutations present as Mendelian errors

a. Germline mutation (maternal or paternal)
b. Somatic mutation (any family member)
c. Genotyping error (any family member)

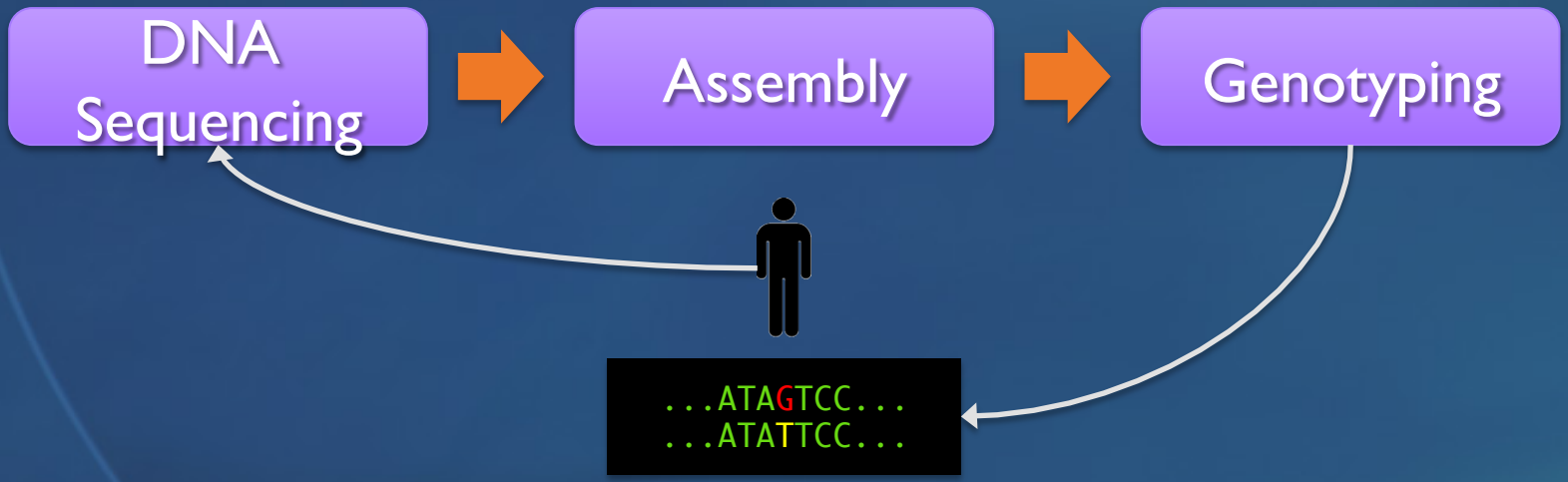


High-throughput DNA Sequencing (HTS)

Obtains complete genotype information

Captures many types of genetic variation

Expensive, difficult Bioinformatics



Bioinformatics of HTS

DNA Sequencing

Output: Reads

de novo Assembly

Velvet

SHARCGS

ABySS

ALLPATHS

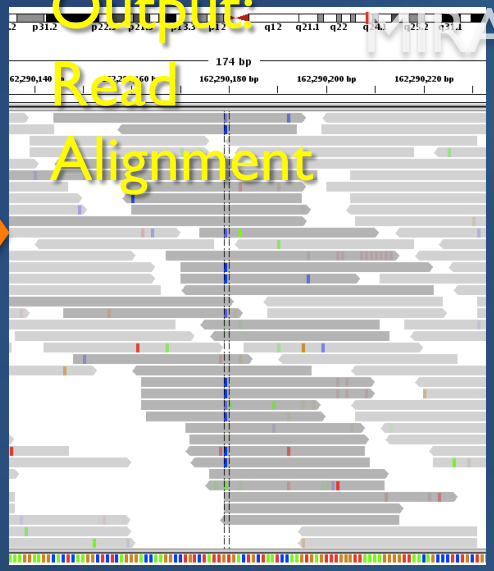
VCAKE

MIRA2

No standard method exists for automated genotyping of HTS assemblies

Assembly Output:

Read Alignment



Referenced-Based Assembly

MAQ

BWA

PerM

BFAST

Bowtie

Bioscope

Genotyping ?

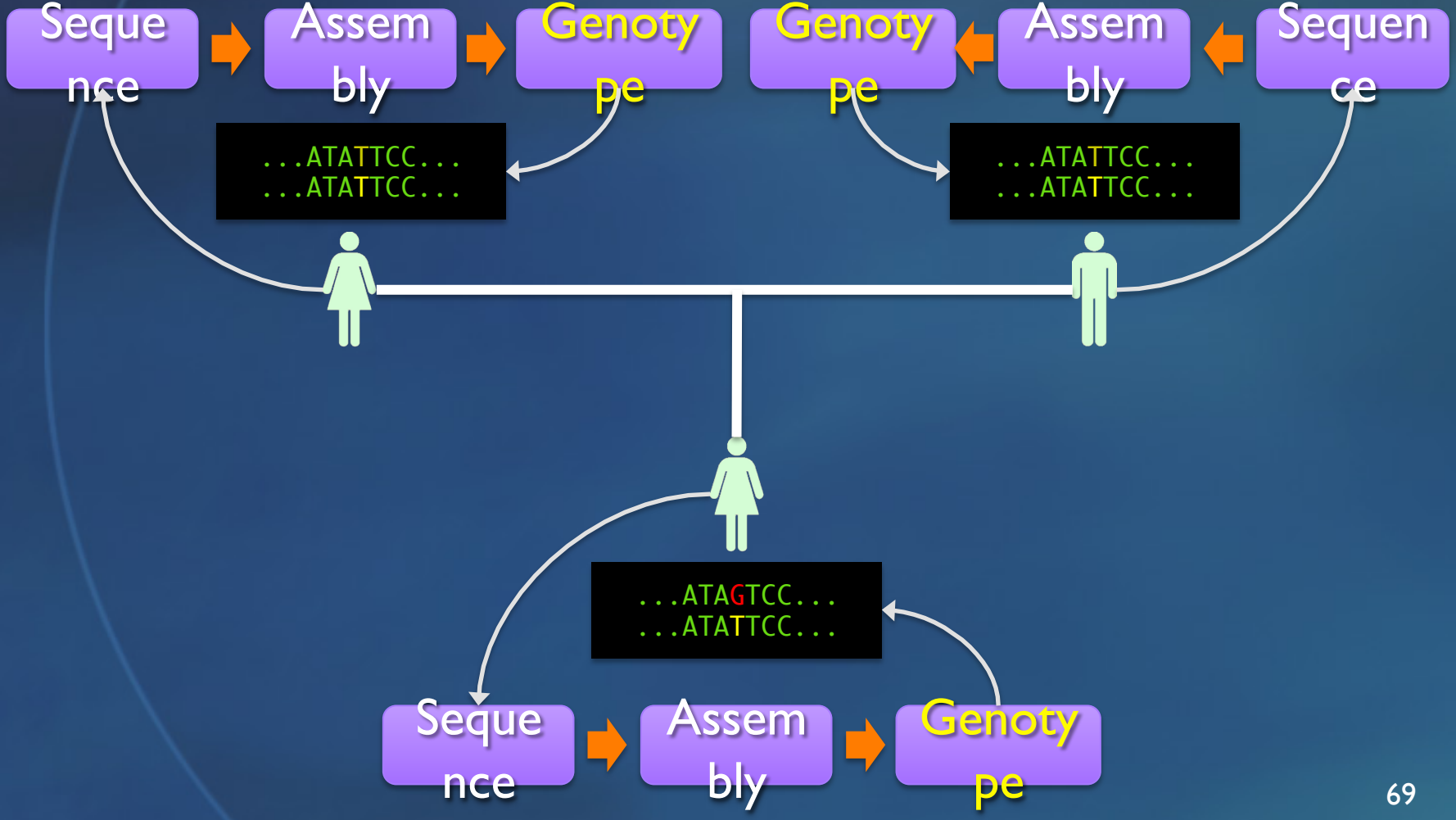
Consensus Sequence

Read Frequency

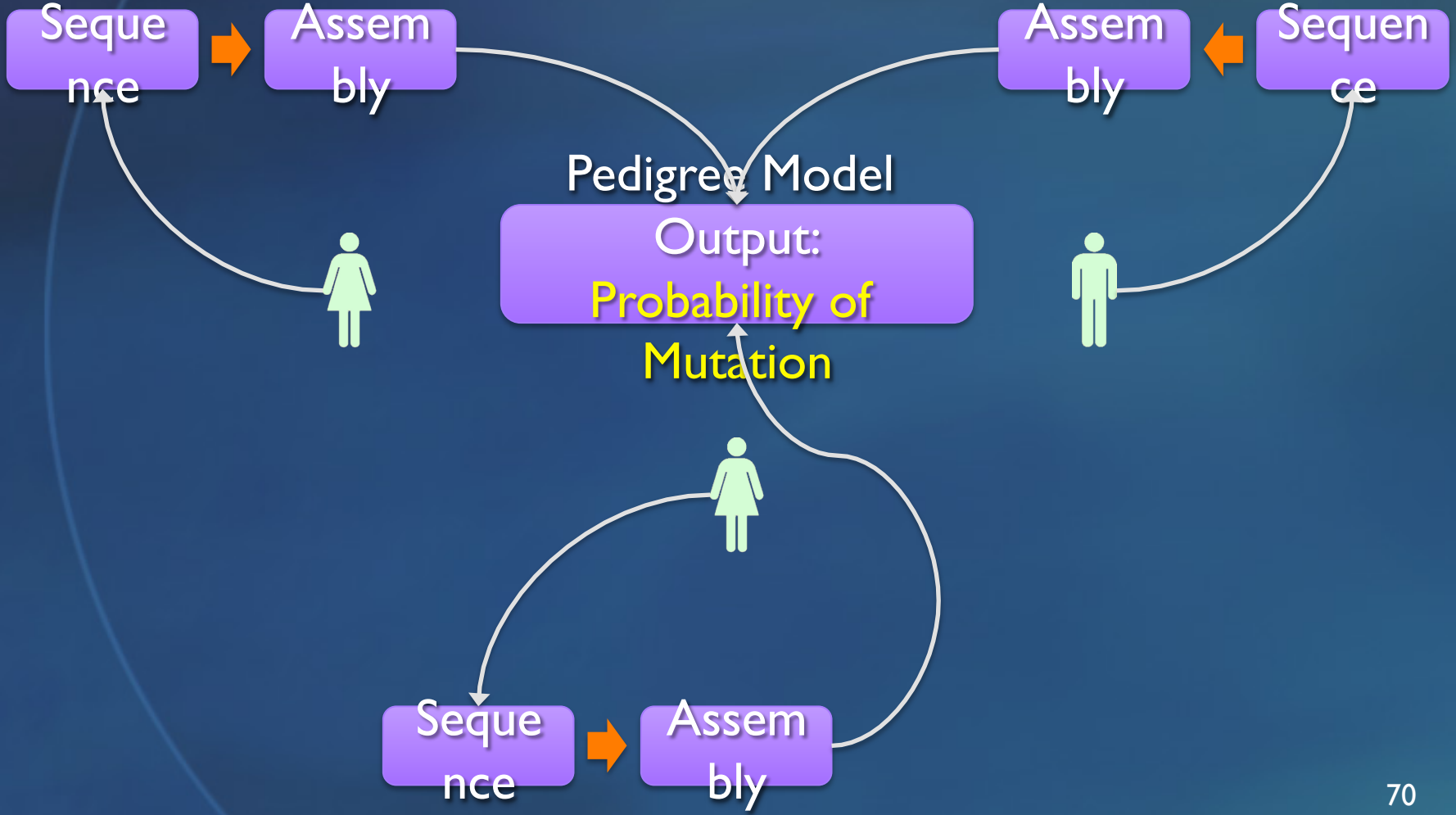
Base Qualities

Mapping Qualities

Standard HTS of a Nuclear Family



Jointly Infer Genotypes to Locate Mutations



Pedigree-Based Method for Mutation Discovery

Parameters of interest

•Somatic Mutations are disentangled from Germ Line Mutations and each are given separate rate estimates

•*D* is the observed data, consisting of the set of observed nucleotide frequencies at every site sequenced for every individual, for all families

•*H* is the 'hidden' data within the model, corresponding to the actual genotypes of the individuals, their inheritance pattern, and the sampling of different chromosomes by resequencing

•Θ is the set of parameters within our model, consisting of:

• θ – Inferred Population mutation rate	*Given
• N_S – Number of sites sampled	*Given
• N_F – Number of families sampled	*Given
• μ_s – Direct Somatic mutation rate parameter	*Estimated
• μ_m – Direct Maternal mutation rate parameter	*Estimated
• μ_p – Direct Paternal mutation rate parameter	*Estimated
• ϵ – Direct Sequencing error rate parameter	*Estimated

Trio Model

Parameters

θ – Population Mutation Rate

μ – Germline Mutation Rate

μ_S – Somatic Mutation Rate

ε – Error Rate

Observed Data

R_M R_O R_F

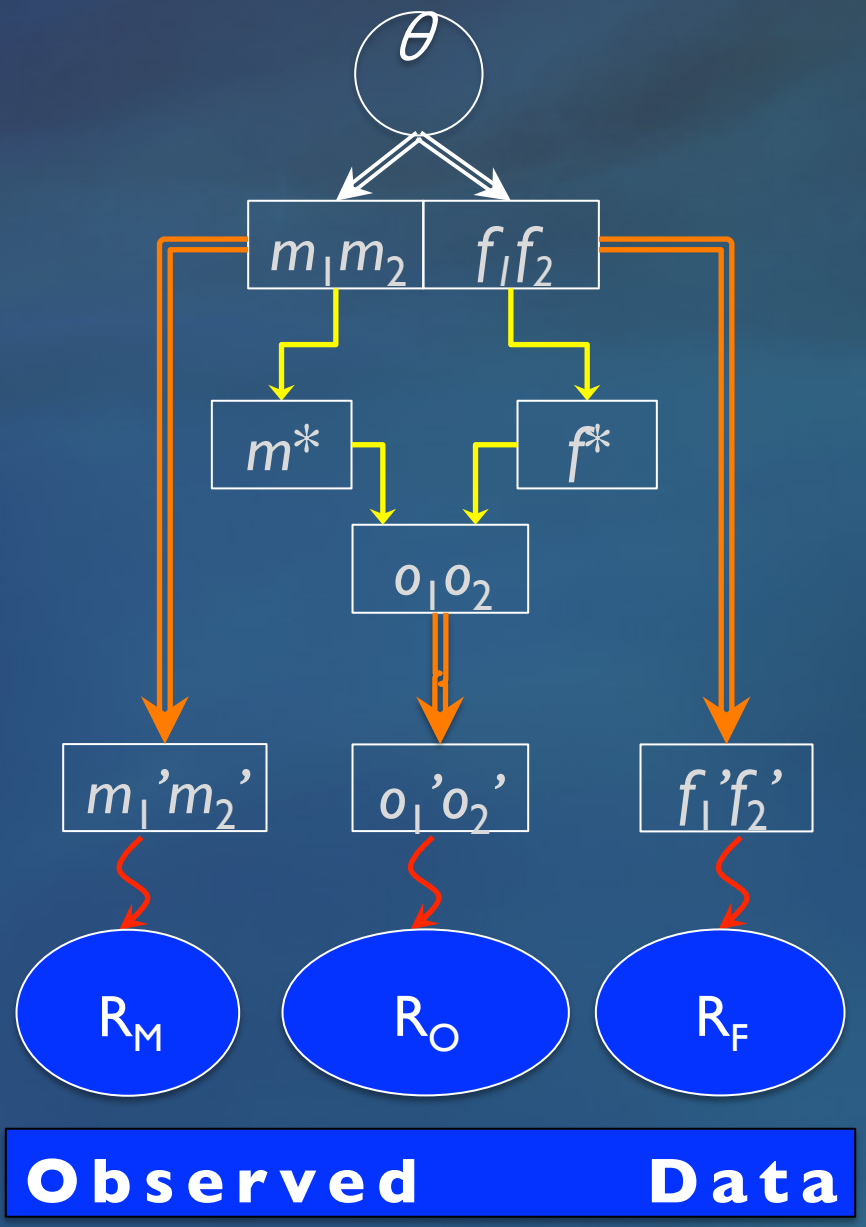
Hidden Data

m_1m_2 – Germ Genotype of Mother

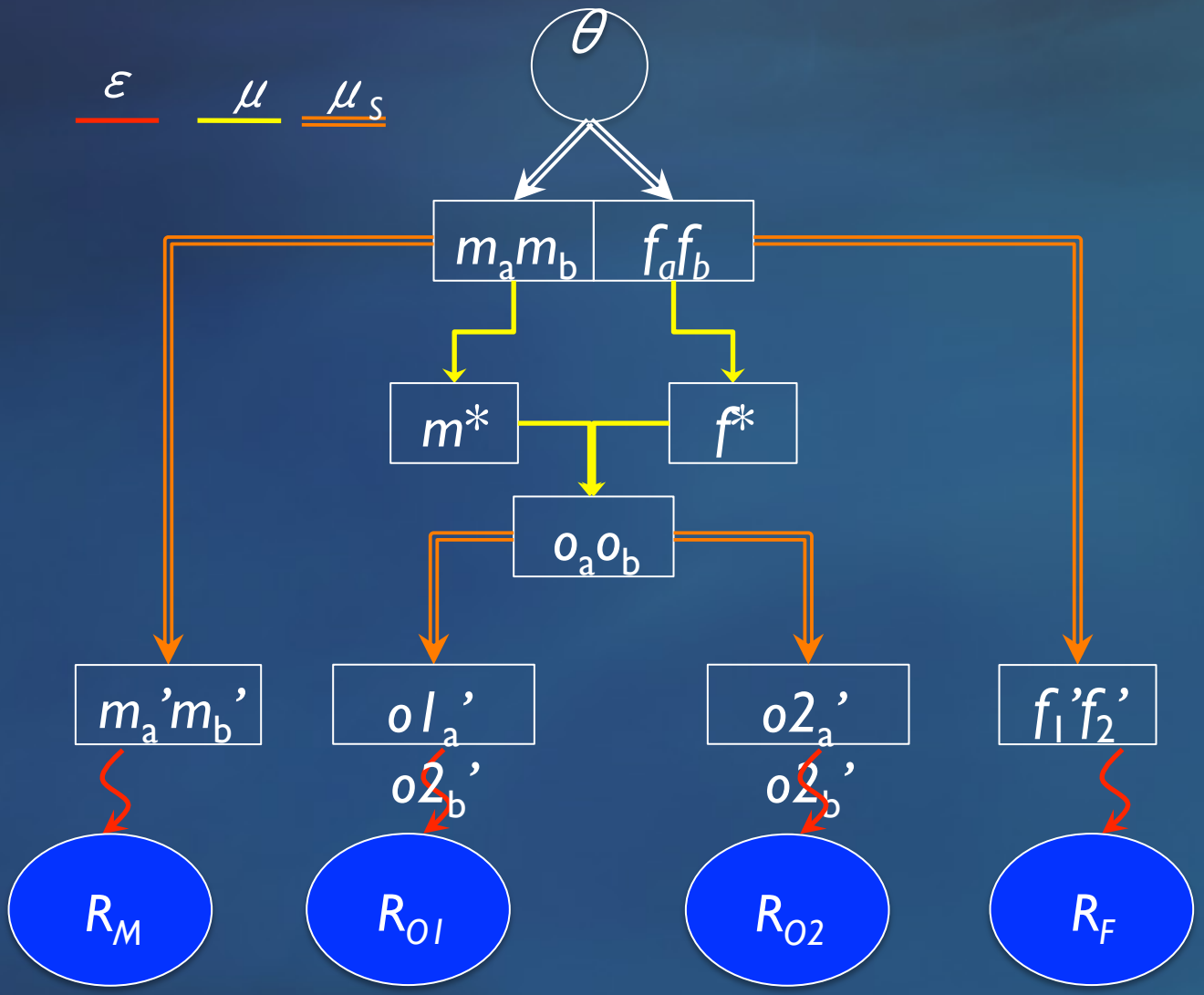
f_1f_2 – Father Germ Genotype

o_1o_2 – Offspring Germ Genotype

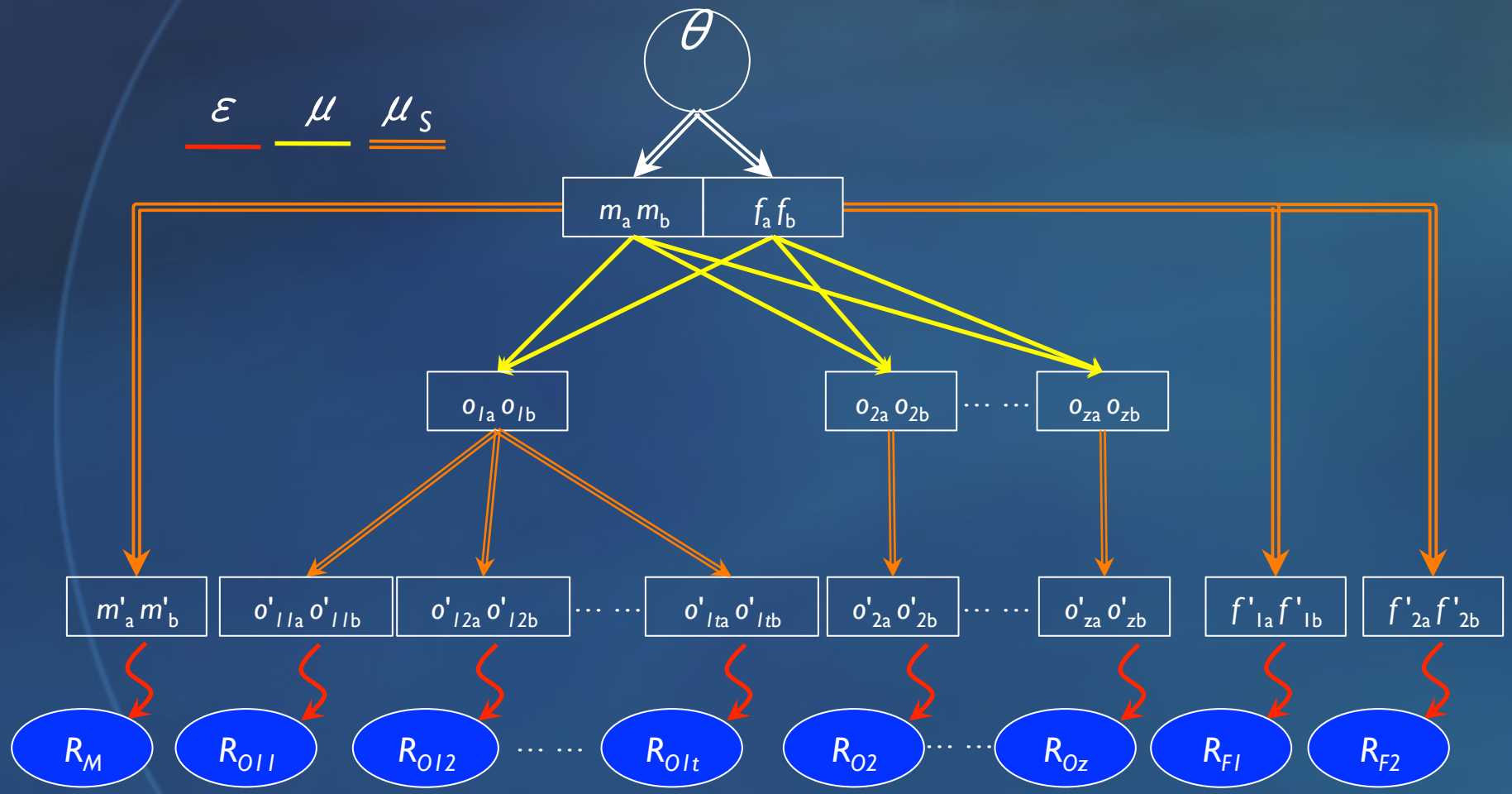
m^* – Transmitted Mother Gamete



Monozygotic Twin Model



Generalized Pedigree Model



Discovery Algorithm

1. Compress / Import Data
   - quality filtering
   - Nx4 data structure

2. Select Parameter Values
   - Maximum Likelihood Estimation with an EM algorithm
   - Expert prior values

3. Evaluate model
   - Rank all sites by Posterior Probability of Mutation δ

Mutation Discovery in The 1000 Genomes Project Pilot 2

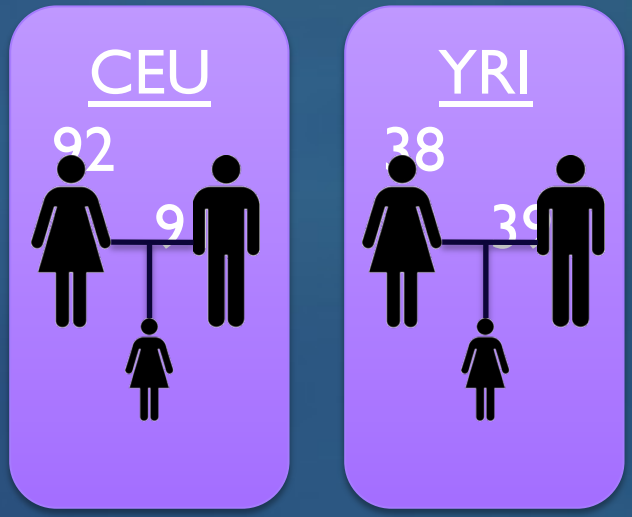
The 1,000 Genomes Project Background

Catalog rare genetic variation in humans

- > 0.1% MAF in coding regions are found
- > 1% MAF in the rest of the genome.

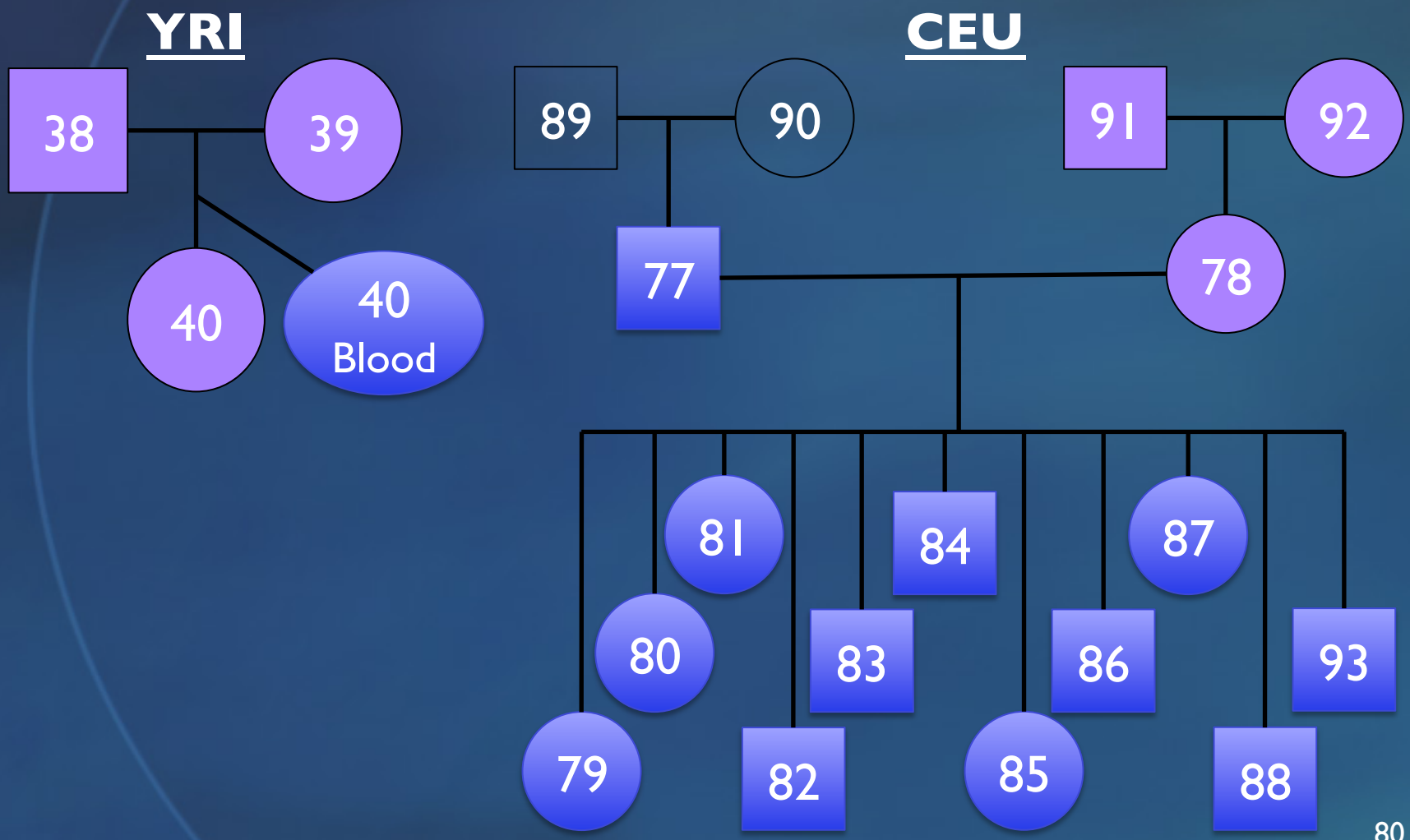
Pilot Projects

- test methodologies and technologies
- Pilot 2 deep resequencing of two nuclear families
- HapMap cell cultures

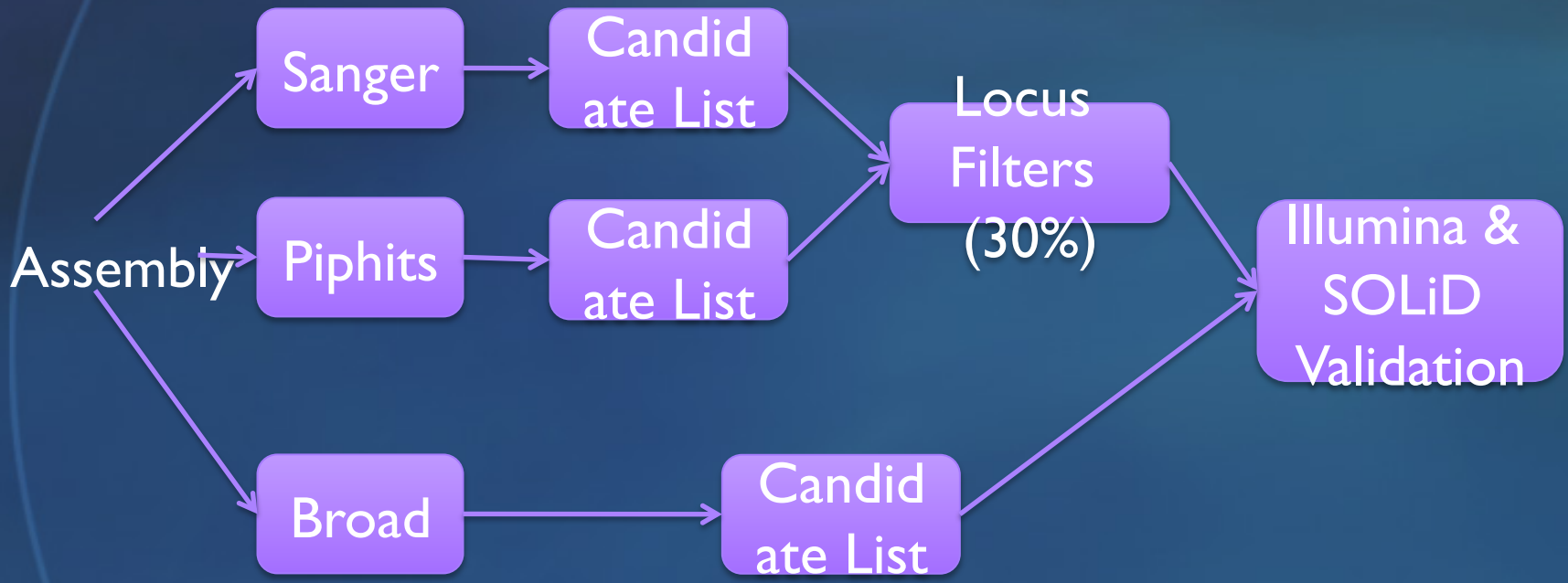


78 40

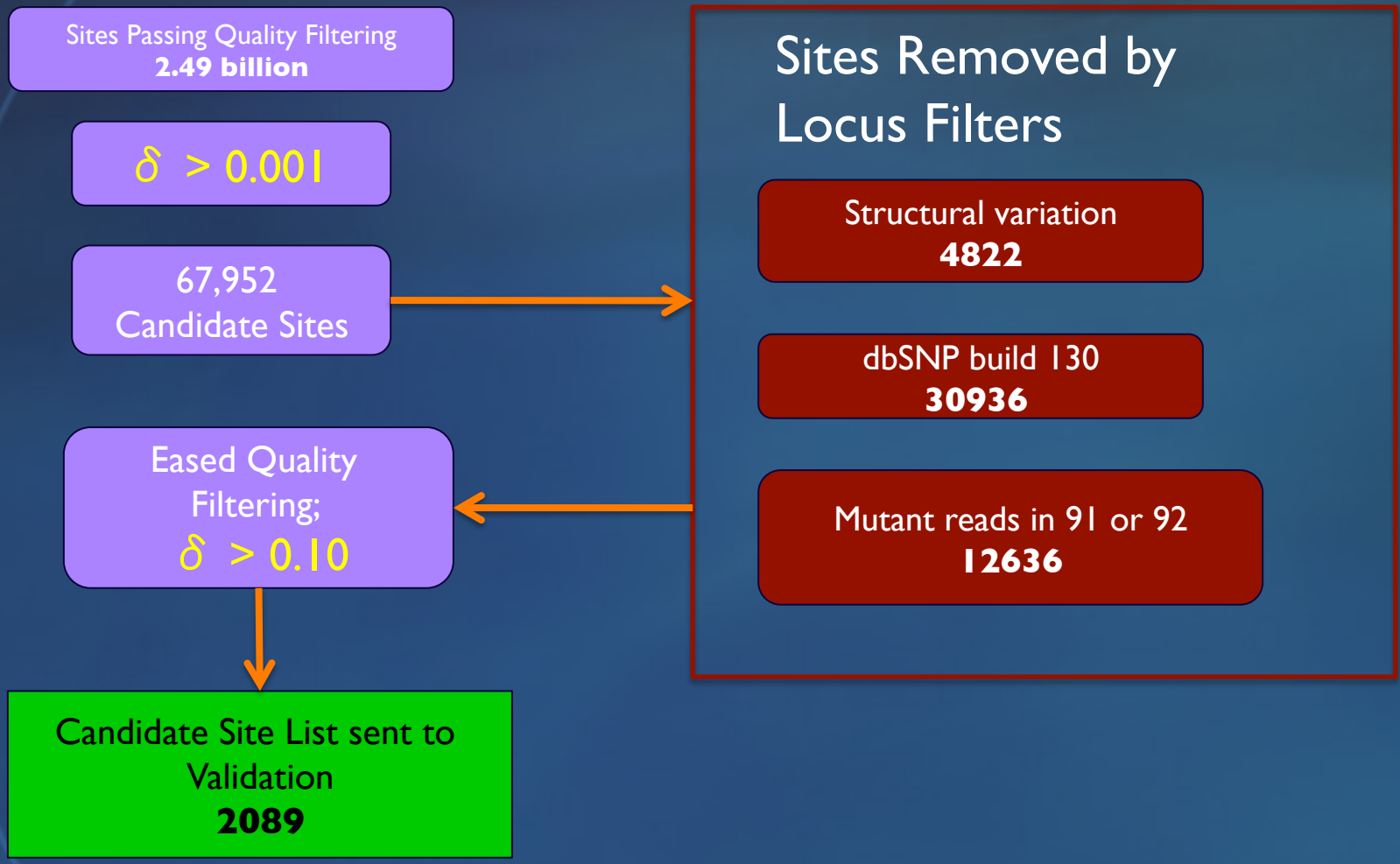
Pilot 2 (1000 Genomes) -25 X Coverage



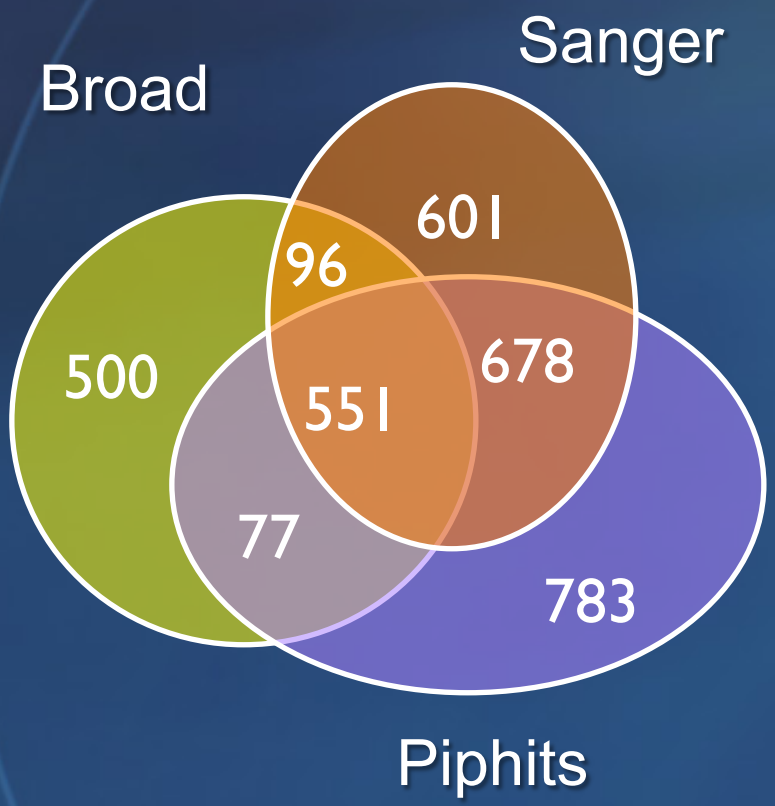
Mutation Finding Collaboration



Piphits Method Summary (CEU)

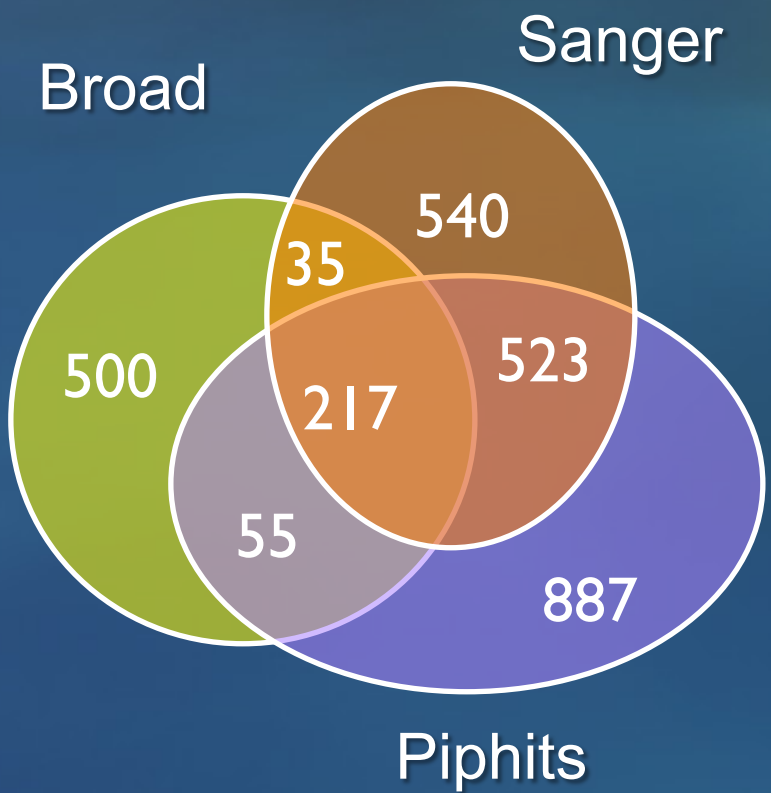


Candidate List Sent to Validation



CEU

N=3,286 sites



N=2,757 sites

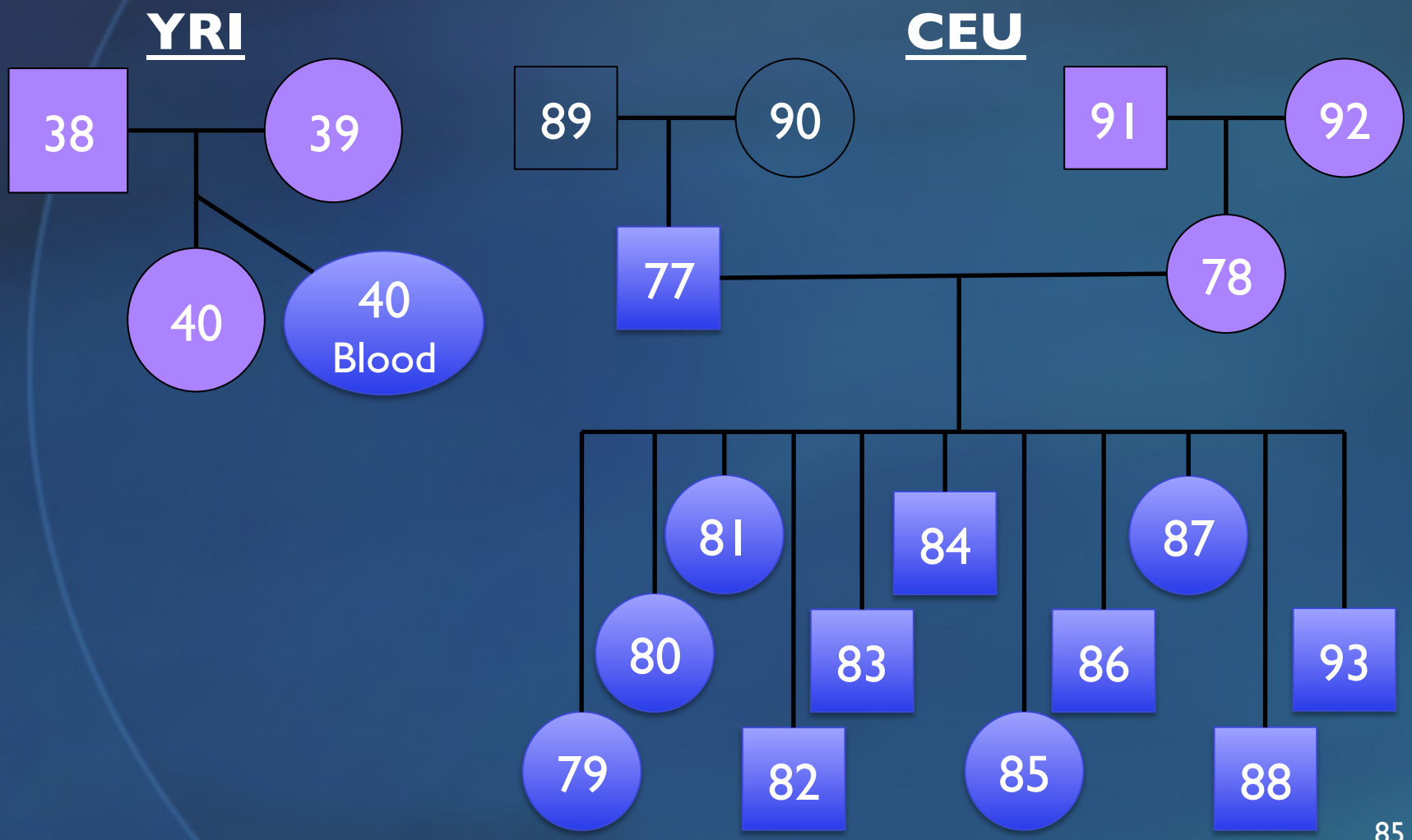
YRI

Validation Strategy

- Philosophy: identify as many *de novo* mutations as possible
  - Over-call to generate a long list of candidates to maximise sensitivity
  - Attempt validation of all candidates
- Validate all candidates by 2 different approaches
  - Agilent hybridisation capture + SOLiD sequencing (19 expts)
  - Nested PCR (1kb then100bp), pooling and sequencing on Illumina GA2 (22,520 PCR primers, ~111,000 PCRs)
  - Informative assays for 99% (SOLiD) and 91% (Illumina) of sites
- Confirm germline:
  - By transmission to 11 CEU grandchildren
  - By presence in Blood DNA (YRI trio)

	Parents	Offspring	3rd generation (CEU) / Blood (YRI)
Germline *de novo*	NO	YES	YES
Somatic *de novo*	NO	YES	NO
Inherited	YES	YES	YES
False positive	NO	NO	NO

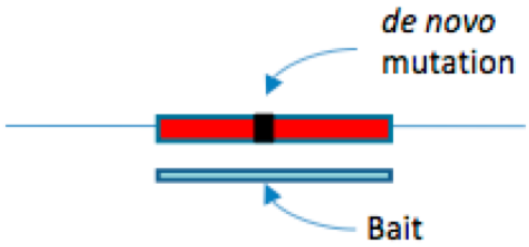
Pilot 2 (1000 Genomes) -25 X Coverage



UdeM Validation Sample Capture

Probe Design

5,838 total sites
- 3135 of 3136 CEU loci
- 2703 of 2750 YRI loci
- 120 bp cRNA probes



Agilent Capture Success

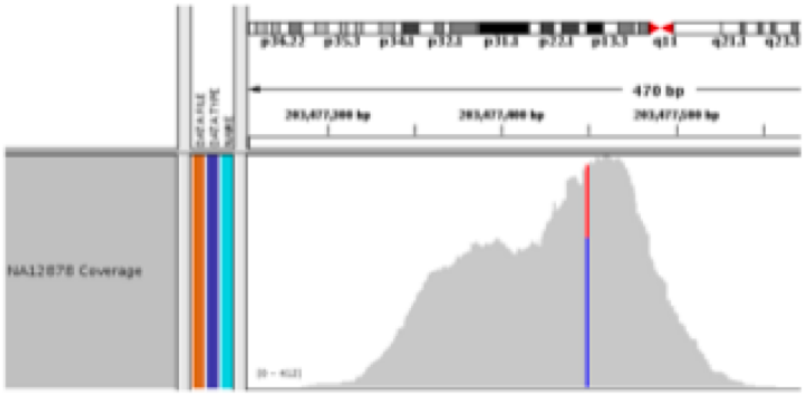
CEU –
91: 95%; 92: 97.2%
78: 95%; 77: 93%
3rd Gen:
92%, 93%(x4), 96%(x5), 97%

YRI –
38: 91%; 39: 91%; 40: 90%

SOLiD Sequencing

3 runs
One plate per run
7 lanes per plate
One sample per lane

19 Total Sequencing Experiments

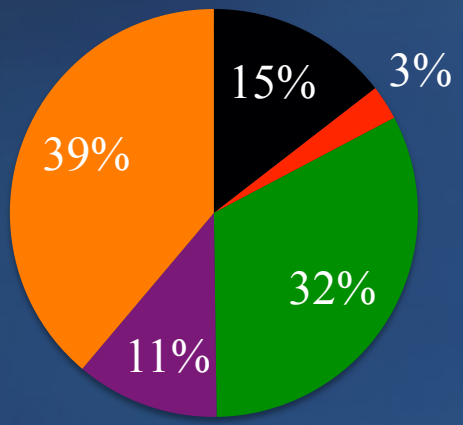


Bowtie / SAMtools Assembly

Avg. 32 million 50 bp color-space reads
Avg. 92.7% sites covered at 1x
Avg. 61.8% sites covered 15x
~60x mean per-locus coverage in CEU
~30x mean per-locus coverage in YRI

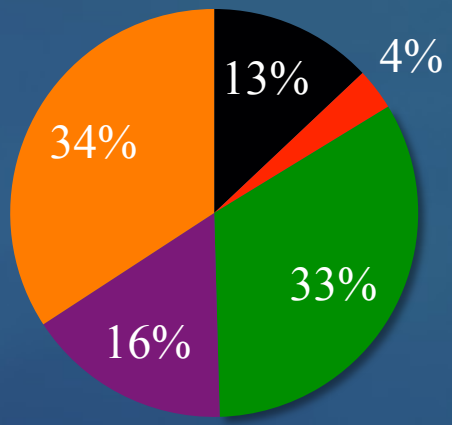
Validation Results – All Candidate Lists

CEU
N=3,236



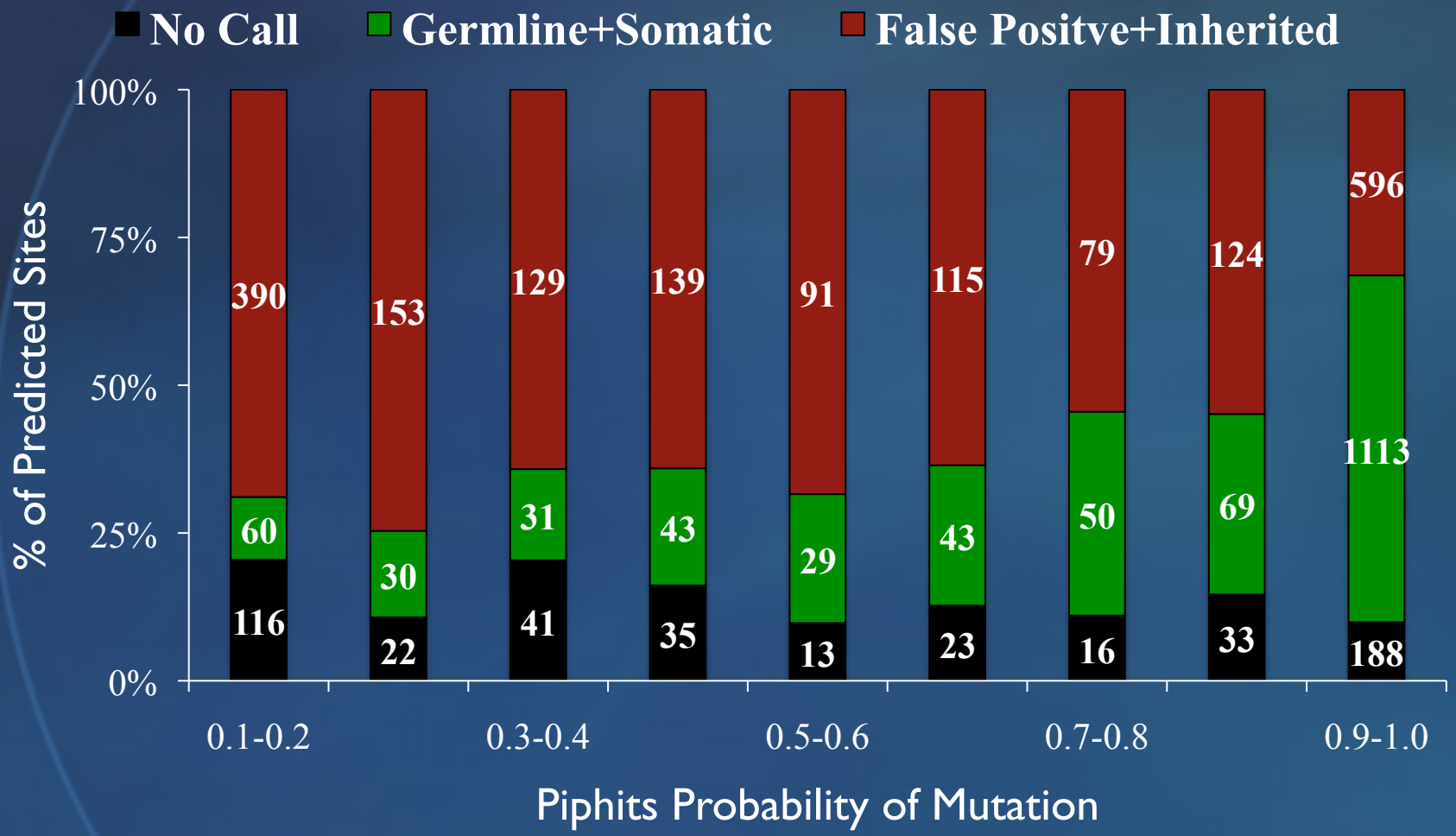
no call
germline
somatic
inherited
false positive

YRI
N=2,750



no call
germline
somatic
inherited
false positive

Utility of Piphits Probability of Mutation



Germline Mutation Rate

Rates in rough agreement between trios

CEU: 2.19×10^{-8}

YRI: 2.96×10^{-8}

Previously reported rate estimates

Haldane, 1935:
~2×10^{-8} (based on hemophilia)

Crow, 1993, 1995:
~ 2×10^{-8}

Kondrashov & Crow, 1993:
~ 2×10^{-8}

Nachman & Crowell, 2000:
~ 2.5×10^{-8}

* Direct estimates
Kondrashov, 2003:

Somatic Mutations

Majority of validated Mendelian errors due to somatic mutation

CEU: 1,052 of 1,141

YRI: 941 of 1,050

Number of somatic mutations suggests two possible findings

Cell cultures are highly mutagenic environments

Somatic mosaicism in humans is common

Total Number of *De novo* mutations in ~500 Mb of DNA surveyed from ASD and SCZ…

- 15 mutations are validated as *de novo* (germ line or somatic) in blood samples

- 6 muts in ASD, 8 in SCZ – including 2 nonsense mutations

2 splice site deletions, 2 frameshift

Gene	Sample and Ref.	Diagnosis	Mutation Type	Chr.	Position	Nucleotide Change & Genomic Context	Amino acid /structural change	CpG	CpG Island	MAPP *p*-value
SHANK3	S00004[a]	Autism Disorder	INDEL	22	49500342	CGAGATTAGC**(G/-)**TAAGGGCCAC	Splice site delG	--	--	--
IL1RAPL1	S00015[a]	Asperger Syndrome	INDEL	X	29869731	CTTGGTGCTA**(TACTCTT/-)**GCTGCTTGTA	I367SfsX6	--	--	--
GSN	S00099[a]	Asperger Syndrome	INTRONIC	9	123104277	GTGAGGCTGG**(C/G)**CCTGCCCAGC	Within intron	YES	NO	--
KLC2	S00036[a]	Autism Disorder	MISSENSE	11	65788196	TACTATCGGC**(G/C)**GGCACTGGAG	R349P	YES	NO	0.001
KIF5C	S00044[a]	Autism Disorder	MISSENSE	2	149575030	GGACCGTAAG**(C/T)**GCTACCAGCA	R802C, R872C	YES	YES	0.001
FLJ16237	S00096[a]	Autism Disorder	MISSENSE	7	15393678	CCATCACTTA**(T/C)**TTTCCATATG	F279L	NO	NO	0.472
NRXN1	S02959[b]	Schizophrenia	INDEL	2	50002821	CAGCACACGG**(-/ACGG)**GTATGGTCGT	G1402DfsX29	--	--	--
MAP2K1	S00237[e]	Schizophrenia	INTRONIC	15	64561310	CTTCTTGTAC**(G/T)**GTCAGGGAGA	Within intron	NO	NO	--
SHANK3	S00161[d]	Childhood Onset Schizophrenia	MISSENSE	22	49484091	GCATGACACA**(C/T)**GGCCTGGTGA	R552W	YES	NO	P=0.051
GRIN2B	S05650[b]	Paranoid Schizophrenia	MISSENSE	12	13611351	CTTCTACATG**(T/G)**TGGGGGCGGC	L825V	NO	NO	P<0.001
SHANK3	S00285[c]	Schizoaffective, mild MR	NONSENSE	22	49506476	TGCCCGAGAG**(C/T)**GAGCTCTGGC	R1133X	YES	YES	--
KIF17	S00215[e]	Schizophrenia	NONSENSE	1	20886681	GGAGCAGATA**(C/A)**TTCCTGGATG	Y575X	NO	NO	--
BSN	S00237[e]	Schizophrenia	SYN	3	49666988	GCACTGCAGT**(G/C)**GTAGACCTCC	V1665V	NO	NO	--
ATP2B4	S00182[b]	Disorganized Schizophrenia	SYN	1	201935404	TCATCCGAAA**(C/T)**GGTCAACTCA	N195N	YES	NO	--
SHANK3	S04261	QNTS - unknown	MISSENSE	22	49507364	GCCACCAGTG**(C/T)**CTCCCAAGCC	P1429S	NO	NO	P=0.107

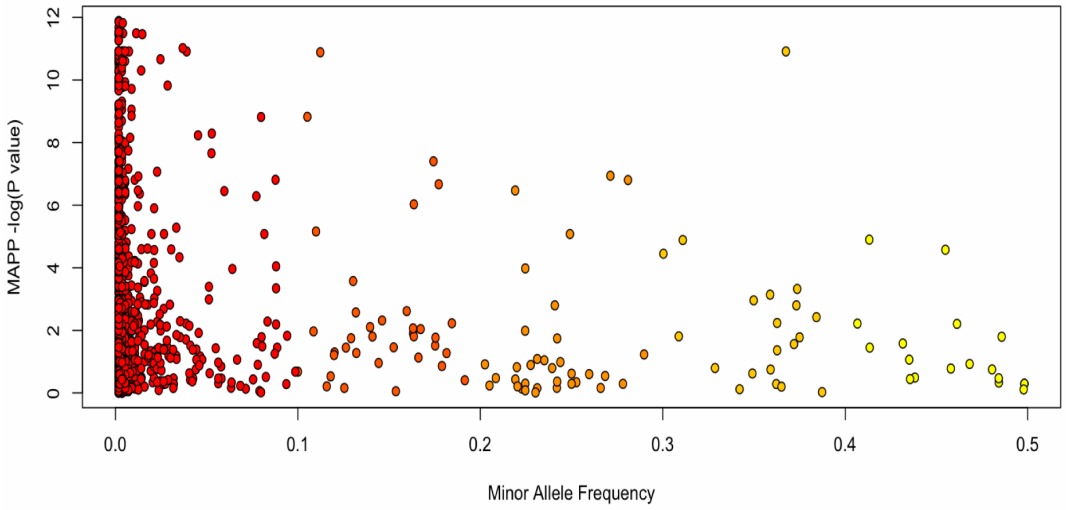
De Novos are deleterious in ASD and SCZ?

Comparison of
De novo Mutations
to Seg. sites

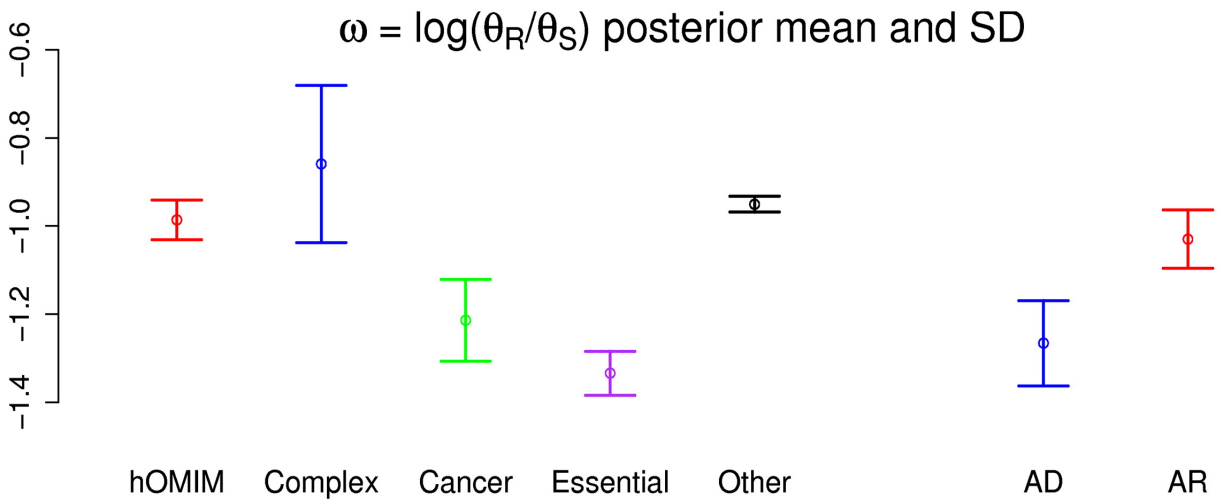
	Functional/Silent Ratios	p-value
De novos	2.5	
Singletons	0.48	0.003*
All segregating	0.27	<0.001*

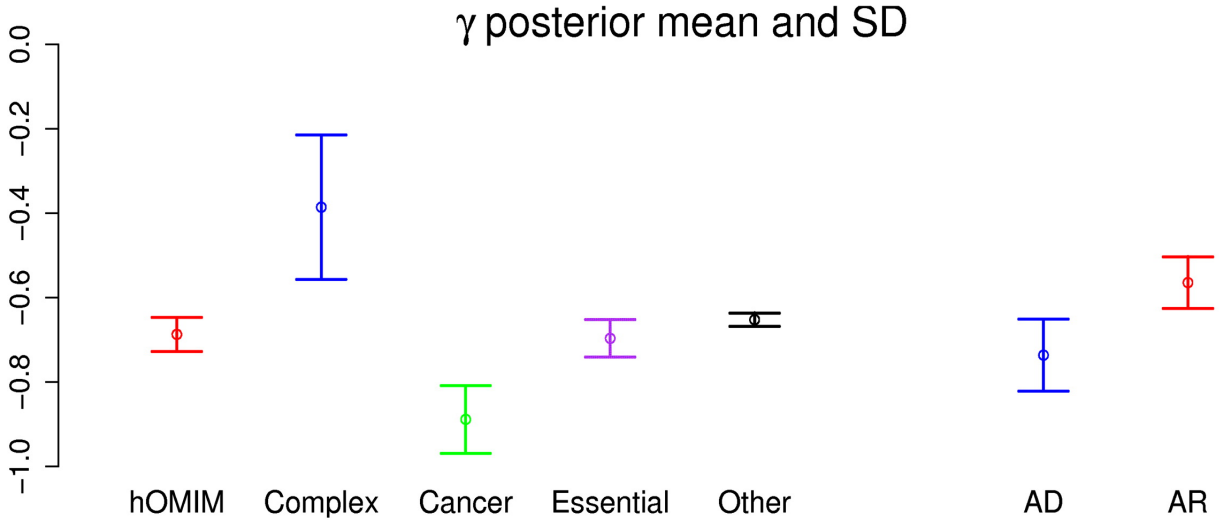
Protein Disruption Predictions (MAPP) of Missense Mutations

-Rare & De novo muts are most disruptive



Selection Parameters Estimated from Polymorphism and Divergence Data for Disease Genes





Complex	Fixed	Segregating
Silent	545.8	51
Replacement	395.6	49

Odds Ratio: 0.7545 (95% CI: 1.1663 – 0.4885)

Other	Fixed	Segregating
Silent	114515.7	9879
Replacement	57440.4	8885

Odds Ratio: 0.5576 (95% CI: 0.5749 – 0.5409)

Essential	Fixed	Segregating
Silent	11492.5	1023
Replacement	3943.6	636

Odds Ratio: 0.5519 (95% CI: 0.6142 – 0.4961)

hOMIM	Fixed	Segregating
Silent	10912.8	961
Replacement	4979.7	798

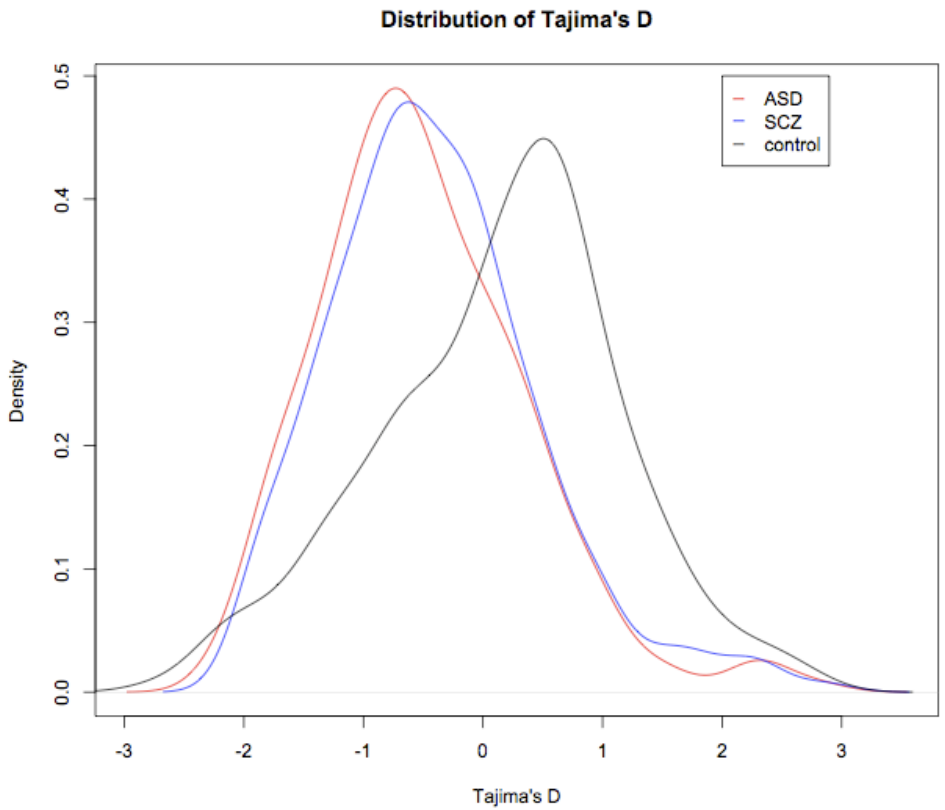
Odds Ratio: 0.5494 (95% CI: 0.6079 – 0.4967)

Cancer	Fixed	Segregating
Silent	2665.3	213
Replacement	872.8	173

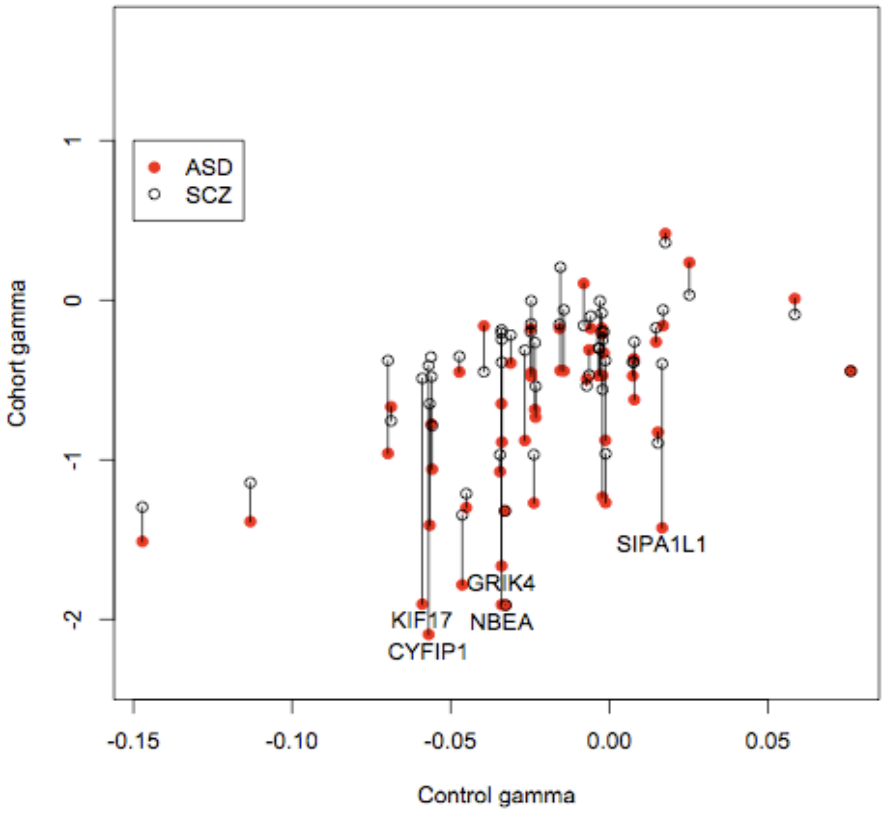
Odds Ratio: 0.4029 (95% CI: 0.5027 – 0.3233)

ASD and SCZ enriched for rare deleterious missense mutations

Across all genes



Per gene inference of selection, cases vs. controls



Plan for Module 21

Wednesday 6/23	1:30-3:00	Introduction	Philip
	3:30-4:00	Introduction (continued)	Philip
	4:00-5:00	Introduction	Mary
Thursday 6/24	8:30-10:00	Recombination	Philip
	10:30-12:00	Recombination practical	Philip
	1:30-3:00	Population size and structure	Mary
	3:30-5:00	Gene flow practical	Mary
	5:00-6:00	Tutorial	Mary/Philip
Friday 6/25	8:30-10:00	Selection	Philip
	10:30-12:00	Selection practical	Philip
	1:30-3:00	Applications and study design	Mary
	3:30-5:00	Coalescent practical	Mary

Details–Friday 6/25

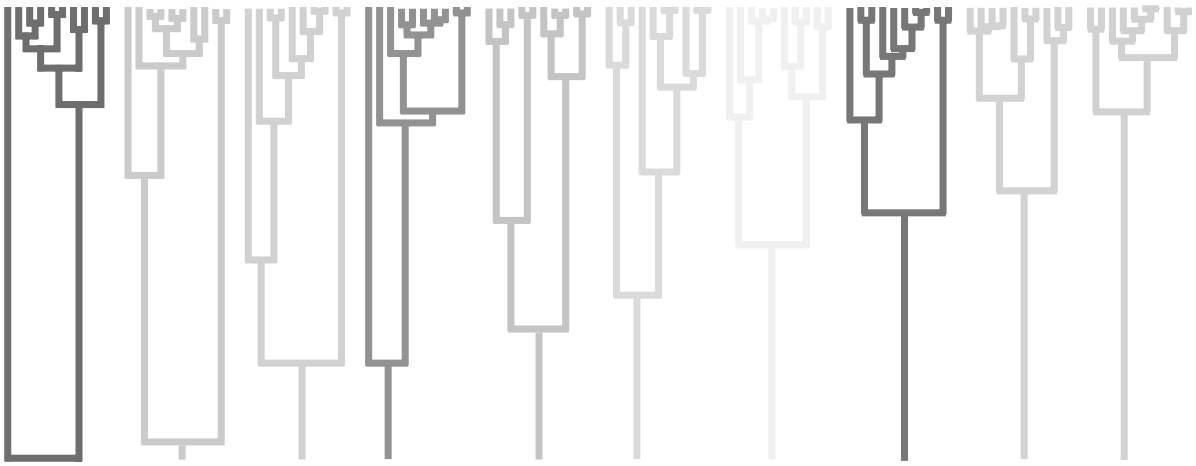
- Friday morning: Selection
  - Phylogenetic approaches
  - Population genetics approaches
  - Coalescent approaches
  - Hands-on selection exercise
- Friday afternoon: Applications of the Coalescent
  - Study design
  - Limits of applicability
  - Validation
  - Hands-on study fine-tuning exercise

Information content of the coalescent

What can best give us more information?

- More individuals?
- More base pairs?
- More loci?

Variability of the coalescent



10 coalescent trees generated with the same population size, $N = 10,000$

Variability of mutations

Does adding more individuals help?

The bottom line

- The information content of a single locus is limited
- Additional sequence length or individuals are only mildly helpful
- Multiple loci allow the best estimates
- If recombination is present, long sequences can partially substitute for multiple loci
- Multiple time points can also help, if significant evolution happens between them

Genealogy samplers to consider

- LAMARC (http://evolution.gs.washington.edu/lamarc.html)
  - Kuhner, Beerli, Felsenstein et al.
  - Estimates:
    - Population size x mutation rate
    - Immigration rates
    - Growth rates
    - Times of divergence
    - Overall recombination rate
  - Likelihood or Bayesian analysis
  - DNA, RNA, SNPs, microsats, elecrophoretic alleles

Genealogy samplers to consider

- MIGRATE (http://popgen.csit.fsu.edu/Migrate-n.html)
  - Beerli
  - Estimates:
    - Population size x mutation rate
    - Immigration rates
    - Tests among different migration models
  - Likelihood or Bayesian analysis
  - DNA, RNA, SNPs, microsats, elecrophoretic alleles

Genealogy samplers to consider

- BEAST (http://evolve.zoo.ox.ac.uk/beast/)
  - Drummond and Rambaut
  - Estimates:
    - Overall population size x mutation rate
    - Overall growth rate
    - With sequential samples, mutation rate and generation time
    - Detailed skyline plots of growth rate
    - Relaxed molecular clock
  - Bayesian analysis
  - DNA, RNA, amino acids, codon data

Genealogy samplers to consider

- IM, IMa, IMa2 (http://lifesci.rutgers.edu/ heylab/HeylabSoftware.htm#IM)
  - Nielsen, Hey, Wakeley et al.
  - Estimates:
    - Population size x mutation rate
    - Immigration rates
    - Size of ancestral populations
    - Times of divergence
    - Daughter population growth rates (IM only)
  - Bayesian analysis
  - DNA, RNA, microsatellites, HapSTRs

- IMa and IMa2 are more efficient; IM has a larger choice of models

Genealogy samplers to consider

- GENETREE (http://mathgen.stats.ox.ac.uk/software.html)
  - Griffiths et al.
  - Estimates:
    - Population size x mutation rate
    - Exponential growth rate
    - Time of most recent common ancestor
    - Times of significant mutations
  - Likelihood analysis (independent genealogies)
  - DNA (infinite sites)

Genealogy samplers to consider

- InferRho (http://sourceforge.net/projects/inferrho)
  - Rannala, Yang and Padhukasahasram
  - Estimates:
    - Recombination rate
    - Gene conversion rate
    - Hotspots
  - Bayesian analysis (correlated genealogies)
  - DNA, SNPs

Useful review paper

Kuhner, MK (2008) Coalescent genealogy samplers: windows into population history. TREE 24:86-93.

The field has changed significantly since 2008 but this review is still somewhat useful.

Designing a study

- What kind of data are available?
- What do you want to know?
- What are the pluses and minuses of available techniques?
- What is expected practice in your subfield?
- How much time do you have?

Designing a genealogy sampler study

- Major questions:
  - Should I be doing this analysis at all?
  - What model should I use?
  - How much data do I need?
  - How long do I have to run the program? (Are we done yet?)

When is a coalescent analysis inappropriate?

Things you can determine in advance:

- Randomly sampled population data are not available
  - A sample of one HIV sequence from each serotype is not usable
  - Data assigned to populations by genetic analysis can't be used to infer migration rates of those populations
- No believable mutational model is available
  - RFLPs
  - AFLPs
  - Insertion/deletion
  - Gene order

When is a coalescent analysis inappropriate?

Things that emerge from analysis:

- Data are too far outside available population models
  - Extremely rapid population change
  - Extremely non-neutral evolution
  - Extremely non-constant gene flow or recombination
- Time-scale of interesting events is much longer or shorter than the organism's coalescent time (approx. $4N_e$ generations)

Some doubtful attempts

- What is the rate of horizontal gene transfer between bacteria and plants?
- How fast did the HIV epidemic spread in the Middle East?
- What is the effective population size of pre-cancer cells in the esophagus?

When is a particular parameter not inferrable?

Θ

- Data should not be invariant
- Data should not be saturated (unalignable)
- Population must be old enough:
  - Expected depth of tree is Θ
  - If population much younger than that, little information on its size
- Unacknowledged recombination or selection can obscure answer
- Don't forget that a big linked locus like mtDNA is still only one locus

When is a particular parameter not inferrable?

Growth rate

- For exponential growth, $4N_e g$ is key parameter
- $4N_e g >> 1$ leads to star phylogenies with little information
- $4N_e g << 0$ can lead to infinite TMRCA! (I don't know what the exact cutoff is)

When is a particular parameter not inferrable?

Migration rate

- If $4N_e m$ much greater than 1, populations homogenize and gene flow hard to measure
- If $4N_e m$ very low migration events are so rare their frequency can't be estimated well
- Very high recombination weakens evidence of migration (haplotypes are too short)

When is a particular parameter not inferrable?

Divergence time

- Needs to be more recent than MRCA (approx $4N_e$ generations
- Very recent divergence not visible (less than 1/10 of this?)
- High gene flow destroys ability to infer divergence (certainly $4N_e > 1$ will have little or no power)

A cautionary tale

Abdo, Crandall and Joyce 2004

- Simulation studies to test inference of migration
- Three Θ values, four M values
- Under many circumstances inference was very poor

A cautionary tale

I resimulated Abdo et al's data:

- Low Θ with high M had no variable sites
- Low M never had more than one (obligatory) migration per tree
- High Θ with low M had mutationally randomized data
- Only a few parameter combinations led to data that could be analyzed at all

A cautionary tale

- Easy mistakes to make (I have made them too)
- Meaningful biological range of these parameters can be narrow
- Bear this in mind when:
  - Choosing types of data
    - If DNA sequences nearly invariant, consider microsatellites
  - Choosing priors
  - Designing simulations

A caveat

- The rest of this talk will focus on genealogy samplers
- Data demands are different for other types of analysis:
  - Allele frequency estimation needs a bigger sample
  - Inference based on infinite-sites needs a low mutation rate
- In general, if data are not rich enough for a genealogy sampler, they are not very informative with any method
- Using both sampler and non-sampler methods is good (remember the red drum study)

What model should I use?

- Mutational models
  - Nucleotide sequences
  - Microsatellites
  - Others
- Population models
  - Growth
  - Migration and subpopulation structure
  - Recombination

What mutational model should I use?

- Nucleotide sequences
  - Optimize model using MODELTEST (Posada and Crandall)
  - Use most nearly optimal model available in your chosen software
  - Using a more complex model will probably not help
  - If sequences are short:
    - Fix mutational parameters at published values
    - Or values from other samples from your organism
    - Or, failing that, from closely related organisms

What mutational model should I use?

- Microsatellites
  - Single-step model probably best available
  - K-Allele model overstates chance of large changes
  - LAMARC offers a mixed model but it is not validated well yet
- Others
  - BEAST offers codon and protein models
  - Codon model best for coding sequence–but SLOW
  - K-Allele model generally useful for unusual types of data

What population model should I use?

- Growth
  - Several programs offer exponential growth
  - Real populations do not grow exponentially forever
  - BEAST offers Bayesian skyline plots, but poor resolution without multiple time-point sampling
  - If growth is very recent, a no-growth analysis will perform better

What population model should I use?

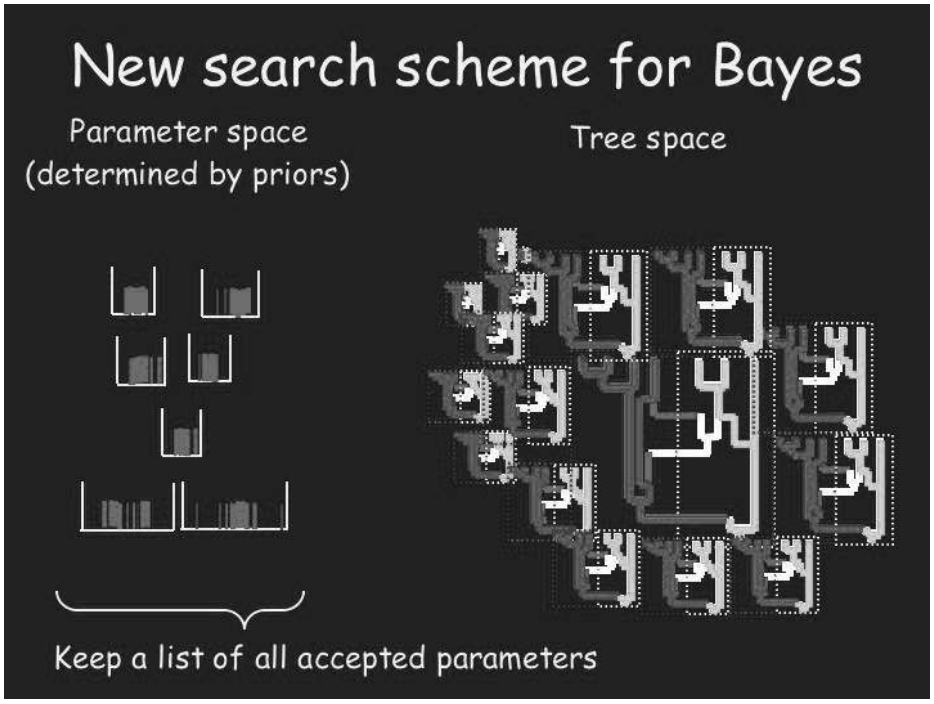
- Defining populations
  - Programs do not perform well unless populations have some structure
  - STRUCTURE (Pritchard) is useful in deciding whether to pool populations
  - Do not use STRUCTURE to assign individuals to subpopulations and then analyze them as if they belonged there!
  - (This has the effect of sending migrants back home....)
- How many populations?
  - More than 2-3 populations too many unless many loci available
  - For cases with many populations, try symmetrical migration rates and/or constrain unneeded rates to 0
- How many parameters to estimate?
  - MIGRATE offers tests based on AIC to help weed out unneeded parameters

What population model should I use?

- Recombination
  - Risky to ignore recombination if it is present
  - "Casting out" apparent recombinants biases Θ downward
  - Four-gamete test can tell whether it is dangerous to disregard recombination
  - If combining recombining and non-recombining loci (eg mtDNA and nuclear DNA) prefer a recombinant analysis
  - May be able to ignore recombination for very short sequences

Bayesian versus likelihood samplers

- Likelihood sampler:
  - Sample genealogies according to driving values
  - Estimate parameter values from stored genealogies
  - Replace driving values with estimates and repeat until satisfied
- Bayesian sampler:
  - Sample genealogies according to current parameter values
  - Sample parameters from prior according to current genealogies
  - Estimate parameter values from histogram of values visited



Bayesian versus likelihood samplers

Which to prefer?

- Kuhner MK, Smith LP, 2007. Comparing likelihood and Bayesian coalescent estimation of population parameters. Genetics 175: 155-165.
- Conclusion: no substantial difference
- Beerli P, 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 22: 341-345
- Conclusion: Bayesian is superior when data are sparse and number of parameters is high

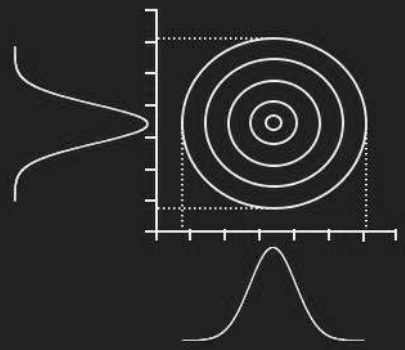
Bayesian versus likelihood samplers

- Likelihood method may have biased (too narrow) confidence intervals when:
  - True parameter value very close to zero
  - Driving values far from truth (run more chains!)
  - Sample of trees inadequate (run more steps!)
  - Data are sparse

Bayesian versus likelihood samplers

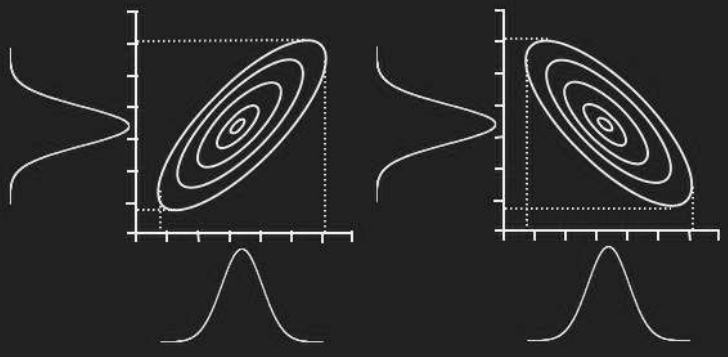
- Bayesian method may be biased when:
  - Prior not appropriate: too narrow, too wide, excludes truth
- In sparse data cases you may appear to get more information from Bayesian than likelihood **because of information in your prior**
- This is only good if your prior is well-founded
- Current Bayesian implementations lose information about correlation among parameters which is available with likelihood

Loss of Correlation Information

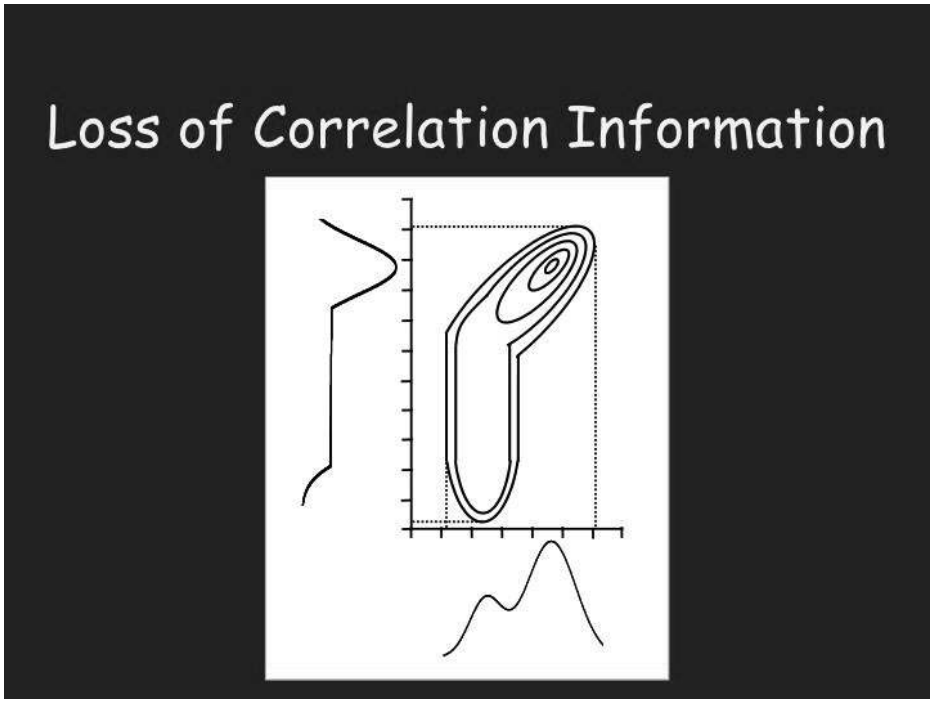


39

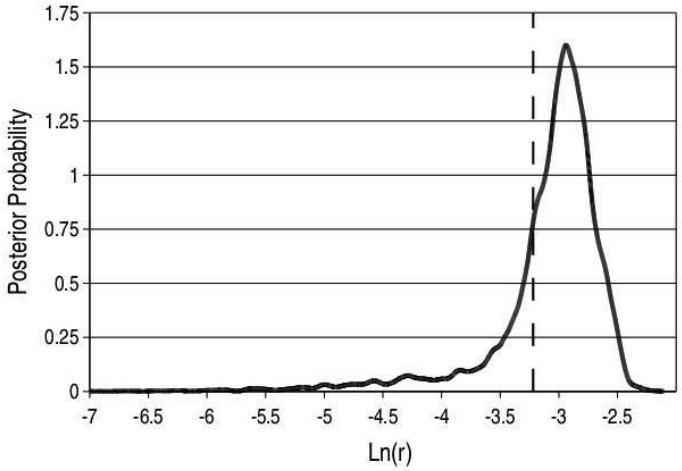
Loss of Correlation Information



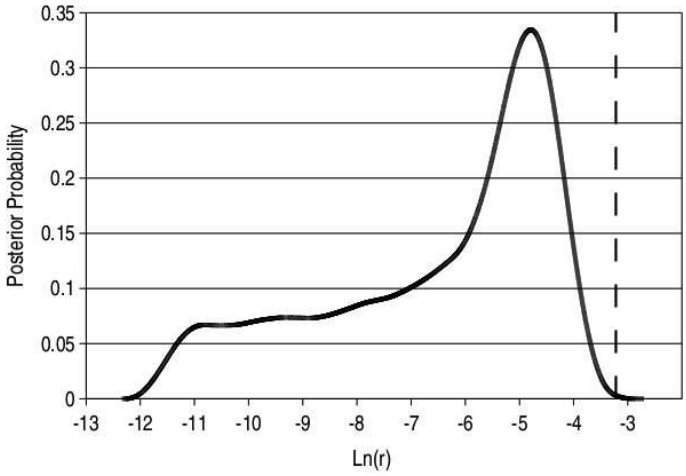
40



Nice outcome: Curve mainly reflects underlying data



Not so nice outcome: Curve strongly influenced by prior



Bayesian versus likelihood samplers

Which to use?

- When things are working well, methods are very similar
- Practical choice often based on software availability
- A Bayesian analysis with well founded prior is probably best
- If priors very unclear, prefer likelihood
- Both methods need adequate number of sampled genealogies!
- Speed difference not substantial

How much data do you need?

- For analyses without recombination:
  - Several unlinked loci are best
  - 2-3 unlinked DNA loci or 5-10 microsats can give reasonable results
  - Multiple time points or long sequences with recombination can compensate for lack of unlinked loci
  - No more than 20-25 samples per population needed
  - Ideally DNA sequences should have at least 10-15 variable sites
  - If polymorphism is low, longer sequences are needed

How much data do you need?

- For analyses with recombination:
  - A single locus can work if it's long
  - Length needed depends on polymorphism level
  - For human DNA levels, 20 KB is a good size
  - Multiple loci are still good if not too short
  - No more than 20-25 samples per population
  - This will take **much longer**

How much data do you need?–Citations

- Pluzhnikov, A. and Donnelly, P. (1996) Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144, 1247-1262

- Felsenstein, J. (2006) Accuracy of coalescent likelihood estimators: do we need more sites, more sequences, or more loci? Mol. Biol. Evol. 23, 691-700.

How much data do you need?

- If you are data-starved:
  - Reduce the number of parameters
  - Do several runs and compare results
  - Pay careful attention to confidence intervals
  - Don't expect the world!

How long to run?

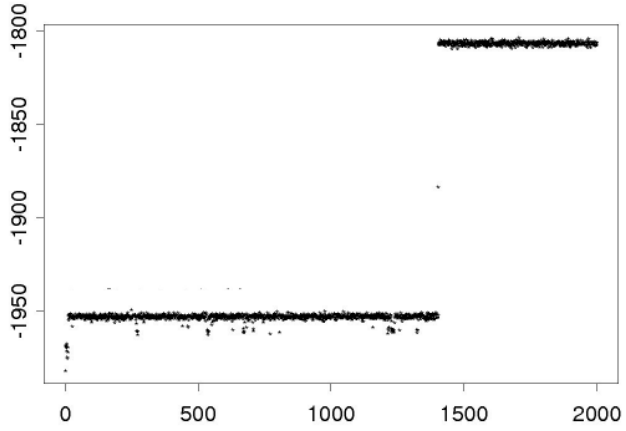
- Some general principles:
  - Results should be broadly similar if program is re-run
  - Longer runs needed for good confidence intervals
  - If run is too short, confidence intervals may exclude the truth
- These programs require informed use
- "Black box" application will lead to misleading results
- Publications must give details of run conditions

Program takes forever to run

- You may be asking too much
- Try restricting your migration model
- Try randomly removing some individuals
  - More than 20 individuals per population doesn't help much
  - Don't systematically remove similar sequences!
- Borrow a faster computer with lots of memory
- Break analysis into parts that can be run separately
- (MIGRATE only) Use several computers in parallel
- (BEAST, soon to be others) Run calculations on graphics card!

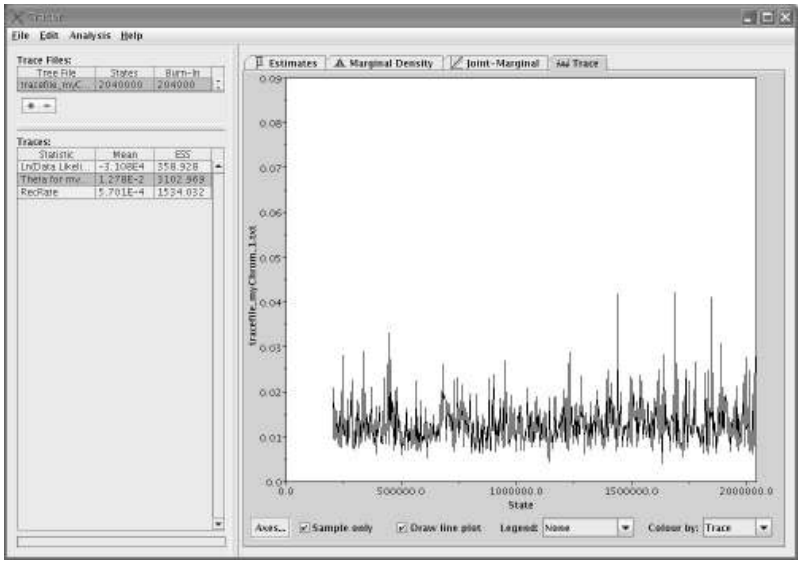
Has the run converged?

- Success can be measured as convergence
- However, a stuck search may appear to converge



Courtesy of Elizabeth Thompson

TRACER analysis



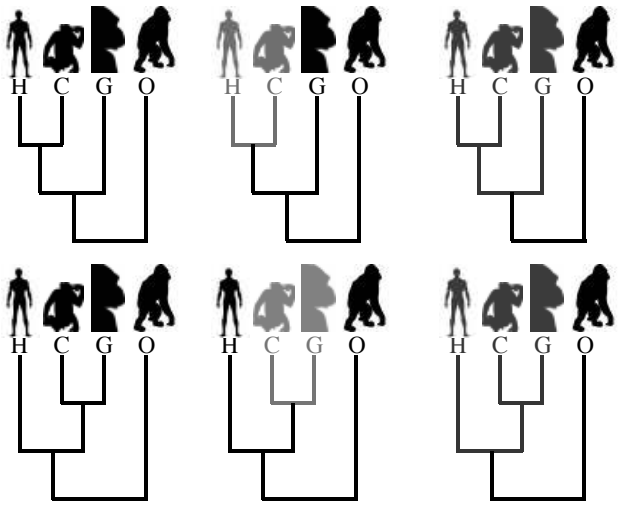
TRACER analysis

- TRACER program of Rambaut and Drummond
- Traces of parameter values over time
- Histograms of posterior probabilities
- ESS (Effective Sample Size) statistic
- Compatible with BEAST, LAMARC, MIGRATE
- IM/IMa have similar functions built in

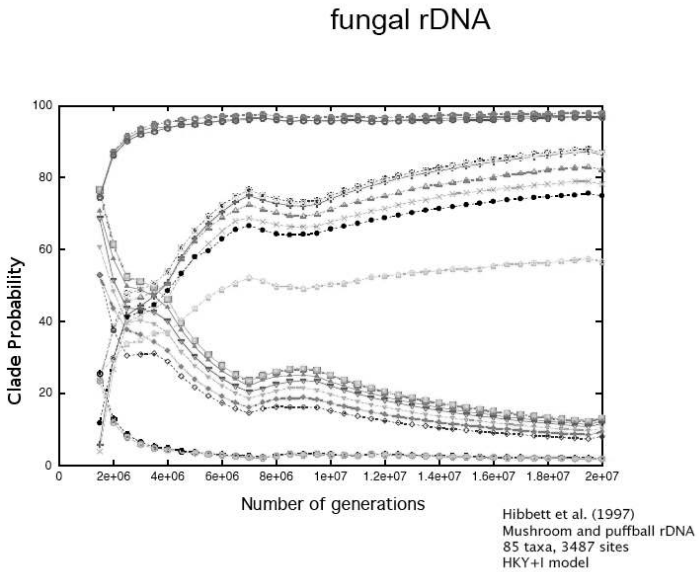
Effective Sample Size

- Effective Sample Size (ESS) corrects sample size for autocorrelation
- ESS = runlength divided by autocorrelation time
- Low ESS is strong evidence of a too-short run
- Unfortunately, high ESS does not guarantee convergence

Clade probabilities with AWTY



Convergence for clade probability

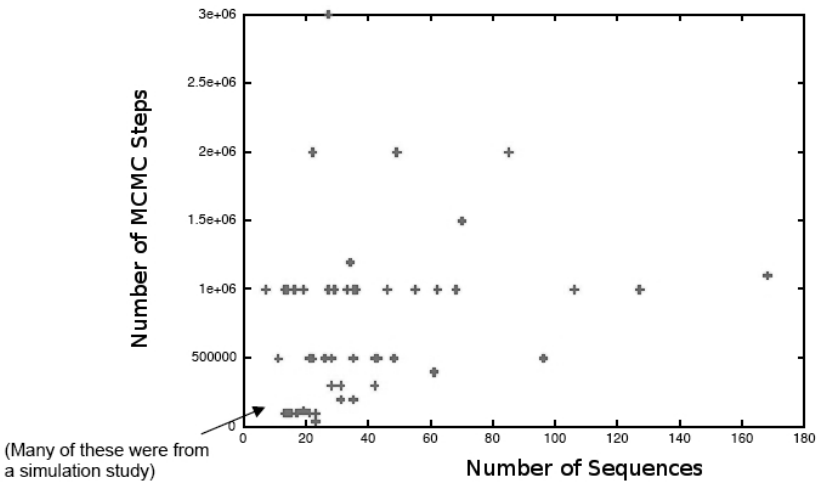


Courtesy of David Swofford, AWTY program

How long are people running their chains?

Literature search for chain lengths used with MrBayes:

- Molecular Biology and Evolution (17 papers)
- Molecular Phylogenetics and Evolution (33 papers)
- Taxon (4 papers)

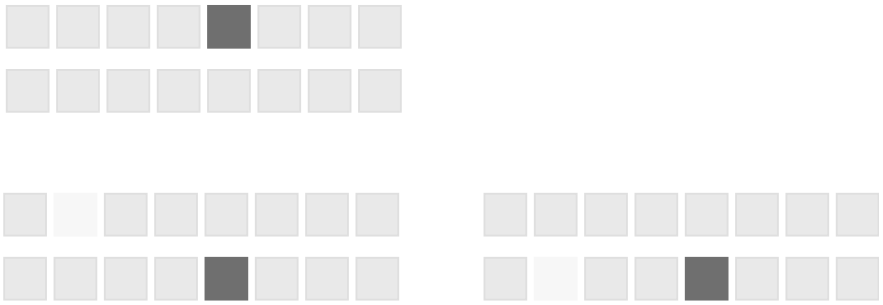


Courtesy of David Swofford, circa 2004

boa: R package for MCMC convergence assessment

- Tests for convergence of Bayesian analyses
- Not yet in common use for geneology samplers, but probably should be!
- Smith, BJ (2007) J Statistical Software 21.
- http://www.public-health.uiowa.edu/boa/

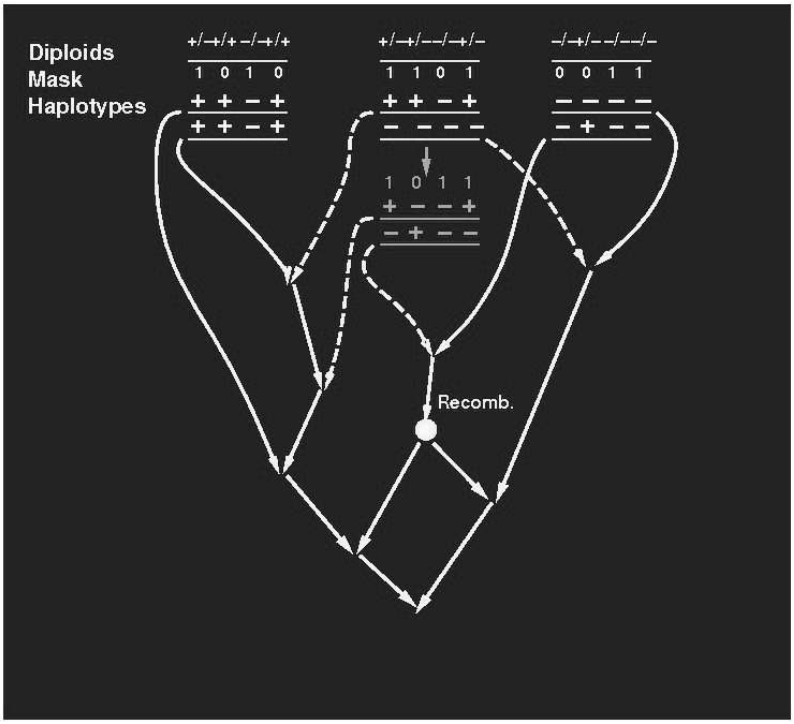
A troublesome example: phase inference



A troublesome example: phase inference

- Some data lack phase information
- Inferring "one best phase" may lead to bias
- MCMC can search simultaneously over:
  - Trees based on current phase assignment
  - Phase assignment based on current tree

A troublesome example: phase inference



A troublesome example: phase inference

Strategy	$\frac{Estimated\Theta}{True\Theta}$
Correct haplotypes	1.0
No haplotype inference	1.65
Haplotype reassignment 10%	1.28
Haplotype reassignment 20%	1.23
Haplotype reassignment 50% (10x search)	1.15
Reassignment with rearrangement	1.33

A troublesome example: phase inference

Strategy	$\frac{Estimated\Theta}{True\Theta}$
Correct haplotypes	1.0
No haplotype inference	1.65
Haplotype reassignment 10%	1.28
Haplotype reassignment 20%	1.23
Haplotype reassignment 50% (10x search)	1.15
Reassignment with rearrangement	1.33
Haplotype reassignment 20%, heated	1.03

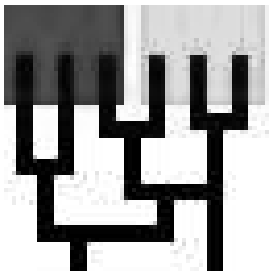
Final thoughts

- Coalescent studies should be carefully designed:
  - Data collection
  - Mutational model
  - Population model
  - Details of analysis
- The strongest studies combine multiple approaches
- Pay as much or more attention to error bars as point estimates

Thanks to

Joe Felsenstein
Peter Beerli
Jon Yamato
Lucrezia Bieler
Elizabeth Thompson
Eric Rynes
Lucian Smith
Elizabeth Walkup

Web site



http://evolution.gs.washington.edu/lamarc.html