Introduction to Coalescent Theory

Instructor: annasapfo.malaspinas@unil.ch

ToC

Table of Contents

Objectives

Brief introduction and definition

The WF model

The coalescent

Coalescence times

Mutations

Site Frequency Spectrum SFS

Demographic inference: how to

Other summary statistics

References

- Textbooks
- Simulation software

Acknowledgements

Lecture's objectives

Goal is to understand/learn about:

- what is the coalescent,
- how coalescent times are distributed, i.e., predict the shape of the tree
- measures of DNA polymorphism (summary statistics)
  - (S_n: # SNPs)
  - (π: diversity)
  - SFS: Site Frequency Spectrum
  - (D-statistics)
- estimating demography from summary statistics

The coalescent - setting the scene

“The discovery of Kingman’s coalescent is one of the most important theoretical discoveries in all of biology over the past 50 years.”

Nielsen, R. Genetics 204, 389–390 (2016).

The coalescent - in dates

The coalescent has played a central role in population genetics for over 30 years. It is the culmination of decades of work.

Key players:

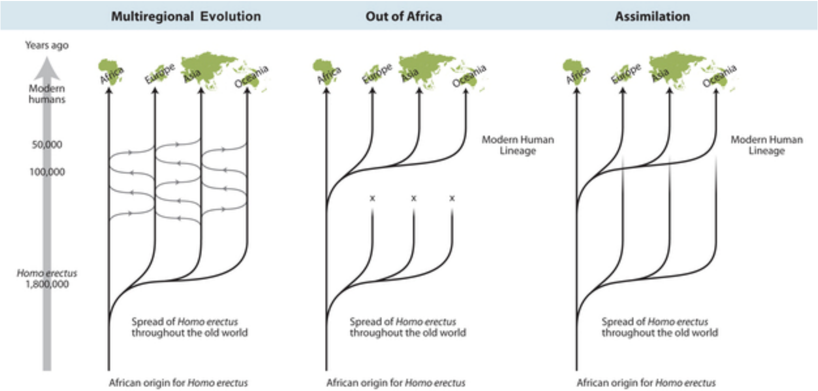
- Kingman (1982) - definitive mathematical treatment
- Hudson (1983) - recombination
- Tajima (1983) -“power of the coalescent“

In comparison, the first survey of DNA sequence polymorphism was published by Kreitman only in 1983.

Importance

- Coalescent models follow the genealogy (ancestry) of “genes” backward in time, starting from the present.
- This turns out to be a very powerful way of thinking about genetic data (polymorphism, SNPs, variants):
  - elegant mathematics,
  - powerful simulation algorithms,
  - explicit likelihood calculations.
- Arguably, an intuitive understanding of coalescent models is essential for anyone analyzing polymorphism data.

Application



We would like to pick the most probable model.
One can compute the likelihood of the data for each model using the coalescent theory.
Tryon, C. & Bailey, S. (2013) Nature Education Knowledge 4(3):4

Intuition

The coalescent is based on the following insights:

- state can be separated from descent,
- the properties of the sample depends only on their genealogy, which can be modeled backward in times.

Definitions

Coalesce (verb) to grow together, merge, combine, fuse.

Coalescence (noun) the joining or merging of elements to form one mass or whole.

Used here to describe the event of ancestral lineages coalescing, merging.

The **neutral** Wright-Fisher model

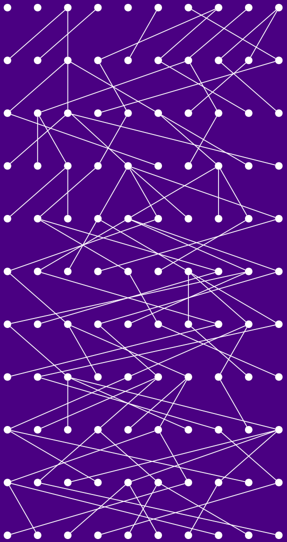
A classical forward in time approach in population genetics to model the evolution of a population (one locus)/the change in allele frequencies through time.

Some of the assumptions:

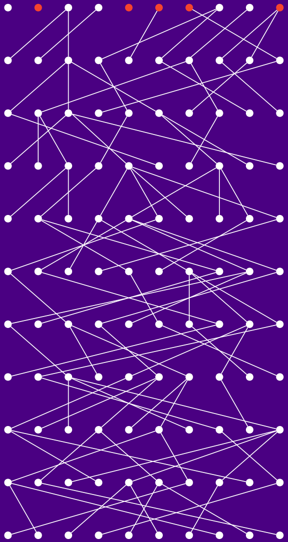
- Two alleles *A1* and *A2* with no differences in fitness.
- Constant population of size N.
- Discrete generations $t = 0, 1, 2, ...$

Note that the constant population size assumption can be relaxed.

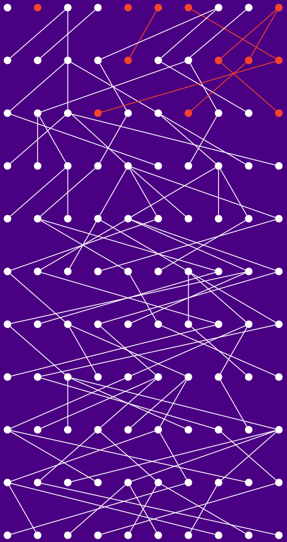
The Wright-Fisher model



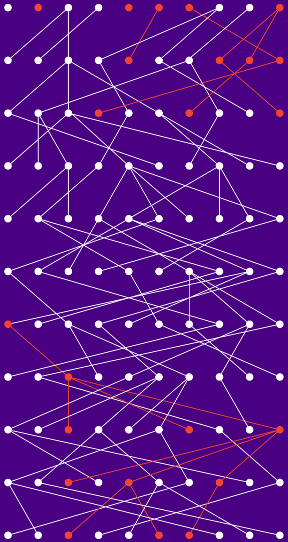
The Wright-Fisher model



The Wright-Fisher model

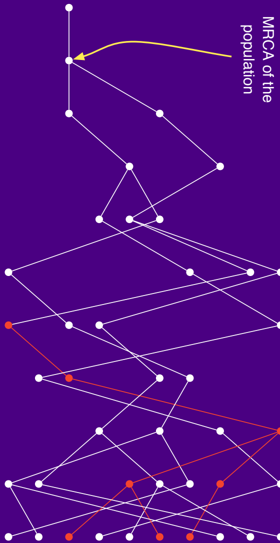


The Wright-Fisher model

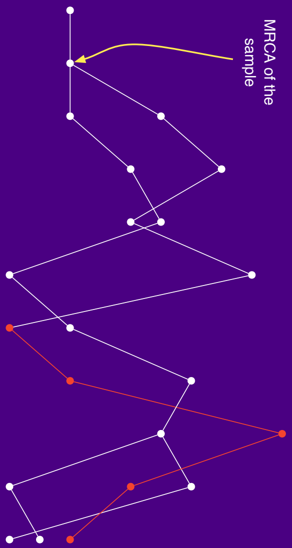


The Wright-Fisher model

MRCA: Most Recent Common Ancestor



The Wright-Fisher model



The **neutral** Wright-Fisher model

Assume:

- Two alleles *A1* and *A2* with no differences in fitness.
- Constant population of size N.
- Discrete generations $t = 0, 1, 2, \ldots.$

In this model, ancestors of the present generation obtained by **randomly sampling with replacement** from the previous generations.

$$\rightarrow$$

Which distribution does that remind you off?

The **neutral** Wright-Fisher model

Assume:

- Two alleles *A1* and *A2* with no differences in fitness.
- Constant population of size N.
- Discrete generations $t = 0, 1, 2, \ldots$.

In this model, ancestors of the present generation obtained by **randomly sampling with replacement** from the previous generations.

$$\rightarrow$$

The number of offspring contributed by a particular individual is Bin(N,1/N).

The **neutral** Wright-Fisher model

The number of offspring contributed by a particular individual is Bin(N,1/N).
Assuming i *A1* alleles and $N - i$ *A2* alleles at generation 0. Then, the probability of having j *A1* alleles in the next generation is given by:

$$P_{ij} = \binom{N}{j} p^j (1-p)^{N-j}$$

If K_t: count of *A1* alleles at generation t, then:

$$E[K_1] = Np = i \text{ and } Var[K_1] = Np(1-p)$$

Summary

Under neutrality, the joint effects of
random reproduction ("genetic drift")
and
mutation
on the distribution of a sample may be modeled by:

- generating the genealogy backward in time,
- superimposing mutation forward in time.

Implications for how we (should) view the genetic data

We want to infer an evolutionary process that gave rise to the data. Often this process affects the genealogy only.

- The observed polymorphisms are of interest because they contain information about the unobserved underlying genealogy,
- the genealogy is of interest because it contains information about the evolutionary process.

The coalescent and classical population genetics

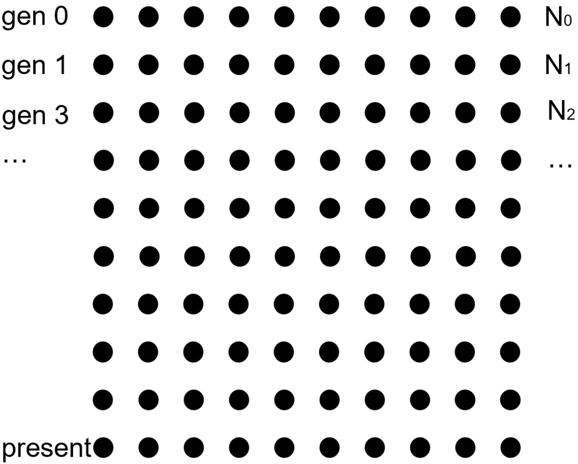
The main difference is one of perspective.

- Classical models: **prospective**
  Given starting conditions, what will happen? Useful for thinking about evolution, (experimental evolution)
- Coalescent models: **retrospective**
  Given the present, what could have happened.
  More natural way of thinking about the data.

The coalescent

Genealogies

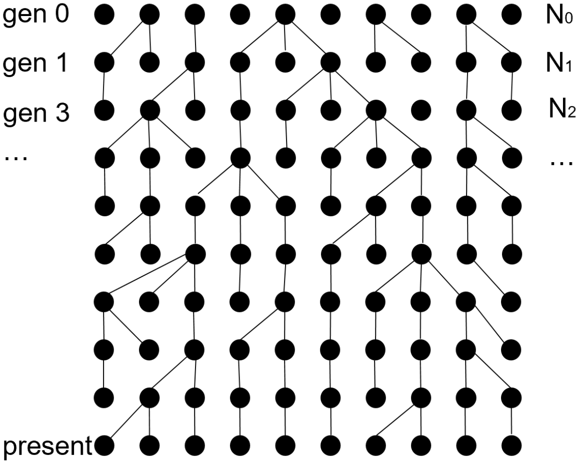
One more time, different visual



“Sampling” of parents from one generation to the next.

Genealogies

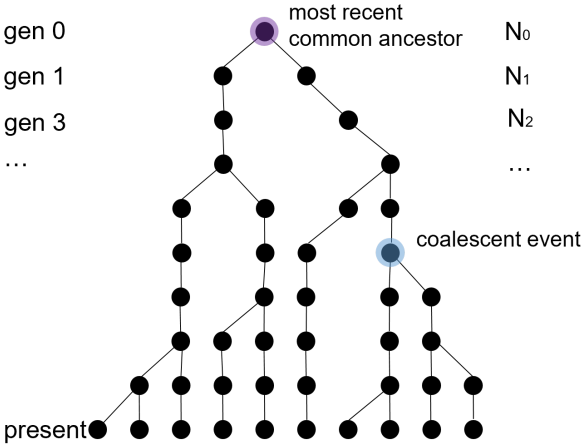
One more time, different visual



"Sampling" of parents from one generation to the next.

Genealogies: a "coalescent" tree

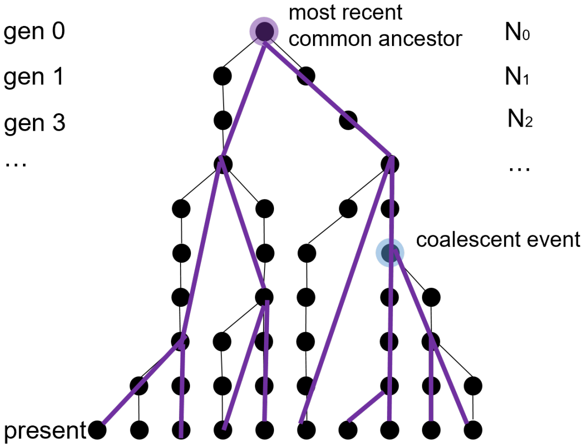
One more time, different visual



If we ignore individuals who did not contribute to current diversity, we easily see coalescent events and a most recent common ancestor.

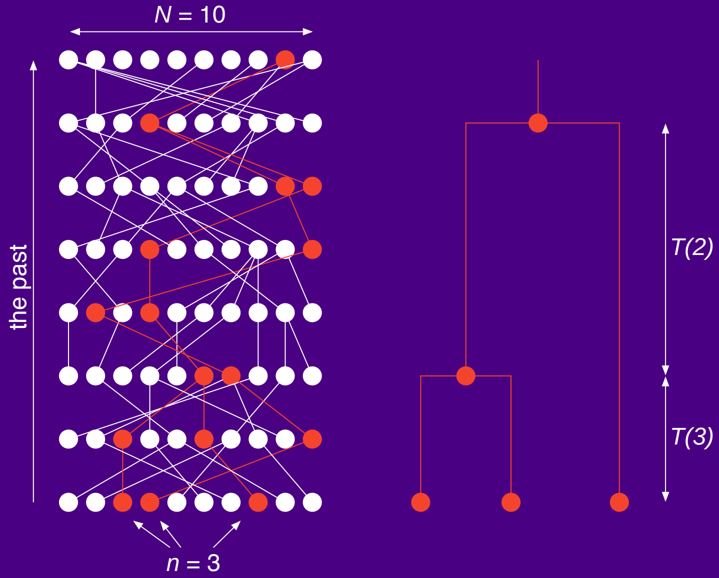
Genealogies: a "coalescent" tree

One more time, different visual



Draw lines to connect these relationships → we have a coalescent tree.

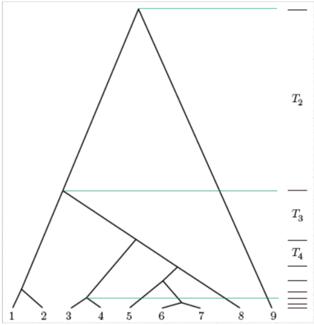
The coalescent



Coalescent trees and coalescent times

time

2 lineages left



Time to most recent common ancestor

3 lineages left

4 lineages left

Under a neutral model (no selection) and if the population is at equilibrium we know a lot about the shape of this tree.

And so what?
We know a lot about the expected patterns of variation.

The standard neutral model

Under highly simplifying assumptions (including neutrality, random mating, and constant population size), we can describe the full expectations of a tree.

This includes:

- Shape of the tree:
    - relative branch lengths,
    - **t**ime to **m**ost **r**ecent **c**ommon **a**ncestor (TMRCA),
- Patterns of variation:
    - number of segregating sites,
    - frequencies of segregating mutations.

WF and the coalescent

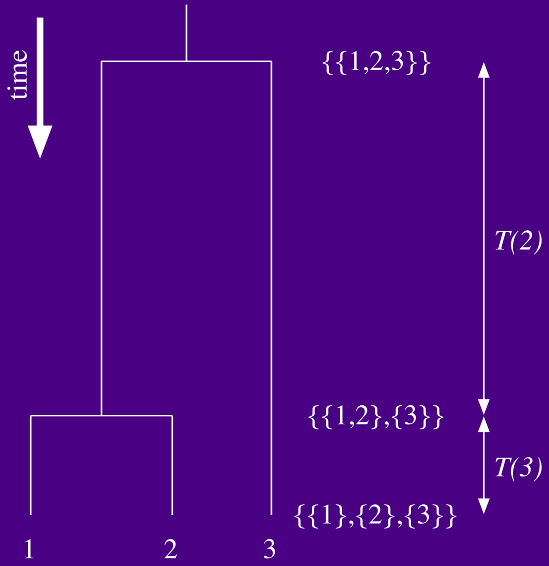
The WF model describes how a population evolves through time, it is an idealized model, that is nevertheless used very frequently.

It is hard to track mathematically.

The coalescent approximates the WF model.

The coalescent is mathematically convenient.

Topology & branch lengths



Labels on the topology

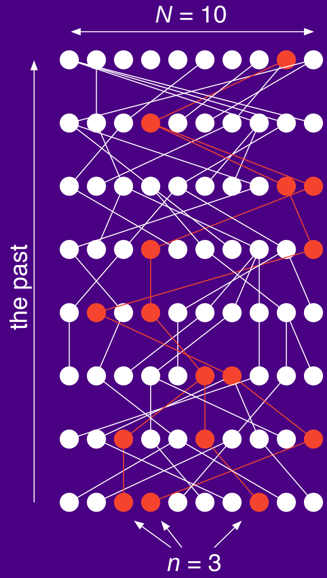
Because of neutrality, individuals are equally likely to **reproduce**. Therefore, all lineages must be equally likely to **coalesce**.

For example,

$$\{\{1\},\{2\},\{3\}\} \quad \text{goes to} \quad \begin{cases} \{\{1,2\},\{3\}\} \\ \{\{1,3\},\{2\}\} \\ \{\{2,3\},\{1\}\} \end{cases}$$

with equal probability of $1/3$
We will keep track of the number of lineages and forget about the individual labels (e.g. measure the coalescence times).

Shape of the tree: WF and the coalescent, blackboard



Derivation of the coalescent time distributions

Let us concentrate on i lineages and trace them back in time. Denoting:

$G_{i,j}^{WF}$:= probability that i lineages have j parents (in one generation).

$$G_{i,i}^{WF} =$$

No coalescence in one generation among the i lineages.
Notation: $\mathcal{O}\left(\frac{1}{N^2}\right)$ represents terms that decrease to zero like $\frac{1}{N^2}$ or faster, as N tends to infinity.

$$G_{i,i-1}^{WF} =$$

One coalescence in one generation among the i lineages

Derivation of the coalescent time distributions

Let us concentrate on i lineages and trace them back in time. Denoting:

$G^{WF}_{i,j}$:= probability that i lineages have j parents (in one generation).

$$G^{WF}_{i,i} =$$

$$1 - \frac{\binom{i}{2}}{N} + \mathcal{O}\big(\frac{1}{N^2}\big)$$

No coalescence in one generation among the i lineages.
Notation: $\mathcal{O}\big(\frac{1}{N^2}\big)$ represents terms that decrease to zero like $\frac{1}{N^2}$ or faster, as N tends to infinity.

$$G^{WF}_{i,i-1} = \frac{\binom{i}{2}}{N} + \mathcal{O}\Big(\frac{1}{N^2}\Big)$$

One coalescence in one generation among the i lineages

Rescaling time, large N and the coalescent

Scale time so that one unit of scaled time corresponds to N generations: $\tau = \frac{t}{N}$.
Probability that the rescaled time T_i is larger than τ (you have i lineages for longer that τ):

$$P(T_i > \tau) = (G_{i,i})^t = (G_{i,i})^{N\tau}$$

With large N (ignoring the N^2 terms):

$$P(T_i > \tau) = (G_{i,i})^{N\tau} \simeq \left(1 - \frac{\binom{i}{2}}{N}\right)^{N\tau}$$

At the limit, $N \rightarrow \infty$:

$$\lim_{N\rightarrow\infty} P(T_i > \tau) = e^{-\binom{i}{2}\tau}$$

Limit of large N, distribution of the coalescence times

We have T_i the (scaled) time til the first coalescence event when there are i lineages.
In the limit of large N:

$$P(T_i > t) = e^{-\binom{i}{2}t}$$

Note that if $f_{T_i}(t) = -\binom{i}{2}e^{-\binom{i}{2}\tau}$, then you have

$$P(T_i > t) = \int_t^\infty f_{T_i}(t')dt' = e^{-\binom{i}{2}t}$$

In other words, T_i is exponentially distributed with mean

$$\frac{2}{i(i-1)}$$

Moreover, the probability that more than two lineages coalesce in a single generation can be neglected, so T_i is the time from i to $i-1$ lineages.

Classic models, large N and the coalescent

Kingman: As the population size N tends to infinity, and with the appropriate rescaling:

- WF $\xrightarrow[\text{time rescaled \& } N\to\infty]{}$ coalescent

See also Wakeley Chapter 3:

- **Moran** $\xrightarrow[\text{time rescaled \& } N\to\infty]{}$ coalescent
- A whole range of population models* $\xrightarrow[\text{time rescaled \& } N\to\infty]{}$ coalescent

*in particular "**exchangeable**-type" population models, i.e., number of offspring of an individual in different generations are independent.

In biological terms: reproductive capacities of every individual in every generation are the same, no transmission of reproduction potential. It must be possible to reassign labels (fitness of alleles, geographic locations, etc) without effect.

the coalescent

The coalescent is continuous-time Markov process, which models the genealogy of a sample of n individuals (genes) as a random bifurcating tree, where the $n-1$ coalescence times:

$$T_2, T_3, ..., T_{\mathbf{n}}$$

are mutually independent, exponentially distributed random variables

Each pair of lineages coalesces independently at rate 1, so the total rate when there are i lineages is i choose 2: $\binom{i}{2}$.

coalescence times

Kingman: As the population site N tends to infinity, the coalescence times T_i are independent and exponentially distributed.

$$f_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}$$

$$t \geq 0; i = 2, ..., n$$

where n is the sample size.

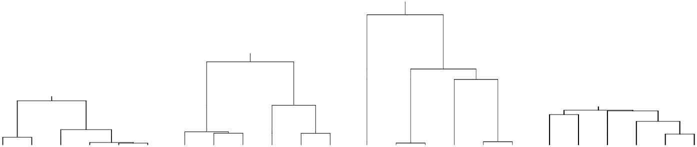
T_i is exponentially distributed with mean

$$\frac{2}{i(i-1)}$$

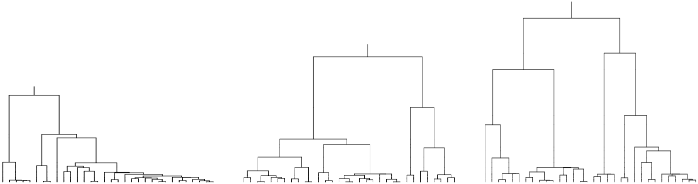
Moreover, the probability that more than two lineages coalesce in a single generation can be neglected, so T_i is the time from i to $i-1$ lineages.

This is very powerful! In particular, exponentials are very easy to work with mathematically and the distribution of the T_i do not depend on the shape of the tree (what comes, before or after).

Realization of the coalescent



The most ancient coalescence time will be the longest.



Increasing the sample size adds only twigs to the tree. Important consequence of this is that increasing the sample size is often ineffective.

Mutations

We cannot observe the genealogies.

We can observe the result of mutations: variations along DNA sequences.

The mutations can inform us about the underlying shape of the tree.

Mutations: shape of the tree does not depend on mutations!

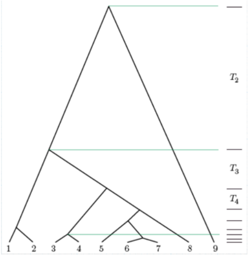
Assuming neutrality, the shape of a coalescent tree is independent of mutation.

If all mutations entering a population are neutral, they do not affect fitness.

They have no impact on whether individuals have greater or fewer numbers of offspring.

Adding mutations to the tree

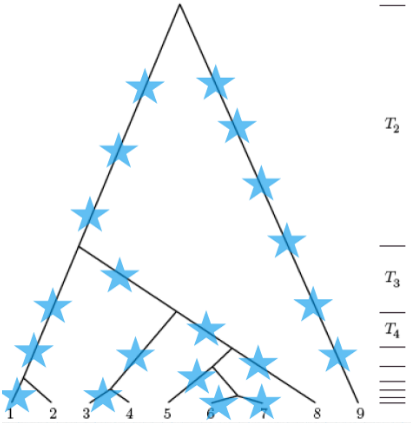
Mutations are independent of tree shape under neutrality.
Throw mutations on the tree.



The longer the length of the branch, the more mutations can "hit" it.

Adding mutations to the tree

The more mutations I have on this tree, the closer we get to understanding the true branch lengths.



The mutation rate

- Let the per-generation probability that an allele (e.g. base) mutates to another allele (base) be μ. The probability of no mutations τ generations back in time is given by the geometric:

$$(1-\mu)^{\tau}$$

- Typically, μ is small, and we can rescale as we did for the coalescence times. We define the rescaled mutation rate as:

$$\mu = \frac{\theta}{2N} \to \theta = 2N\mu$$

- The probability of no mutation in t units of scaled time ($t = \frac{\tau}{N} \to \tau = Nt$) is:

$$(1-\mu)^{Nt} = (1-\frac{\theta}{2N})^{Nt} \xrightarrow[N\to\infty]{} e^{-\theta t/2}$$

- The time until the first mutation is exponentially distributed with rate $\frac{\theta}{2}$

Number of mutations

Mutations occur at a rate $\theta/2$ per unit of time.

Number of mutations per unit time $\sim$ Poisson($\theta/2$)

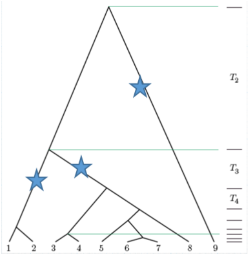
For a fixed time length t ("piece of genealogy"), the number of mutations is the sum of t independent Poisson($\theta/2$)

This (number of mutations) is in itself a Poisson distributed with mean ($t\theta/2$)

$$P(K = k|t) = \frac{\left(\frac{\theta t}{2}\right)^k}{k!} e^{-\frac{\theta t}{2}}$$

SFS: Adding mutations to the tree and allele frequencies

Throw mutations on the tree.

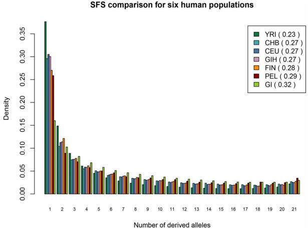


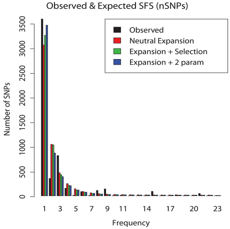
Three mutations:

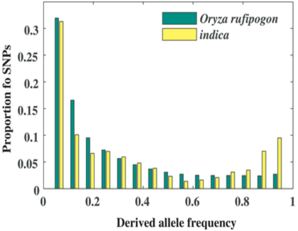
- we have one present in 1 individual,
- we have one present in 2 individuals,
- we have one present in 6 individuals.

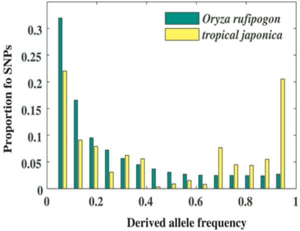
the Site Frequency Spectrum (SFS)

SFS $= \vec{\xi} :=$ count of the number of sites that have i copies of the mutant allele. i can take values from 1 to $n-1$. The SFS is an array, often visualised as a histogram:









Refs: Pedersen et al. , Genetics, 2017; Caicedo et al, Plos Genetics, 2007; Pepperell et al., Plos Pathogens, 2013

the Site Frequency Spectrum (SFS)

SFS $= \vec{\xi} :=$ count of the number of sites that have i copies of the mutant allele. i can take values from 1 to $n-1$. The SFS is an array.

$$\vec{\xi} = (\xi_1, \xi_2, ..., \xi_{n-1})$$

What are the expected counts for the number of sites where we have the allele in i copies?
This will be proportional to the branches with i 'subtended leaves' ("children at present"). Denoting such branches τ_i. We have:

$$E[\xi_i] = \frac{\theta}{2} E[\tau_i]$$

Turns out, $E[\tau_i] = \frac{2}{i}$, we therefore have:

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

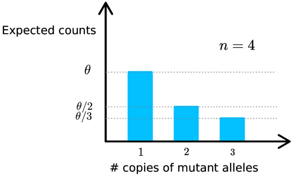
Shape of the SFS

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

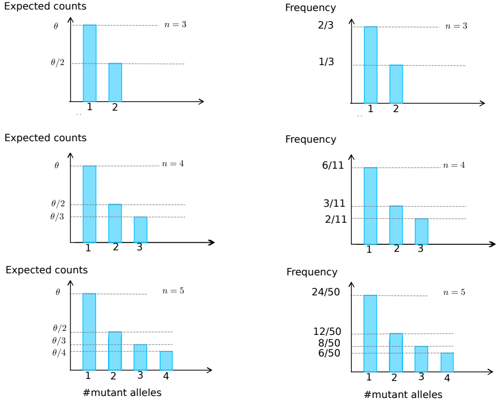
There are: θ singletons, $\theta/2$ doubletons, $\theta/3$ tripletons...
In general, for n samples

$$E[\vec{\xi}] = (E[\xi_1], E[\xi_2], ..., E[\xi_{n-1}]) = (\theta, \theta/2, \theta/3, ..., \theta/(n-1))$$

Example for $n = 4$:

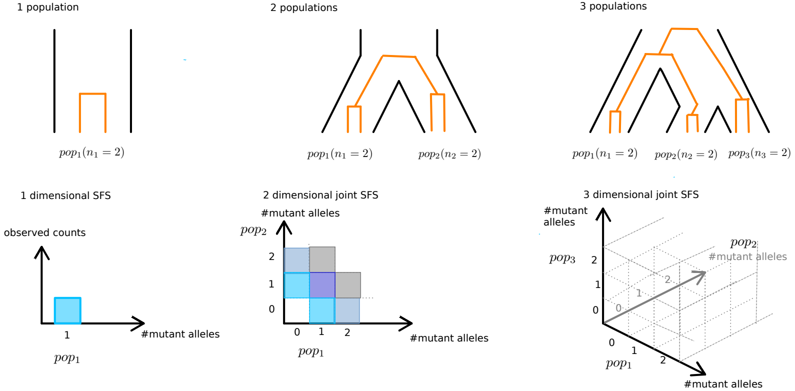


SFS is often renormalized (counts are replaced by frequencies)



Examples of the expected renormalized SFS for $n = 2, 3, 4$ haploid individuals (chromosomes) under the standard neutral coalescent model with the infinite site mutation model.

The joint (multidimensional) Site Frequency Spectrum



1 population

2 populations

3 populations

1 dimensional SFS

2 dimensional joint SFS

3 dimensional joint SFS

$(\xi_i)_{i\in(0\ldots n_1)}$

$(\xi_{i,j})_{i\in(0\ldots n_1),j\in(0\ldots n_2)}$

$(\xi_{i,j,k})_{i\in(0\ldots n_1),j\in(0\ldots n_2),k\in(0\ldots n_3)}$

Shape of the SFS, diversity, number of SNPs and demography

The observed :

- frequency of variants sequenced in a population (SFS),
- diversity,
- number of SNPs,
- ...

tell you about the **underlying shape** of the coalescent tree and the relationships between samples, populations

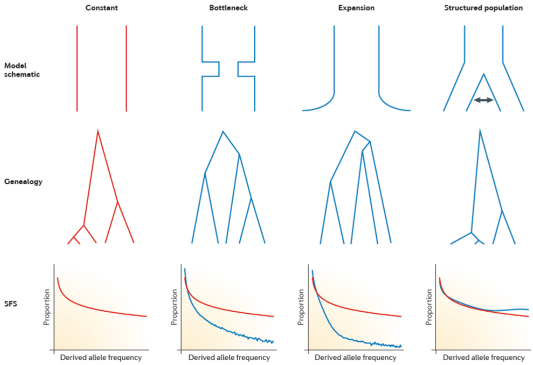
Demographic inference: overall idea

Specific demographic models result in specific coalescent trees. Coalescent trees can be inferred from the summary statistics (e.g., the SFS).

Statistical inference and modeling:
Which type of population history may explain observed levels and patterns of variation?

$$P(\text{data}|\text{model})$$

Demography, tree shapes: SFS



Source: Schraiber & Akey 2015 (Nat Rev Genetics)

Summaries statistics and the Site Frequency Spectrum

The three measures

- S_n: number of segregating sites, SNPs
- π: diversity
- D-statistics

are all functions of the site frequency spectrum.

S_n, π are **linear** functions of the site frequency spectrum ξ_i.

D-statistics (which is based on 4 populations) is a (**non linear**) function of the **joint** 3 dimensional SFS:=
$(\xi_{i,j,k})_{i\in(0,1),j\in(0,1),k\in(0,1)}$.

Four summary statistics

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

- Diversity π
  $E[\pi] = \theta$

- SFS
  $E[\xi_1], E[\xi_2], ..., E[\xi_{n-1}] =$
  $\theta, \frac{\theta}{2}, ..., \frac{\theta}{n-1}$

- D-statistic
  $D = \frac{\xi_{0,1,1} - \xi_{1,0,1}}{\xi_{0,1,1} + \xi_{1,0,1}}$

Take-home message

The coalescent is a powerful model to infer demographic (evolutionary) parameters.

Textbooks

- Nielsen, R., Slatkin, M., 2013. An Introduction to Population Genetics: Theory and Applications, 1 edition. ed. Sinauer Associates, Inc., Sunderland, Mass.
- Wakeley, J., 2008. Coalescent Theory: An Introduction, 1st Edition edition. ed. Roberts and Company Publishers, Greenwood Village, Colo.
- Norborg, M., Chapter 5: Coalescent Theory in Balding, D.J., Moltke, I., Marioni, J. (Eds.), 2019. Handbook of Statistical Genomics, 4th edition. ed. Wiley, Hoboken, NJ.

Some simluations software

ms
Hudson, R. R. "Generating Samples under a Wright-Fisher Neutral Model. Bioinformatics" 18, no. 2 (February 2002): 337–38.
`http://home.uchicago.edu/~rhudson1/source/mksamples.html`

scrm
Staab, Paul R., Sha Zhu, Dirk Metzler, and Gerton Lunter. "Scrm: Efficiently Simulating Long Sequences Using the Approximated Coalescent with Recombination." Bioinformatics 31, no. 10 (May 15, 2015): 1680–82. https://doi.org/10.1093/bioinformatics/btu861.

msms
Ewing, Gregory, and Joachim Hermisson. "MSMS: A Coalescent Simulation Program Including Recombination, Demographic Structure and Selection at a Single Locus." Bioinformatics 26, no. 16 (August 15, 2010): 2064–65. https://doi.org/10.1093/bioinformatics/btq322.

msprime
Kelleher, Jerome, Alison M. Etheridge, and Gilean McVean. "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes." PLOS Computational Biology 12, no. 5 (May 4, 2016): e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

Acknowledgments

Magnus Norborg

Jeffrey Jensen

Rasmus Nielsen

Yun S. Song