

# MCMC background

Workshop given at MBARI  
February 10, 2006  
C. R. Young

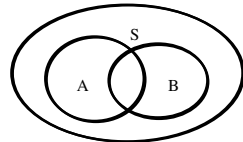
# What we'll talk about

- Quick review of conditional probability
- Bayes theorem
  - Derive from conditional prob.
  - Likelihood, Prior, Posterior
- Statistical distributions
  - Uniform
  - Normal
  - Gamma (Exponential)
  - Poisson
- Markov chains
  - Difference between prob and markov
  - Discrete time
  - One dimensional walk
  - Stationary
  - Residence times
  - Continuous time
- Coalescent theory
  - (from notes)
- Model vs. observational uncertainty
- Monte Carlo sampling

## Conditional Probability

Define:

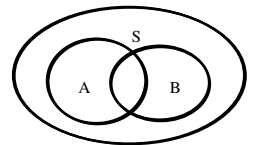
- $\Pr\{A\}$  = probability that event A occurs
- $\Pr\{A\} = (\text{area of } A) / (\text{area of } S)$
- $\Pr\{A \text{ or } B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \text{ and } B\}$



$\Pr\{A\} = (\text{area of } A) / (\text{area of } S)$   
- However area might be defined

$\Pr\{A, B\}$

## Conditional Probability



•  $\Pr\{B \text{ occurred given that } A \text{ occurred}\} = (\text{area common to } A \text{ and } B) / (\text{area of } A)$

•  $\Pr\{B|A\} = \frac{\text{area common to } A \text{ and } B}{\text{area of } A} = \frac{\text{area common to } A \text{ and } B}{\text{area of } S} / \frac{\text{area of } A}{\text{area of } S}$

$$\Pr\{B|A\} = \Pr\{A, B\} / \Pr\{A\}$$

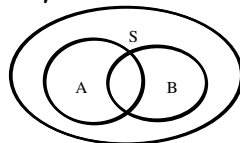
## Conditional Probability

Conditional probability

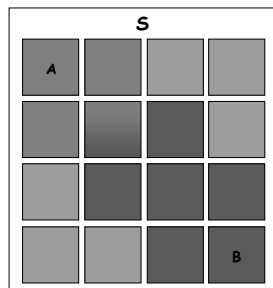
- $\Pr\{B|A\} = \Pr\{A, B\} / \Pr\{A\}$
- $\Pr\{A|B\} = \Pr\{A, B\} / \Pr\{B\}$

Rearrange for Probability of A and B

- $\Pr\{A, B\} = \Pr\{B|A\} \Pr\{A\}$
- $\Pr\{A, B\} = \Pr\{A|B\} \Pr\{B\}$

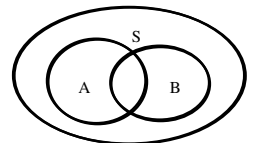


Event	Area
S	16
A	4
B	8
A&B	1



- $\Pr\{B|A\} = \Pr\{A, B\} / \Pr\{A\}$
- $\Pr\{B|A\} = (1/16) / (4/16)$
- $\Pr\{B|A\} = 0.25$

## Independence



• Events are independent if:

- $\Pr\{A|B\} = \Pr\{A\}$
- $\Pr\{B|A\} = \Pr\{B\}$
- i.e., information that A (or B) occurred does not tell you anything about the probability of B (or A) occurring

$$\Pr\{B|A\} = \Pr\{A, B\} / \Pr\{A\}$$

(rearrange)

$$\Pr\{A, B\} = \Pr\{B|A\} \Pr\{A\}$$

(substitute)

$$* \Pr\{A, B\} = \Pr\{B\} \Pr\{A\}$$

\* This equation is often given as the definition of independent events, but it is actually derived from the definition based on conditioning...

## Bayes' Theorem

- Likelihood (classical statistical analyses)
  - Pr{data | parameters}
  - Parameters are fixed
  - Data is a random quantity
- But we have *this* data
- And our *parameters* seem to be uncertain
- How do we get:
  - Pr{parameters | data}

## Bayes' Theorem

- Pr{data | parameters}
  - Say  $Y = \{\text{collection of data}\}$
  - Say  $\Theta = \{\text{collection of parameters}\}$
- Pr{Y |  $\Theta$ } is likelihood
- We know:
  - $\Pr\{A, B\} = \Pr\{A|B\} \Pr\{B\}$
  - $\Pr\{B|A\} = \Pr\{A, B\} / \Pr\{A\}$

## Bayes' Theorem

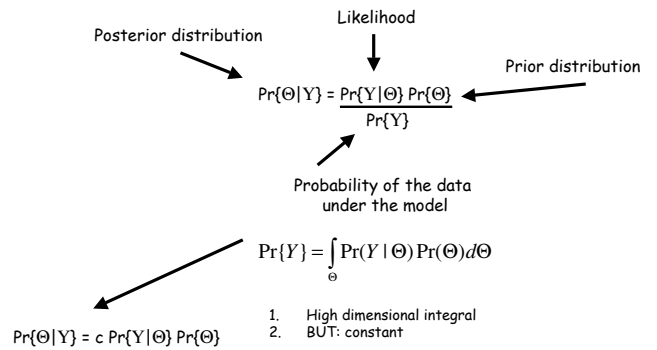
- Pr{data | parameters}
  - Say  $Y = \{\text{collection of data}\}$
  - Say  $\Theta = \{\text{collection of parameters}\}$
- Pr{Y |  $\Theta$ } is likelihood
- We know:
  - $\Pr\{A, B\} = \Pr\{A|B\} \Pr\{B\}$
  - $\Pr\{B|A\} = \Pr\{A, B\} / \Pr\{A\}$

Bayes Theorem

$$\Pr\{B|A\} = \frac{\Pr\{A|B\} \Pr\{B\}}{\Pr\{A\}}$$

$$\Pr\{\Theta|Y\} = \frac{\Pr\{Y|\Theta\} \Pr\{\Theta\}}{\Pr\{Y\}}$$

## Bayes' Theorem

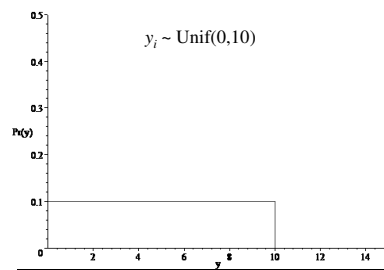


## Statistical distributions

- Uniform
- Normal
- Gamma/Exponential family
- Poisson

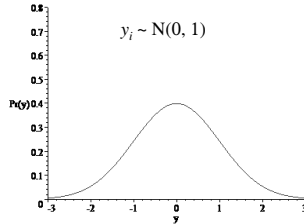
## Uniform Distribution

- $y_i \sim \text{Unif}(\min, \max)$
- Lives on all real numbers



## Normal Distribution

- $y_i \sim N(\mu, \sigma)$
- Lives on all real numbers

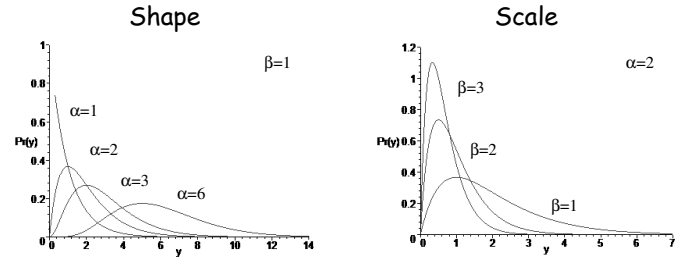


## Gamma

- $y_i \sim \Gamma(\alpha, \beta)$
- $y_i > 0$
- Lives on positive numbers

$$p(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{(\alpha-1)} e^{(-\beta y)}$$

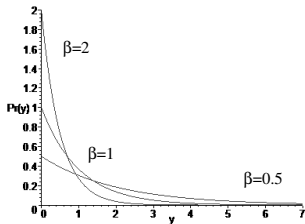
$$p(y) = c y^{(\alpha-1)} e^{(-\beta y)}$$



## Exponential (Gamma with $\alpha=1$ )

- $y_i \sim \varepsilon(\beta)$
- Lives on positive numbers

$$p(y) = \begin{cases} \beta e^{-\beta y} & \text{for } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

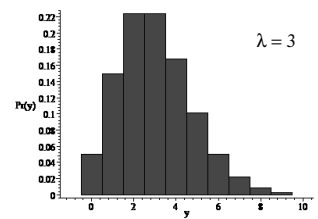
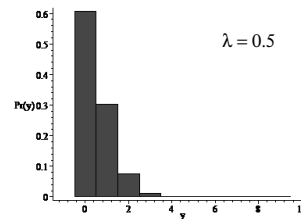


## Poisson

- $y_i \sim \text{Poisson}(\lambda)$
- Lives on integers zero or greater

$$\Pr(Y_i = y_i) = \begin{cases} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} & \text{for } y_i = 0, 1, \dots \\ 0 & \text{otherwise} \end{cases}$$

$\lambda > 0$



## Finding means of distributions

- Find the expectation (mean) of a distribution

$$E(y) = \sum_y y \Pr(y | \lambda) \quad E(y) = \int_y y \Pr(y | \lambda) dy$$

$$p(y | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{(\alpha-1)} e^{(-\beta y)}$$

## Monte Carlo Sampling

- Can also find the mean by drawing from that distribution
- Draw, say, 1000 independent and identically distributed (IID) samples

$$E(y_m) = \frac{1}{M} \sum_{m=1}^M y_m$$

$$E(y_m) = \frac{1}{1000} \sum_{m=1}^{1000} y_m$$

```
> p := (alpha, beta, y) -> beta^alpha * y^(alpha-1) * exp(-beta*y) / GAMMA(alpha);
> integrate( y*p(alpha, beta, y), y = 0..infinity);
      alpha
      ----
      beta
```

# Markov Chains

- Stochastic process
  - A collection of random variables  $\{y_t, t \in T\}$  where  $t$  indexes time
  - $T$  can be either discrete or continuous
  - $S$  can be either discrete or continuous
- A stochastic process is a Markov chain if for any  $A \in S$

$$\Pr(y_{t+1} \in A \mid y_0, \dots, y_t) = \Pr(y_{t+1} \in A \mid y_t)$$

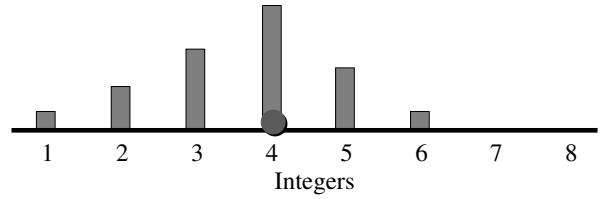
The past and future states of the process are independent given the present

# Random walk

$$\Pr(y_{t+1} \in A \mid y_0, \dots, y_t) = \Pr(y_{t+1} \in A \mid y_t)$$

- At some state in  $S$ 
  - $y_t = 4$
  - $S = \{1, 2, \dots, 8\}$

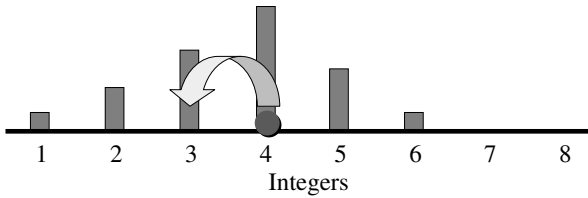
Move to a new state defined by probability:  
 $\Pr(y_{t+1} \mid y_t)$   
 Note: It does not matter how you got here, only matters that you are here



# Random walk

$$\Pr(y_{t+1} \in A \mid y_0, \dots, y_t) = \Pr(y_{t+1} \in A \mid y_t)$$

- At some state in  $S$ 
  - $y_t = 4$
  - $S = \{1, 2, \dots, 8\}$

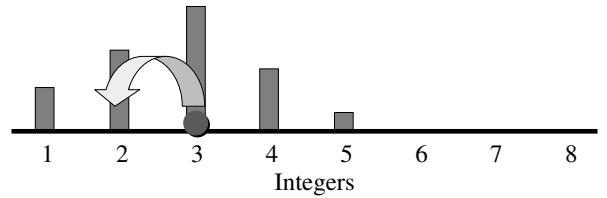


# Random walk

$$\Pr(y_{t+1} \in A \mid y_0, \dots, y_t) = \Pr(y_{t+1} \in A \mid y_t)$$

- At some state in  $S$ 
  - $y_t = 3$
  - $S = \{1, 2, \dots, 8\}$

Move to a new state:  
 $\Pr(y_{t+1} \mid y_t)$



# Random walk

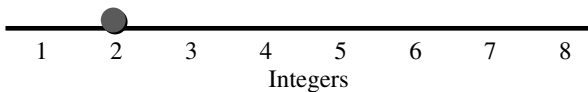
$$\Pr(y_{t+1} \in A \mid y_0, \dots, y_t) = \Pr(y_{t+1} \in A \mid y_t)$$

- At some state in  $S$ 
  - $y_t = 2$
  - $S = \{1, 2, \dots, 8\}$

$$Y = \{y_t, t \in T\}$$

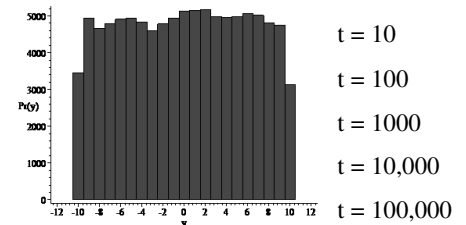
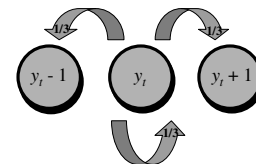
$$Y = \{4, 3, 2\}$$

Time series



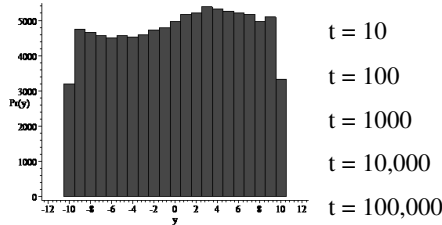
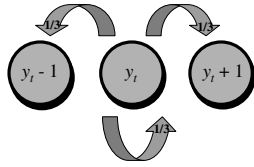
# A Simple Markov Chain

- $S = \{-10, \dots, 10\}$
- $y_0 = 0$
- $\Pr(\Delta y = -1) = 1/3$
- $\Pr(\Delta y = 0) = 1/3$
- $\Pr(\Delta y = 1) = 1/3$



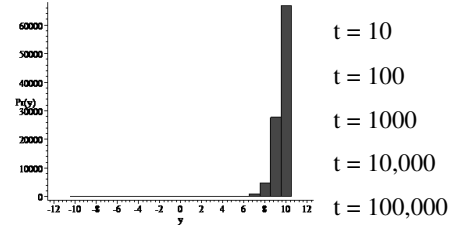
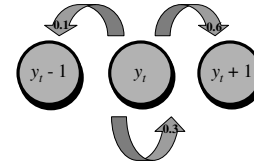
## A Simple Markov Chain

- $S = \{-10, \dots, 10\}$
- $y_0 = -10$
- $\Pr(\Delta y = -1) = 1/3$
- $\Pr(\Delta y = 0) = 1/3$
- $\Pr(\Delta y = 1) = 1/3$



## A Simple Markov Chain

- $S = \{-10, \dots, 10\}$
- $y_0 = -10$
- $\Pr(\Delta y = -1) = 0.1$
- $\Pr(\Delta y = 0) = 0.3$
- $\Pr(\Delta y = 1) = 0.6$



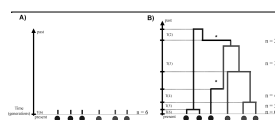
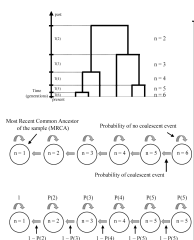
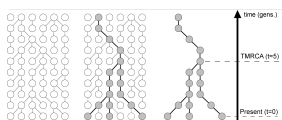
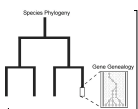
## Markov Chains

- What is the probability that the chain is in state  $y_i$ ? (stationary distribution)
- What is the probability that a chain visits some interval  $y=[a,b]$ ? (residence times)
- What is the expected state of the chain over a large amount of time? (average of the stationary distribution)
- How long do we have to wait to get from  $y=a$  to  $y=b$ ? (waiting times)

## Markov Chains: recap

- Past and future states are independent given the present
- Chain is sequence of states
  - State vector
  - Time series
- Transition probabilities given current state
- Some chains settle into an equilibrium distribution
  - Independent of initial state
  - Can be described as a probability distribution

## Coalescent Theory



## Forward thinking: the predictive approach

Random variation in **reproduction** causes random fluctuation in **allele frequencies**

Wright (1931) showed that the equilibrium distribution of the allele frequency of  $p$ , where alleles mutate:  $P \rightarrow Q$  at rate  $\mu$  and  $Q \rightarrow P$  at rate  $\nu$ , can be described by:

$$\Pr(p) = cp^{(4N_e\nu-1)}q^{(4N_e\mu-1)}$$

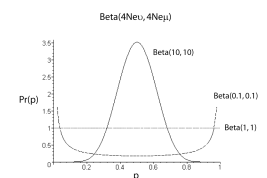
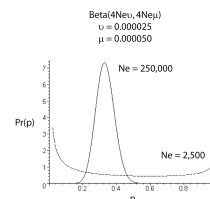
$$\theta = 4N_e\mu$$

**Drift** is more efficient in **small populations**

$\theta \ll 1$ : **drift dominates**

(pushes alleles toward fixation: 0 or 1)

$\theta \gg 1$ : **mutation prevents fixation**



# Coalescent Theory

- Model of sequence evolution that considers only sampled individuals
- For what is it used? To answer the questions:
  - Given some population parameters, what should I expect my sample to look like?
  - Given my sample, what are the most likely parameters that gave rise to my sample?

## Given some population parameters, what should I expect my sample to look like?

- Simulation of gene evolution
  - Create gene sequences under specified conditions
    - Effective populations size
    - Migration (non-random mating)
- Helps build our intuition about population genetics processes
- Statistical Power Analysis
  - Explore the behavior of statistics under different parameters
  - What is the probability of correctly rejecting a null hypothesis?
- Experimental design
  - How many individuals should I sample?
  - How many genes do I need to sample from an individual?
  - How long should each gene sequence be?

## Given my sample, what are the most likely parameters that gave rise to my sample?

- Statistical inference
  - Bayesian and Maximum Likelihood Methods
    - $P(X|\theta) \Leftrightarrow P(\theta|X)$
    - Simulations to determine likelihood surfaces
- Ultimately allows one to:
  - Extract information about the **evolutionary history** of organisms
  - Define a level of **confidence** about that inference

## What are we after?

Hopefully, our models capture the **important evolutionary processes** that produce the data that we see.

- We gain intuition about these processes by creating models of **idealized scenarios**
- **Test predictions** (or assumptions) of these models in an attempt to validate them.
- Make **inferences** about **evolutionary processes** that maintain or erode genetic variation in natural populations

As empirical population geneticists (or molecular ecologists), we are usually interested in evolutionary processes operating in **specific organisms** or groups of organisms.

- So we might like to estimate **parameters** of the model that are specific to our organism:
- Effective population size
  - Migration rates
  - Times of divergence between populations or species

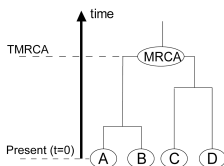
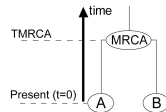
"All models are wrong, some models are useful." - George Box (1976)

## The history of coalescent theory

Gustave Malécot (in the 1940's):

- following a pair of gene copies back to **common ancestor**
- the notion of **identity by descent**:

If we pick two genes from a Wright-Fisher (WF) population, how long ago on average did the two genes share their **most recent common ancestor (MRCA)**?



Genealogical approaches to **samples larger than two** appeared in response to the first direct measurements of molecular variation (Harris 1966, Lewontin and Hubby 1966).

We might expect the TMRCA for a sample of size four to be **deeper** in the past than a sample of size two

## The history of coalescent theory

• Ewens (1972): Derived the **distribution of allele counts** in a sample under the infinite-alleles model of selectively neutral mutation

• Watterson (1975): gave an explicitly genealogical derivation of the **number of segregating sites**, or polymorphic sites, in a sample of sequences under the infinite-sites model of mutation without recombination;  $\theta_s = 4N\mu$

• Kingman (1982a,b,c): proved the existence of the **coalescent process**. Showed that the 'n-coalescent' holds for a wide range of populations with different breeding structures

• Tajima (1983): Derived the expectation of the **average number of pairwise differences** in a sample. As it turns out, this is an estimate of the composite parameter:  $\theta_d = 4N\mu$

• Reviews

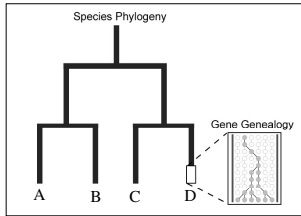
• Hudson (1990) wrote a wonderful review of coalescent theory and made available an **algorithm** to simulate data under different population models

• Fu and Li (1999)

• Nordborg (2001)

## A look at gene genealogies

- **Gene genealogies** underlie the data that we see
- Coalescent theory capitalizes on the long-standing familiarity of evolutionary biologists with **tree structures**
- For example, the only figure in *The Origin of Species* (Darwin, 1859) is a hypothetical phylogeny, i.e. a tree representing **patterns of descent among species**
- Here, gene genealogies represent **patterns of descent among genes** within species



## Genealogical vs. sampling variance

Due to the **randomness of reproduction**: even if we had perfect knowledge of the genealogy, we would still have some **uncertainty** about the parameters that we are interested in:

- $\theta = 4N_e\mu$
- $M = 4N_e m$

The variability in data that we see has two components:

1. **Genealogical variance\***  
(reduced by sampling more loci)
2. **Sampling variance**  
(reduced by sampling more nucleotides at a given locus)

## Separating the Genealogical process from the mutation process

If we assume that mutations are **neutral**, then we can separate the process that leads to gene genealogies (reproduction) from the mutational process.

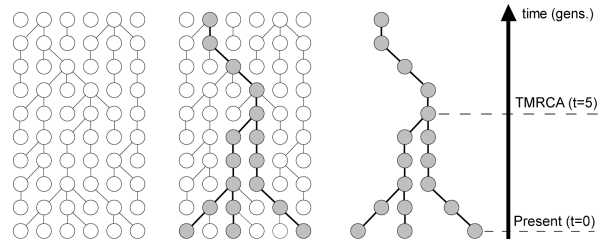
Why? By definition, a neutral mutation has **no effect** on the survival or reproduction of an individual.

In the case of neutral mutations, **demographic processes** (reproduction, migration, etc.) affect our data only through their effects on the **shapes of genealogies**

You can think of mutation as the **fuzzy lens** through which we are informed of the **genealogy**.

## A look at gene genealogies

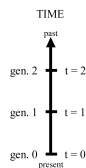
- **Forward time** approach models alleles in **populations**
- **Coalescent** approach considers **samples** of genes
- The genealogical process is **time reversible** meaning that it is the **same forwards and backwards**
- If we can describe properties of samples
  - Can ignore the rest of the population
  - Saves computational effort



## The Kingman n-coalescent (standard coalescent)

We assume that reproduction in the population follows the standard WF model:

1. Nonoverlapping generations
2. Random reproduction
3. Constant population size of  $N$  individuals
4. Random mating (no structure)



Variable	Definition
$t$	time step (the present starts at zero; see figure on the above)
$N$	haploid population size
$n$	number of ancestors present in the sample
$\Pr(n)$	probability no ancestors share a parent in the previous generation if $n$ ancestors are present
$C(t n)$	probability of a coalescent event after the $t^{\text{th}}$ generation, conditional on $n$ ancestors in the sample
$T(i)$	Time interval between coalescent events when the sample contains $i$ ancestors
$T$	Total length of the genealogy

## Important Parameters

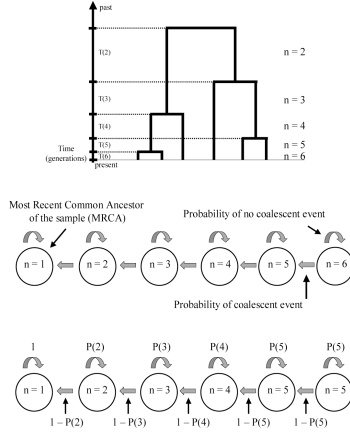
- **Haploid Wright-Fisher Population:**
  - $N_e$  = effective population size
  - $\mu$  = mutation rate (per sequence per gen.)
  - $\theta = 2N_e\mu$  (mutation, compound parameter)

For diploids:  $\theta = 4N_e\mu$   
For haploid/maternally-transmitted:  $\theta = N_e\mu$

# The Kingman n-coalescent (standard coalescent)

## Information in a genealogy

- Who's related to whom
  - Branching order
  - Which individuals share a most recent common ancestor (MRCA)
- Time\*
  - Branch lengths
    - Defined as time between pairs of sequences
    - T(i) = time during which there are i distinct lineages
  - Estimated by the number of mutations separating sequences



Start with a sample of, say, six genes

We will building a genealogy from the present time, t = 0, backwards in time until we are left with the most recent common ancestor (MRCA) of the sample.

## Finding the distribution of coalescent times

What about 3 ancestors in the sample?

We need to know the probability that none of the three share a common ancestor.

Pr(3 distinct parents) = Pr(2 distinct parents) \* Pr(3<sup>rd</sup> parent is distinct from other two)

We already know this

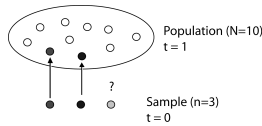
There are N-2 remaining individuals to pick as the ancestor for our 3<sup>rd</sup> individual.

Probability that the 3<sup>rd</sup> has a distinct parent from the first two (given that the first two have distinct parents) is:

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

Therefore, the probability that all three sampled individuals have distinct parents in the previous generation is:

$$\Pr(3) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right)$$



## Finding the distribution of coalescent times

First, pick a random individual from the population of size N

Now ask, what is the probability that the second randomly chosen individual shares a parent with the first?

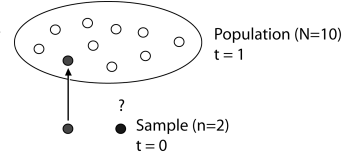
Under the WF model every individual is equally likely to be the parent of either individual

The probability that they both share the same parent is:

$$\frac{1}{N}$$

The probability that neither share a parent in the previous generation is:

$$\Pr(2) = 1 - \frac{1}{N}$$



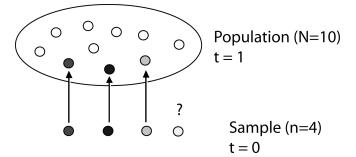
## Finding the distribution of coalescent times

We can continue to build on our previous results in this way - a pattern emerges:

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

$$\Pr(3) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right)$$

$$\Pr(4) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right) \left(1 - \frac{3}{N}\right)$$



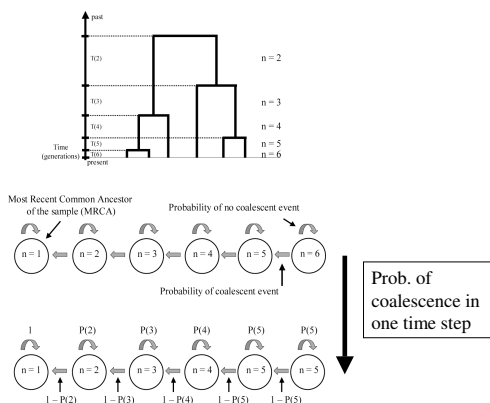
So we can make a general statement of the probability that n ancestors have n distinct parents in the previous generation:

$$\Pr(n) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{N}\right) \approx \frac{n(n-1)}{2N}$$

- Probability that no ancestors in a sample of size n share a common ancestor

Now we need to consider how long we will have to wait before the lineages coalesce...

## Recap:



## Finding the distribution of coalescent times

Since the outcome of each generation is independent, the probability that no ancestors in a sample share a common ancestor in t generations (no coalescent event) is:

$$\Pr(n)^t$$

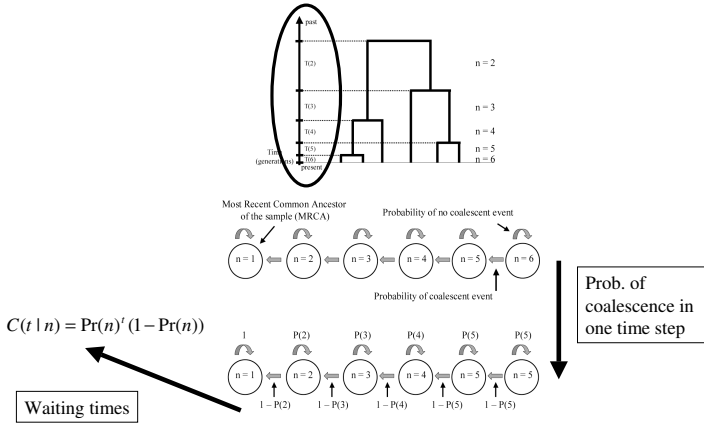
The probability, C(t | n), that there is a coalescent event in the t<sup>th</sup> generation conditional on there being n lineages in the sample is:

$$C(t | n) = \Pr(n)^t (1 - \Pr(n))$$

That is, there were no coalescent events for the first t generations, with probability Pr(n)<sup>t</sup>, then there was a coalescent event in the (t+1)<sup>th</sup> generation, with probability 1-Pr(n)

This is important: it describes the T(i)'s (waiting times)!

## Recap:



## Finding the distribution of coalescent times

This is important: it describes the  $T(i)$ 's (waiting times)!

$$C(t | n) = \Pr(n)^t (1 - \Pr(n))$$

$C(t | n)$  looks a lot like Binomial sampling:

$$p^x (1 - p)^y$$

In our case, we have:

$$p^t (1 - p)^1$$

This is a specific type of binomial sampling

- $(1-p)$  is a very small probability
- the number of trials,  $t$ , must be very large for an event of type  $(1-p)$  to occur
- Poisson process
- waiting time for a Poisson process,  $t$ , can be approximated by an exponential distribution:

$$\Pr(t) = \lambda e^{-\lambda t}$$

## Finding the distribution of coalescent times

$$\Pr(t) = \lambda e^{-\lambda t}$$

If we measure time in units of  $N$  generations, then we arrive at a general model for the times of coalescent events:

Rate of coalescent events:

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

Probability of coalescent event at time,  $t$ :

$$C(t | n) = \frac{n(n-1)}{2} e^{-\frac{n(n-1)t}{2}}$$

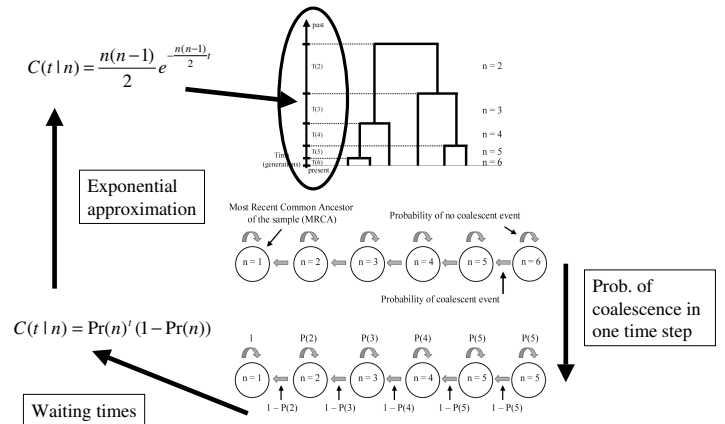
Mean time of coalescent event for  $i$  distinct lineages:

$$E[T(i)] = \frac{2}{i(i-1)}$$

Mean time of coalescent event for sample of size  $n = 2$ :

$$E[T(2)] = \frac{2}{2(2-1)} = 1$$

## Recap:



## The effect of the rescaling

Average coalescence times for  $i$  ancestors.  $E[T(i)]$ : time measured in generations;

$E[T(i)/N]$ : time scaled by population size (measured in units of  $N$  generations).

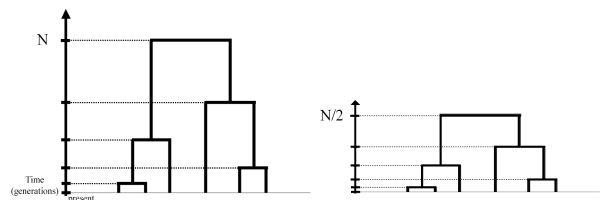
$i$	$E[T(i)]$			$E[T(i)/N]$		
	$N=100$	$N=200$	$N=1000$	$N=100$	$N=200$	$N=1000$
6	6.7	13	67	0.07	0.07	0.07
5	10	20	100	0.10	0.10	0.10
4	17	33	167	0.17	0.17	0.17
3	33	67	333	0.33	0.33	0.33
2	100	200	1000	1.00	1.00	1.00

Three bits of intuition can be gained from examining this table:

1. Coalescence times (measured in generations) are **shorter** for small  $N$
2. Coalescence times are **very short** when there are **many** ancestors ( $n \gg 2$ ) compared to when there are few ancestors
3. The **relative timings** of coalescent events are **not affected** by population size, only the total height of the genealogy

## The effect of the rescaling

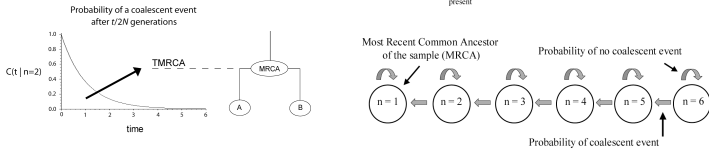
Genealogies of populations with **different effective sizes** differ only in their **scale** (height) not their **shape** (topology or relative coalescence times)



# Simulating genealogies

Under the WF model, all lineages are equally likely to share a common ancestor. Starting with an arbitrary number of genes,  $n$ :

1. Draw a coalescence time from  $C(t | n)$
2. Randomly choose 2 samples to coalesce
3. Create a parent lineage at  $T(n)$
4. Decrement  $n$   
 <Repeat until  $n = 1$ >

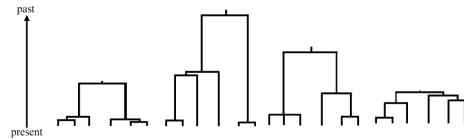


# Simulating genealogies

By repeating the simulation algorithm many times, we can get a feel for the **range of genealogies** that are probable for a sample of size  $n$ .

There is **considerable variation** among genealogies.

The figure below shows a sample ( $n = 6$ ) of four genealogies that were created with the same time scale ( $N$ ).

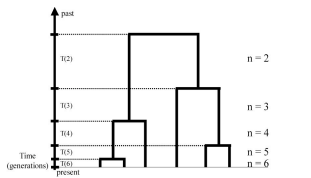
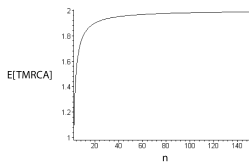
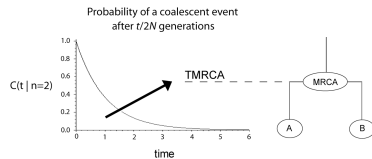


## TMRCA

The times of coalescent events,  $T(i)$ , are random draws from  $C(t)$ .

The  $T(i)$ 's are independent, so to find the average TMRCA of a sample of size  $n$ , we simply sum all of the expectations of the  $T(i)$ 's. The figure above shows this expectation for various values of  $n$ .

$$E[TMRCA] = \sum_{i=2}^n E[T(i)] = 2 \left( 1 - \frac{1}{n} \right)$$



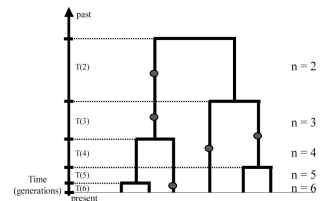
## Adding mutations

Mutations are randomly placed on branches proportional to their length.

For instance, we can compute that total length of the genealogy,  $T$ , by summing over the product of the coalescent intervals,  $T(i)$ , and the number of lineages that share that interval,  $i$ :

$$T = \sum_{i=2}^n iT(i)$$

The number of mutations on the genealogy is proportional to total length,  $T$ :



## Estimating $\theta$

Assuming that the per generation mutation rate is  $\mu$  under the infinite-sites model, the expected number of polymorphic sites,  $E[S]$ , is:

$$E[S] = N\mu T$$

If we define  $\theta = 2N\mu$  for a haploid locus, then we see that:

$$\begin{aligned} E[S] &= \frac{\theta}{2} \sum_{i=2}^n iT(i) \\ &= \frac{\theta}{2} \sum_{i=2}^n i \frac{2}{i(i-1)} \\ &= \theta \sum_{i=1}^{n-1} \frac{1}{i} \end{aligned}$$

You might recognize this equation after a bit of rearrangement:

$$\hat{\theta}_S = \frac{S}{a_1} \quad \text{where:} \quad a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \quad \text{This is Watterson's } \theta$$

## Relaxing the assumptions of the WF model

We have assumed so far that the population follows the WF model.

Much of the work on coalescent theory in the past 5–10 years explores the coalescent when some of the **assumptions** of the WF model are **relaxed**.

Some results have been obtained from the **Moran model**, which allows overlapping generations. This model results in a change in the time scaling (much like changes in  $N$ ), but otherwise, they are largely the same.

Some of the other assumptions that have been relaxed include:

- Fluctuations in population size \*
- Migration/isolation models (structured coalescent) \*
- Recombination (recombination graph)
- Selection (ancestral selection graph)
- Metapopulations (extinction/recolonization)

# Fluctuations in population size

Changes in population size that occur on the time scale of the coalescent **change the shape** of the genealogies (more rapid fluctuations simply change  $N_e$  - the harmonic mean).

Examples of this effect are shown in the figure to the right.

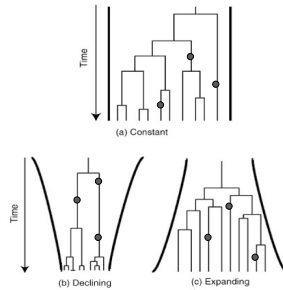
**Expanding population:**

- Produces long external branches on the genealogy.
- Long external branches result in an excess of singleton mutations.

**Declining populations:**

- Longer internal branches.
- Long internal branches result in deficiency of singleton mutations.

Statistical methods have been developed that detect this signal in the data.



# D statistics

One statistic that is sensitive to departures from the WF model is Tajima's D

Tajima's D is defined as the difference between two estimators of theta:

- Watterson's theta,  $\hat{\theta}_s$
- Average number of pairwise differences,  $\hat{\theta}_x$

$$D = \frac{\hat{\theta}_x - \hat{\theta}_s}{\sqrt{\text{Var}(\hat{\theta}_x - \hat{\theta}_s)}}$$

$\hat{\theta}_s$  is sensitive to the number of singletons (mutations on external branches)

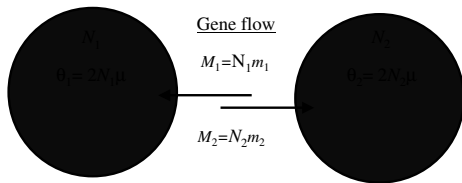
$\hat{\theta}_x$  is less sensitive to the number of singletons

- D: excess of rare mutations
- +D: deficiency of rare mutations

Tajima's D identifies genealogies with **unexpected shapes**

# Equilibrium Migration Model

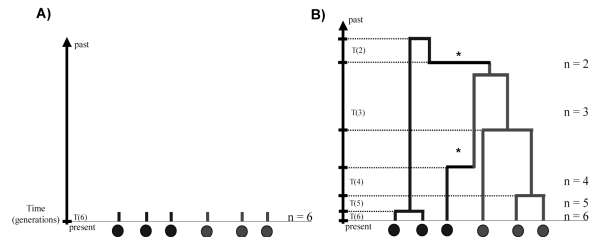
- Drift migration equilibrium
- Population sizes may be different
- Migration rates may be asymmetric
- 4-demographic parameter model



You've seen the parameter  $4N_e m$  before, I use  $N_e m$  because:  
 1. We have been talking about a **haploid** population  
 2. These rates are **decomposed** into immigration rates

# Structured Coalescent

- Start at time  $t = 0$  with labeled individuals (A in figure)
- Draw coalescent event from  $C(t | n)$  as before but the rate,  $\lambda$ , is the sum of the coalescent and migration rates
- To decide what event, pick an event proportional to it's rate:
- If it's a coalescent event, pick two individuals in the same population
- If it's a migration event, change an individual's label



# Geneological interpretation of $F_{ST}$

**Allele Frequency:**

$$F_{ST} = \frac{\sigma_p^2}{p(1-p)} \quad \sigma_p^2 = \overline{p_i^2} - \bar{p}^2$$

$$F_{ST} = 0.50$$

**Geneological:**

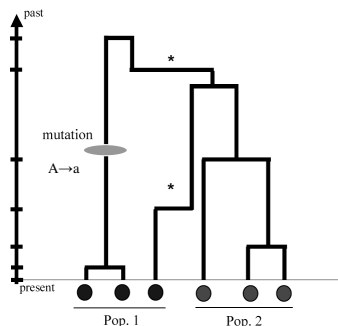
$$F_{ST} = \frac{\bar{t} - \bar{t}_w}{\bar{t}} = 1 - \frac{\bar{t}_w}{\bar{t}}$$

**Pairwise differences:**

$$K_{ST} = \frac{K_T - K_S}{K_T}$$

$K_T$  = A.P.D. among all indivs. in sample  
 $K_S$  = A.P.D. among indivs. within subpopulations

Frequency of alleles		
	A	a
Pop 1	0.33	0.66
Pop 2	1.00	0.00



# Geneological interpretation of $F_{ST}$

**Pairwise differences:**

$$K_{ST} = \frac{K_T - K_S}{K_T}$$

$K_T$  = A.P.D. among all indivs. in sample  
 $K_S$  = A.P.D. among indivs. within subpopulations

**Pairwise differences between sequences**

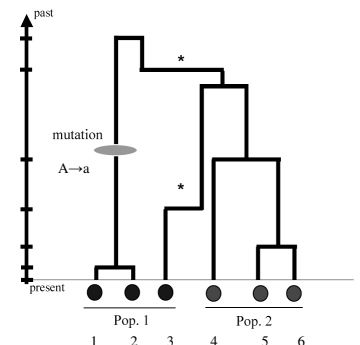
	1	2	3	4	5	6
1	-					
2	0	-				
3	1	1	-			
4	1	1	0	-		
5	1	1	0	0	-	
6	1	1	0	0	0	-

$$K_T = 0.53$$

$$K_S = 0.33$$

$$K_{ST} = 1 - \frac{K_S}{K_T} = 1 - \frac{0.33}{0.53} = 0.38$$

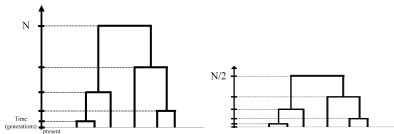
Frequency of alleles		
	A	a
Pop 1	0.33	0.66
Pop 2	1.00	0.00



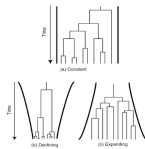
# The effects of demographic processes on genealogies

Demography	Effect
$N_e$ :	scale
Growth:	shape
Migration:	scale and shape

Scale effect (height)

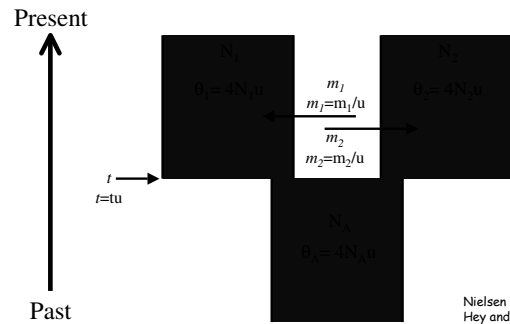


Shape effect (topology or  $T(i)$ 's)



# Isolation / Migration Model

- Ancestral population splits into two populations
- May or may not have exchanged migrants since that time
- Population sizes may be different



Nielsen and Wakeley 2001  
Hey and Nielsen 2004

# Isolation model

- Ancestral population splits into two populations
- Populations have **not** have exchanged migrants since that time
- Population sizes may be different
- 4-demographic parameter model

