# Statistical population genetics

## *Lecture 6: Mutations*

Xavier Didelot

Dept of Statistics, Univ of Oxford

didelot@stats.ox.ac.uk

# Occurrence of mutations

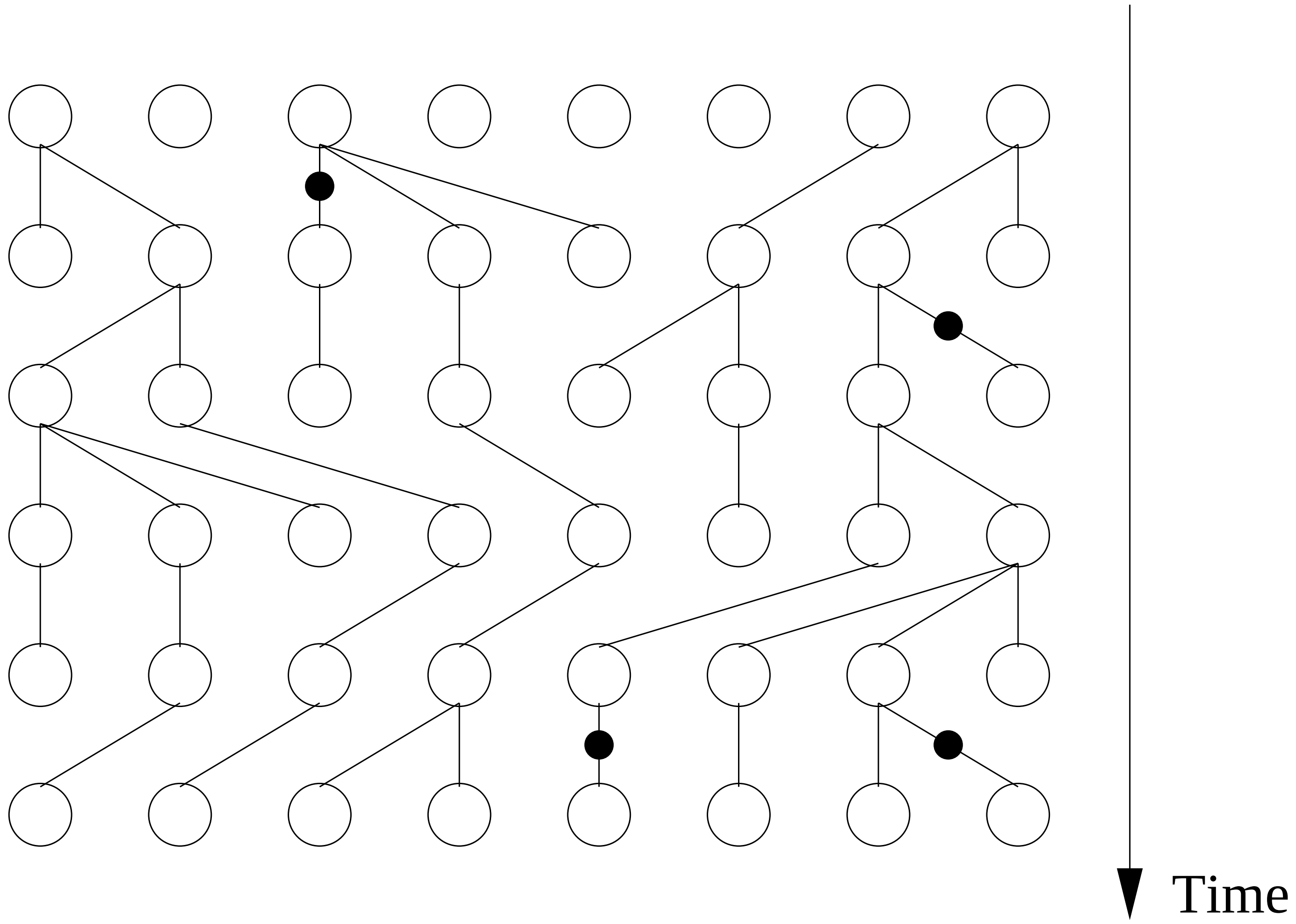
- In this lecture we discuss the **occurrence** of mutations without worrying about their **effect**.
- This is possible because we assume that mutations are **neutral**, ie. they do not change the probabilities of death and reproduction.
- Two models for the **effect** of mutations will be considered in the next two lectures: the **infinite alleles model** and the **infinite sites model**.

# Occurrence of mutations

**Definition** (Wright-Fisher model with mutation)**.**
*In the Wright-Fisher model with mutation, mutations occur with probability $u$ on offspring between generations.*

- The number of mutations occurring in the whole population at each generation is distributed as Binomial$(M, u)$.
- A similar definition could be given for the Moran model with mutation, with the same consequences in the coalescent.

# Occurrence of mutations



# Mutations in the coalescent

**Theorem** (Mutations in the coalescent model)**.**
*In the coalescent model, mutations happen as a Poisson process on the branches of the coalescent tree with rate* $\theta/2 = Mu$*.*

# Mutations in the coalescent

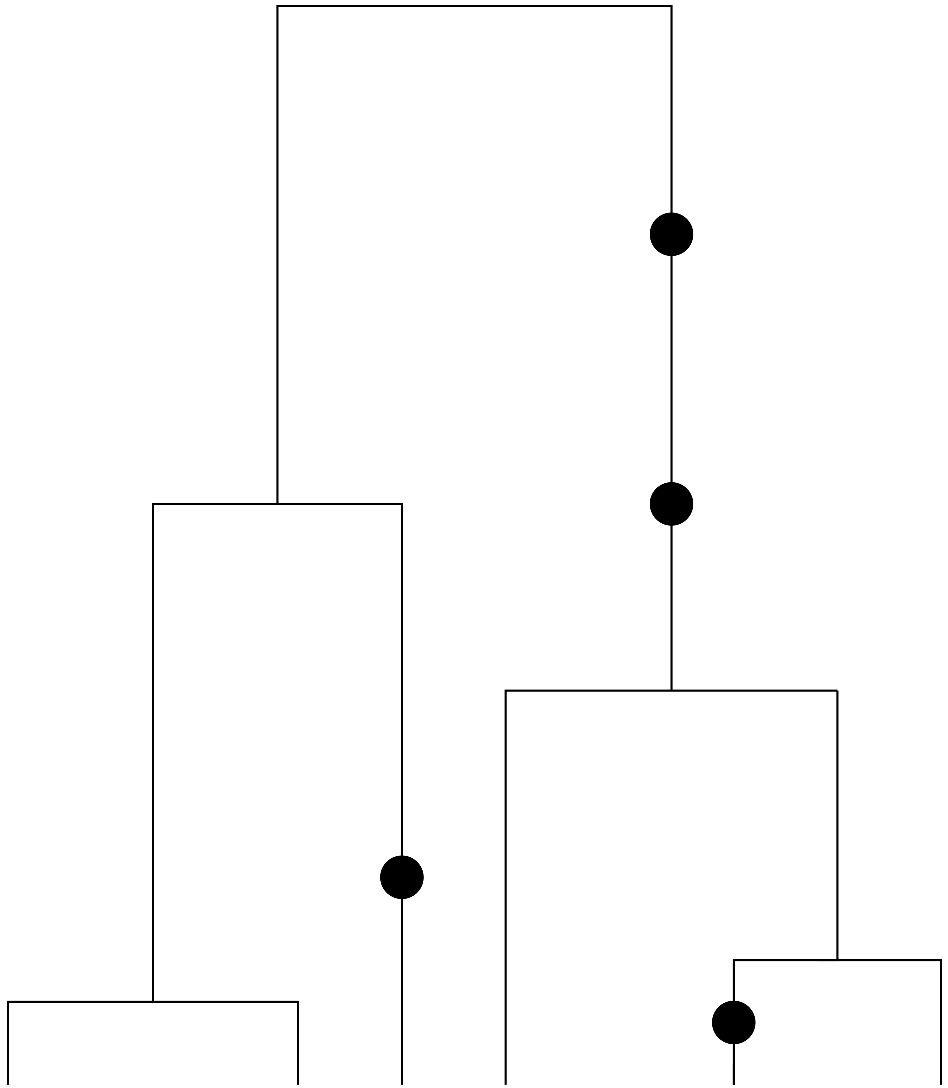
**Proof.**

- If we consider a single branch of the coalescent model, the time $T$ (in units of $M$ generations) before the first mutation satisfies:

$$\mathbb{P}(T > t) = (1-u)^{tM} = \left(1 - \frac{\theta}{2M}\right)^{tM} \xrightarrow[M\to\infty]{} \exp(-\theta t/2)$$

- Thus $T$ is exponentially distributed with parameter $\theta/2 = Mu$.
- Mutations occur independently on the branches of the coalescent since they occur independently on disjoint lineages of the Wright-Fisher model.
- Mutations therefore occur as a Poisson process on the branches of the coalescent tree. □

# Mutations in the coalescent



# Simulation algorithm

- The number of mutations occurring on a branch of length $l$ is **Poisson** distributed with mean $\theta l/2$.
- The following algorithm can be used to **simulate** the coalescent model with mutation:

  **Algorithm** (Coalescent with mutations)**.**

  1. *Simulate a coalescent tree using the algorithm without mutations;*
  2. *For each branch of length $l$, draw the number of mutations from Poisson($\theta l/2$);*
  3. *For each branch the times of the mutations are chosen uniformly on the branch.*

# Coalescence and mutation

**Theorem** (Combining coalescence and mutation)**.**
*In the coalescent with mutation, events (either mutation or coalescence) occur at rate* $k(k-1+\theta)/2$ *where* $k$ *is the number of lineages. When an event happen, it is a mutation with probability* $\theta/(\theta+k-1)$ *and a coalescence with probability* $(k-1)/(\theta+k-1)$.

- Combining mutation and coalescence is extremely useful to establish **recursion equations** in the coalescent.
- We will see many examples of this!

# Coalescence and mutation

**Proof.**

- If $X$ and $Y$ are exponentially distributed with parameters $\lambda_1$ and $\lambda_2$, min$(X, Y)$ is exponentially distributed with parameter $\lambda_1 + \lambda_2$:

$$\mathbb{P}(\min(X,Y) < t) = \mathbb{P}(X < t) + \mathbb{P}(X > t)\mathbb{P}(Y < t) = 1 - \exp(-(\lambda_1 + \lambda_2)t)$$

- Thus the waiting time before the first event (either coalescence or mutation) is Exponential($k(k-1)/2 + \theta k/2$)).
- Furthermore the probability that each event is either a mutation or a coalescence follows from:

$$\begin{aligned}\mathbb{P}(X < Y) &= \int_0^\infty f_X(x)(1 - F_Y(x))\mathrm{d}x \\ &= \int_0^\infty \lambda_1 \exp(-\lambda_1 x)\exp(-\lambda_2 x)\mathrm{d}x = \frac{\lambda_1}{\lambda_1 + \lambda_2}\end{aligned}$$

□

# Simulation algorithm

The following algorithm can be used to **simulate** the coalescent model with mutation:

**Algorithm** (Coalescent with mutations version 2)**.**

1. *Start with* $k = n$ *lines where* $n$ *is the sample size;*
2. *Wait an exponentially distributed amount of time with parameter* $k(k-1+\theta)/2$*;*
3. *With probability* $(k-1)/(k-1+\theta)$ *the event is a coalescence event, otherwise it is a mutation event;*
4. *If the event is a coalescent event, choose a pair of lines randomly and join them. Decrease the value of* $k$*;*
5. *If the event is a mutation, choose uniformly a line to mutate;*
6. *If* $k > 1$*, go back to step 2.*

# Mutations on a coalescent tree

The following theorem was first obtained by Watterson (1975) and later by Tavaré (1984) using coalescent theory.

**Theorem** (Mutations on a coalescent tree)**.**
*Let $S_n$ denote the number of mutations on a coalescent tree of $n$ genes. Then:*

$$\mathbb{P}(S_n = s) = \frac{n-1}{\theta} \sum_{i=1}^{n-1} (-1)^{i-1} \begin{pmatrix} n-2 \\ i-1 \end{pmatrix} \left( \frac{\theta}{i+\theta} \right)^{s+1}$$

# Mutations on a coalescent tree

**Proof.**

- On each branch of length $l$, the number of mutations is Poisson distributed with rate $\theta l/2$.
- Furthermore, the convolution of Poisson distributions with rates $\lambda_1, ..., \lambda_m$ is a Poisson distribution with rate $\sum_{i=1}^{m} \lambda_i$.
- Therefore, $S_n$ is Poisson distributed with parameter $\theta T_{\text{total}}/2$. Integrating over the distribution of $T_{\text{total}}$ gives:

$$\mathbb{P}(S_n = s) = \int_{t=0}^{\infty} \frac{(\theta t/2)^s}{s!} e^{-\theta t/2} \mathbb{P}(T_{\text{total}} = t)\mathrm{d}t$$

- Injecting the formula for the distribution of $T_{\text{total}}$ gives the required result.

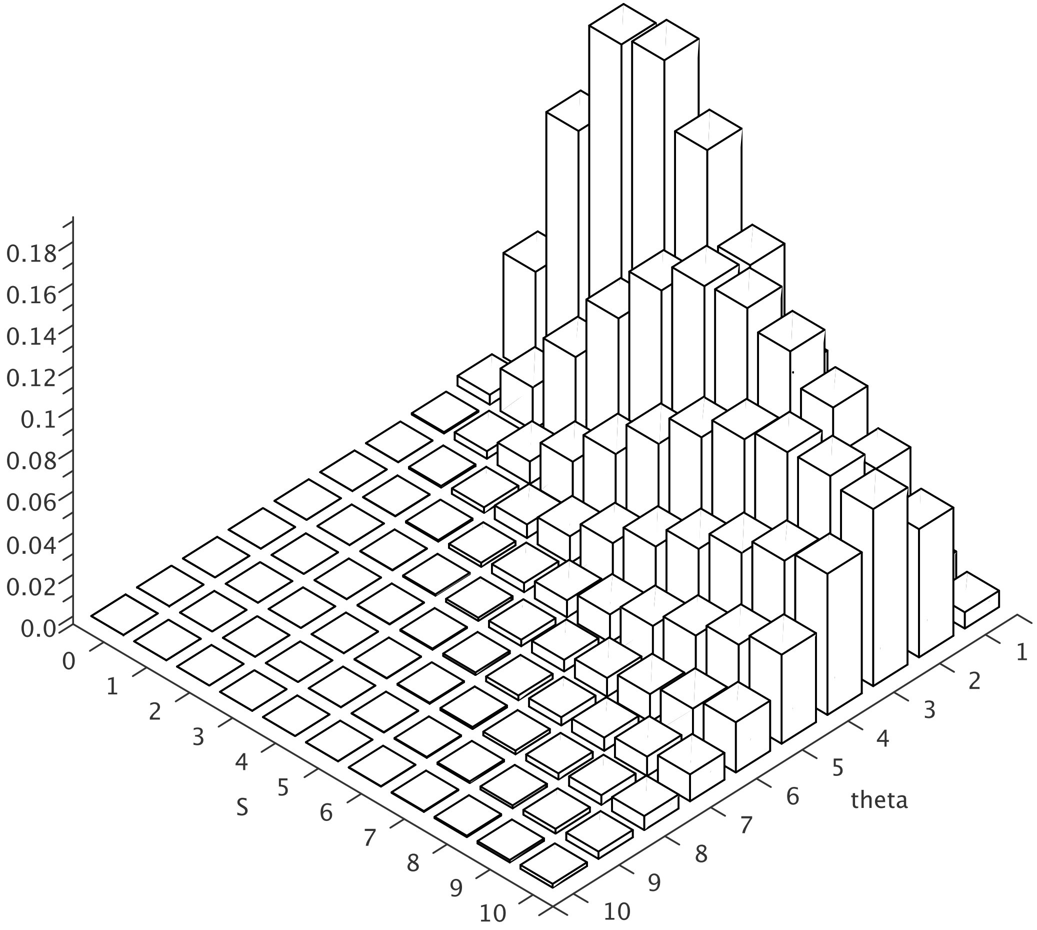
# Mutations on a coalescent tree

- Another approach is to use the recursive form of the coalescent with mutations.
- The $s$ mutations can occur in two ways: with the last event being either a coalescence or a mutation.
- If the last event was a mutation, then just before that we had $n$ lineages and $s-1$ mutations in the tree.
- If the last event was a coalescence, then just before that we had $n-1$ lineages and $s$ mutations in the tree.
- We deduce from this the following recursion Equation:

$$\mathbb{P}(S_n = s) = \frac{n-1}{n-1+\theta}\mathbb{P}(S_{n-1} = s) + \frac{\theta}{n-1+\theta}\mathbb{P}(S_n = s-1)$$

- This can be solved with limiting condition $\mathbb{P}(S_1 = 0) = 1$ to give the desired result. □

# Mutations on a coalescent tree



# Mean and variance

**Theorem** (Mean and variance of the number of mutations)**.**

*Let $S_n$ denote the number of mutations on a coalescent tree of $n$ genes. Then:*

$$\begin{aligned}\mathbb{E}(S_n) &= \theta \sum_{i=1}^{n-1} \frac{1}{i} \\ \mathrm{var}(S_n) &= \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}\end{aligned}$$

# Mean and variance

**Proof.**

- The mean and variance of $S_n$ can be calculated from the probability density function above.
- It is also possible to use the fact that $S_n$ is Poisson distributed with parameter $\theta T_{\mathrm{total}}/2$.
- We can also use the fact that $S_n = \sum_{i=2}^{n} s_i$ where $s_i$ is the number of mutations occurring when there are $i$ lineages. We have:

$$\mathbb{P}(s_i = 0) = \frac{i-1}{\theta + i - 1} \text{ and } \mathbb{P}(s_i = s > 0) = \frac{\theta}{\theta + i - 1}\mathbb{P}(s_i = s - 1)$$

# Mean and variance

- By induction this leads to:

$$\mathbb{P}(s_i = s) = \left(\frac{\theta}{\theta + i - 1}\right)^s \frac{i-1}{\theta + i - 1}$$

- This is a shifted geometric distribution with parameter $p = (i-1)/(\theta + i - 1)$, so that the mean is $(1-p)/p$ and the variance $(1-p)/p^2$.
- The mean of $s_i$ is therefore equal to $\theta/(i-1)$ and the variance to $\theta/(i-1) + \theta^2/(i-1)^2$.
- Summing from $i = 2$ to $n$ gives the result. □

# Example

- Dorit *et al.* (1995) **sequenced** a sample of 38 ZFY genes from the human population.
- They observed **no mutation** between the sequences.
- Donnelly *et al.* (1996) used this data in a **Bayesian coalescent framework** to estimate $T$, the TMRCA of the human population.

# Example

- Let $T_i$ denote the time during which $i$ ancestral lines are present and $S_i$ the number of mutations occurring during that time.
- We have $T = \sum_{i=2}^{38} T_i$ and $\forall i \in [2..38], S_i = 0$.
- We want to compute $\mathbb{E}(T|S = 0)$.
- The prior distribution of $T_i$ is exponential with parameter $i(i-1)/2$:

$$\mathbb{P}(T_i = t) = \frac{i(i-1)}{2} \exp\left(\frac{-ti(i-1)}{2}\right)$$

- Furthermore:

$$\mathbb{P}(S_i = 0|T_i = t) = \exp\left(\frac{-t\theta i}{2}\right)$$

# Example

- Using Bayes' rule, we get:

$$\mathbb{P}(T_i = t | S_i = 0) = \frac{\mathbb{P}(S_i = 0 | T_i = t)\mathbb{P}(T_i = t)}{\mathbb{P}(S_i = 0)} \propto \exp\left(\frac{-ti(\theta + i - 1)}{2}\right)$$

- Thus the conditional distribution of $T_i | S_i = 0$ is exponential with mean $2/(i(\theta + i - 1))$.
-

$$\mathbb{E}(T | S = 0) = \sum_{i=2}^{n} \frac{2}{i(\theta + i - 1)}$$

- Taking $u = 2 \cdot 10^{-5}$ and $M = 5000$, we get $\theta = 2Mu = 0.2$.
- This implies $\mathbb{E}(T|S = 0) = 1.72$.
- If we assume that each generation lasts on average 20 years, we get an estimate of 172,000 years for the TMRCA of the sample.

# Summary

- **Mutations** occur as a Poisson process with rate $\theta/2$ on the branches of the coalescent tree
- **Combining** mutation and coalescence is a powerful tool to derive **recursion equations**
- We have found a recursion to calculate the **number of mutations on a coalescent tree**
- The Dorit dataset is a first example of **inference** from genetic data