Evolutionary Genetics: Part 2
Wright-Fisher model and neutral theory

\[ f_i = \frac{\theta}{i} \binom{n-1}{i}^{-1} \sum_{k=2}^{n-i+1} \binom{k}{i} \binom{n-k}{i-1} E(T_k), \quad 1 \leq i \leq n-1. \]

\[ E(T_k) = \sum_{j=k}^{n} (-1)^{j+k} \frac{(2j-1)n!(n-1)!(j+k-2)!}{k!(k-1)!(n-j)!(n+j-1)!(j-k)!} \int_0^\infty \exp\left(-\frac{\beta^2}{2}\int_0^t 1/p(s)ds\right) dt. \]

Winter Semester 2012-2013

Prof Aurélien Tellier
FG Populationsgenetik
Color code

Color code:

Red = Important result or definition

Purple: exercise to do

Green: some bits of maths
Population genetics: 4 evolutionary forces

random genomic processes
(mutation, duplication, recombination, gene conversion)

natural selection

molecular diversity

random spatial process (migration)

random demographic process (drift)

Lets look at genetic drift
Wright-Fisher model, genetic drift and neutral theory
Neutral theory

- Introduced by Motoo Kimura in 1960s, big controversy at the time.
- Can we explain all polymorphism data without the action of natural selection?
- Kimura: Most polymorphisms that occur do not influence the fitness of an individual,
- thus these polymorphisms are not subjected to selection
- these mutations would evolve neutrally

- mutations at silent or degenerate sites do not change the Amino Acid BUT may or may not evolve neutrally
- some non-synonymous mutations do not affect fitness (change in Amino Acid does not affect the fitness) BUT may or may not evolve neutrally
Neutral theory

- Neutral theory = most changes in allele frequencies in a population can be attributed to **genetic drift**

- Why?

- When a mutation arise in a gamete of an individual, many things can happen:
  - the carrier of the gamete must survive to reach the reproductive age,
  - that gamete must be fertilized and develop an embryo,
  - the embryo has to be viable to be at the next generation.

- Genetic drift means that mutation creates new alleles which by chance
  - can rise in frequency and spread in a population,
  - or they can get lost.
Neutral theory

- We need a model to explain how genetic drift occurs
- and then use it to derive expectations on what polymorphism we should observe in DNA sequences

- This model is based on how a population of individuals reproduce over time

- Important: to demonstrate that a given trait or polymorphism pattern is due to selection, you MUST disprove alternative neutral explanations!
The Wright-Fisher model
The Wright – Fisher model

- Fundamental model in population genetics

**Assumptions (check list):**

- Constant population size
- Discrete and non-overlapping generations
- Random mating (= panmixia)
- Equal sex-ratio
- $2N$ haploid individuals $= N$ diploid individuals (with two allele each)
- One locus
- No recombination
The Wright – Fisher model

- How does it work?
- Let us assume 10 haploid individuals at generation $t$
- The offspring generation is obtained from the parents as follows:
  - constant population => 10 individuals at generation $t+1$
  - each individual from the offspring picks a parent at random from generation $t$
  - connect parent and child by a line
  - each offspring inherits the genetic information of the parents
The Wright – Fisher model

- all individuals have the same fitness (= expected number of offsprings)
- this means that there is an equal probability for all individuals to be picked as a parent

- each individual has \(2N\) chances to become ancestor to generation \(t+1\)
- and the probability to be picked at each draw is \(1/2N\)

- Game 1: you are the offspring choose one parent at random from previous generation \((2N=6)\) => throw dice
- Game 2: you are the parents, for each offspring you have \(1/2N\) chance to have it as yours => throw dice
Figure 2.1: The 0th generation in a Wright-Fisher Model.

Figure 2.2: The first generation in a Wright-Fisher Model.
Figure 2.3: The tangled and untangled version of the Wright-Fisher Model after some generations.
The Wright – Fisher model

- all individuals have the same fitness (= expected number of offsprings)
- this means that there is an equal probability for all individuals to be picked as a parent

- each individual has $2N$ chances to become ancestor to generation $t+1$
- and the probability to be picked at each draw is $1/2N$
The Wright – Fisher model

Maths 1: Binomial distribution

Maths 2: Poisson distribution
for the Wright – Fisher model:

- $n = 2N$ and $p = 1/2N$

Can you calculate the expectation and variance for the binomial distribution for this model?

- thus $\lambda = np = 1$

Can you calculate the expectation and variance for the Poisson distribution for this model?

What is the difference between $\text{Exp}$ and $\text{Var}$ of the Binomial and Poisson?

these are approximations
From the Poisson distribution, the probability of an individual not to leave descendants is:

\[ P[X=0] = e^{-1} \approx 0.37 \]

A fraction 1–0.37 = 63% of all individuals have descendants at each generation.

in a randomly mating population, the present day population descent from a small fraction of individuals a few generations ago.

this fraction is approximately \( \approx 0.63^t \)

For a population of size \( 2N = 10,000 \)

The population comes from: \( 10,000 \times 0.63^{15} \approx 10 \) individuals from 15 generations ago.

The 9,990 other individuals did not leave descendants today.
Genetic drift
Genetic drift

- It is a random stochastic process!!!!

- What it means for evolutionary biology: even if we know everything about a population and its biology, we cannot predict the state of the population in the future.

- Game 3: Two alleles (Red and White)
- you are the offspring in populations of sizes \(2N = 6\), \(2N = 12\) and \(2N = 20\)
- \(p = 0.5\) use the sheet with table to keep track of frequencies
- at each generation, the numbers on the dice for successful reproduction change according to frequency of red allele at previous generation
- \(\Rightarrow\) throw dice
Genetic drift

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 2N = 12

- Throw dice for each offspring
- Chance of being red = \#red / 12
- Here from 1 to 4 on the dice = reproduction
In fact we do not need to pick a random parent for each individual one by one:

- we can pick up this number from the binomial distribution, and use this as the frequency at next generation

\[ p = \text{frequency of allele } A \quad 1-p = \text{frequency of allele } a \]

- frequencies at last generation in population with \(2N\) individuals

- if more than two alleles we use the multinomial distribution
Genetic drift

- calculate from formula of the binomial distribution the probabilities:
  - of losing an allele $P[X=0]$ with $2N = 10$ for $p = 0.5$; $p = 0.2$; $p = 0.01$
  - of losing an allele $P[X=0]$ with $p = 0.5$ and $2N = 10$; $2N = 100$; $2N = 1000$

- calculate from formula of the binomial distribution the probabilities:
  - of fixing an allele $P[X=2N]$ with $2N = 10$ for $p = 0.5$; $p = 0.2$; $p = 0.01$
  - of fixing an allele $P[X=2N]$ with $p = 0.5$ and $2N = 10$; $2N = 100$; $2N = 1000$

- Can you correlate this with what you observe during the dice experiments???
Genetic drift

- Genetic drift = random change of allele frequency between generations

- Key facts about genetic drift:
  - The probability of loosing/fixing alleles is HIGHER for small $N$ (population size)
  - The probability of loosing/fixing alleles is HIGHER for small $p$ (frequency)
Genetic drift

- Use Populus

- What about the time to fix or loose alleles?

- Look at the graphs, and try different values of
  
  - $N = 100, 500, 1000, 10\,000$
  
  - *initial frequency* = 0.01, 0.1, 0.5, 0.99

- You can look at one or several loci: what does this mean?
 Genetic drift

- Genetic drift = random change of allele frequency between generations

- Key facts about genetic drift:
  - The probability of losing/fixing alleles is HIGHER for small \( N \) (population size)
  - The probability of losing/fixing alleles is HIGHER for small \( p \) (frequency)

- If loci are independent in the genome (physically not linked), each locus has independent changes in allele frequency under genetic drift!!!
The coalescent - 1
Kingman JF (1982)
Until now we have predicted the state of population at $t+1$ based on time $t$

This is the process forward in time

Useful because it is logical and intuitive

However, we can also follow the genealogy backward in time from present to past

Why???

because most data we collect come from present day populations that we can sample

The question becomes: what are the forces that have shaped the observed patterns of diversity in our data (SNPs,...)?

These forces have acted in the history of the population, so we look at the genealogy
Figure 2.3: The tangled and untangled version of the Wright-Fisher Model after some generations.
Back to the Wright – Fisher model

Figure 2.6: The coalescent of two lines in the Wright-Fisher Model
Back to the Wright – Fisher model

The genealogy of a sample of the population
Back to the so-called:

MRCA = Most Recent Common Ancestor of the sample

Because only a limited proportion of individuals leave descendants

Figure 2.6: The coalescent of two lines in the Wright-Fisher Model
The coalescent

- When two lines fuse = coalescent event
- The idea is that we care only now about individuals which left descendants that are found in our sample

- go to:

- **Animation** shows the Wright-Fisher model with only the important lineages
- **Trees** = rearrangements of the lines so that they do not cross
- **Rho** = 0

- Vary $n$ = number of samples (individuals at present)
- How does the tree look like? Where are most of the coalescent events (recent or old?)
The coalescent

- **General question:** What is the probability that two or more individuals share a common ancestor?

- What is the probability that two individuals at generation $t$ have the same ancestor at generation $t-1$?

- In the Wright-Fisher model:
  - $P[\text{common parent one generation ago}] = 1/2N$
  - $P[\text{NO common parent one generation ago}] = 1 - 1/2N$
  - As one offspring chooses its parent, and this is the proba that a second offspring chooses the same parent.
The coalescent

- **General question:** What is the probability that two or more individuals share a common ancestor?

- What is the probability that two individuals at generation $t$ have the same ancestor at generation $t-1$?

- In the Wright-Fisher model:
  - $P[\text{common parent one generation ago}] = \frac{1}{2N}$
  - $P[\text{NO common parent one generation ago}] = 1 - \frac{1}{2N}$
  - As one offspring chooses its parent, and this is the proba that a second offspring chooses the same parent...

- Can be iterated into the past for two lineages
  - $P[\text{two lineages coalesce } t \text{ generations ago}] = \left(\frac{1}{2N}\right) \times \left(1-\frac{1}{2N}\right) \times \left(1-\frac{1}{2N}\right) \times \ldots \times \left(1-\frac{1}{2N}\right)$

- $P[\text{two lineages coalesce } t \text{ generations ago}] = \left(\frac{1}{2N}\right) \times \left(1-\frac{1}{2N}\right)^{t-1}$
The coalescent: two lineages

- \[ P[\text{two lineages coalesce} \ t \ \text{generations ago}] = \left(\frac{1}{2N}\right) \times (1-\frac{1}{2N})^{t-1} \]
- If you remember the geometric function \( \approx \) exponential

- \[ P[\text{two lineages coalesce exactly} \ t \ \text{generations ago}] = P[X = t] = p \times (1-p)^{t-1} \]
- \[ P[X \geq t] = (1-p)^t \approx e^{-pt} \] with \( p = \frac{1}{2N} \)

Waiting time for coalescent two lineages

Coalescence is faster in small populations!!!
For a sample of size \( n \), there are possible coalescent pairs

\[
\binom{n}{2} = \frac{n!}{(n-2)!2!} = \frac{n(n-1)}{2}
\]

The coalescent coalescent probability per generation is :

\[
P[\text{coalescence in sample of size } n] = \frac{\binom{n}{2}}{2N}
\]

So if \( T_n \) is the time until the first coalescent event:

\[
P[T_n > t] = \left[ 1 - \frac{\binom{n}{2}}{2N} \right]^t \approx \exp \left( -\frac{t \binom{n}{2}}{2N} \right)
\]

What does this mean?
The coalescent

- We can calculate many aspects of a genealogical (coalescent) tree for a population of size $2N$

- Time to MRCA: $E[T_{MRCA}] = 4N (1 - 1/n)$

- Time of coalescence of last two lineages: $E[T_2] = 2N$

Most of the coalescence time happens for low number of lineages far in the past

Remember your exercise before

Times are functions of $N$
The coalescent

- We can calculate many aspects of a genealogical (coalescent) tree for a population of size $2N$.
- The probability that a sample of size $n$ contains the MRCA of the whole population is $rac{n-1}{n+1}$.
- So typically a sample of 20 individuals is enough to study a population!

![Graph showing the probability of a sample containing the MRCA of the whole population as a function of sample size.](image-url)
We can calculate many aspects of a genealogical (coalescent) tree for a population of size $2N$.

- Beware we calculated only the Expectations.
- However, the variance also exist, so that for the same $n$ and $2N$ values, each tree is be different in size and shape.

- The size of the tree is given in units of $2N$ generations !!!!
Important assumption of the coalescent model:

We assume that $n << 2N$

So that the probability of more than 2 lineages to coalesce is very small

$P[3 \text{ lineages to coalesce}] = (1/2N)^2$

This is small enough to be be neglected if $N$ is big, but not always true
How to use the coalescent to simulate sequence data?

We need to introduce mutations on the tree