HST.508/Biophysics 170: Quantitative genomics
Module 1: Evolutionary and population genetics
Lecture 2: the coalescent & what to do with it

Professor Robert C. Berwick

Topics for this module

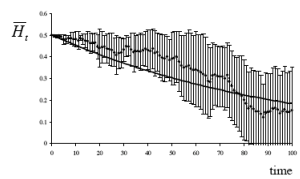
1. The basic forces of evolution; neutral evolution and drift
2. Computing 'gene geneaologies' forwards and backwards; the coalescent; natural selection and its discontents
3. The evolution of nucleotides and phylogenetic analysis
4. Measuring selection: from classical methods to modern statistical inference techniques

But first, a few more words about drift…

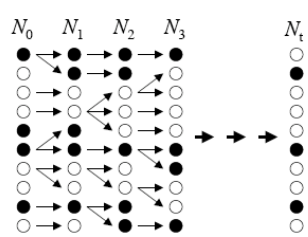
The key to evolutionary thinking: follow the money;

money= variation

- We saw how the Fisher-Wright model lets us keep track of variation (= differences, *heterozygosity*) going *forward in time,* alternatively, similarity, *homozyogosity*)
- Second we can add in the drip, drip of mutations and see what the account ledger balance says



Last Time: The Wright-Fisher model & changes in expected variability



We get a binomial tree that depends on frequency, p, and total population size, N.

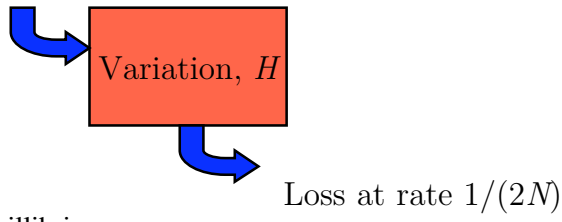
→ Binomial sampling $\Pr\{j \mid i\} = \frac{2N!}{j!(2N-j)!}\left(\frac{i}{2N}\right)^j\left(1-\frac{i}{2N}\right)^{2N-j}$

What is the pr that a particular allele has at least 1 copy in the next generation?
Well, what is the pr of *not* picking an allele on *one* draw?
Ans: 1-(1/2N). There are 2N draws (why?). So, pr of *not* picking for this many draws is $[1-(1/2N)]^{2N} = e^{-1}$ for large N

Let's explore the consequences…

Adding mutations – the mutation-drift balance

Mutation gain $2Nu$



$\Delta H = 0$ at equillibrium, so

$$\hat{H} = \frac{4Nu}{1+4Nu}$$

$4Nu = \theta$ basic level of variation

The forces of evolution…

$$E[H] = \frac{4N_e u}{1+4N_e u}$$

Goal: understand relation between forces: u, $1/N$

signal → u

noise → $\frac{1}{N}$

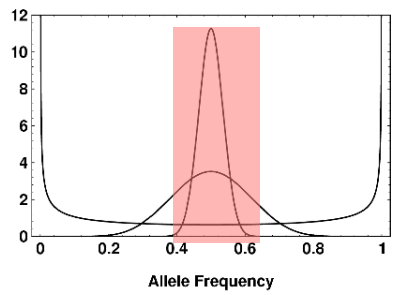
$> ?1$

Mutation vs. drift: the key number is $4N\mu$ vs. 1

$N\mu > 1$, diversity *increases*
heterozygosity maintained around 0.5

Gain heterozygosity →
variance stays high

Population "large" wrt genetic drift



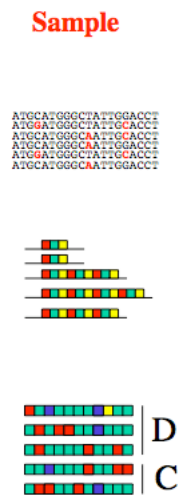
"Follow the variation"

Heterozygosity= $\hat{H} = \dfrac{4Nu}{1+4Nu}$ $\quad 4Nu = \theta$

Homozygosity (identity)= $1-H = G = 1/\theta$

<u>These are the key measures of how</u> 'variant' two genes (loci), sequences, etc. are
What can we learn about their distributions?
How can we estimate them from data?
How can we use them to test hypotheses about evolution?

The F measure already
tells us something about expected variation



$G = 1/\theta = 1/4Nu =$ *measure* that 2 sequences (or alleles, or…) differ in exactly *zero* ways

Compute π = nucleotide diversity = # of diffs in 2 sequences (informally)

What is E[π]?

We shall see that E[π]= $4Nu$ *ie,* θ

But why <u>this</u> pattern of variation?
Drift? Mutation? Selection? Migration?

"Follow the variation": some famous data about individual variation in *Drosophila melanogaster* (Marty Kreitman)

Table removed due to copyright reasons.

Kreitman 1983 original data set for melanogaster Adh sequences
Kreitman,M (1983): Nucleotide polymorphism at the alcohol
#ehydrogenase locus of Drosophila melanogaster.
Nature 304, 412-417.

Kreitman data

11 alleles; 14 sites polymorphic
1.8 every 100 sites segregating
(typical for *Drosophila*)
Variation in 13 out of 14 silent; position #578 is a replacement polymorphism

Q: why this pattern of variation?
Q: is 11 alleles a big enough sample?
(The answer is Yes, actually, as we shall see)

The key to the bookkeeping of evolution is: Follow the money – keeping track of *variation*

Because this is a binomial draw with parameters p, $2N$, the mean of this distribution (the expected # of A_1 alleles drawn) is just $2Np$, i.e., mean frequency is p
And its variance is $2Np(1-p)$
What abour the mean and variance <u>not</u> of the # of alleles, but of the frequency itself, p' ?

$$E[p'] = E[X]/2N = 2Np/2N = p$$

The variance of p' <u>goes down</u> as the population size increases, as we would expect:

$$\mathrm{Var}[p'] = \mathrm{Var}[X]^2/4N^2 = 2Np(1-p)/4N^2 = p(1-p)/2N$$

Key point: drift is important <u>when the variance is large</u>

Second consequence: new mutations, if neutral…

What is the probability that a particular allele has at <u>least 1 copy</u> in the next generation? In other words: that a brand-new mutation survives?

Well, what is the *pr* of *not* picking an allele on *one* draw?
Ans: $1-(1/2N)$. There are $2N$ draws (why?).
So, pr of *not* picking for this many draws is:

$$[1-(1/2N)]^{2N} = e^{-1} \text{ for large } N$$

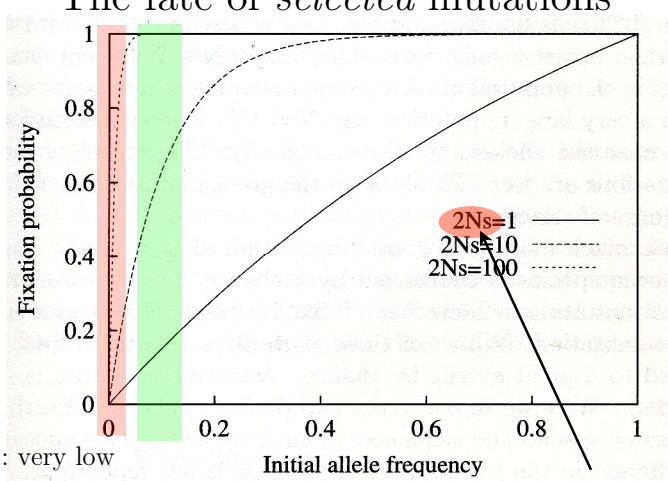
So: probability of a new mutation being lost simply due to 'Mendelian bad luck' is $1/e$ or 0.3679

Why doesn't population size N matter?
Answer: it's irrelevant to the # of offspring produced initially by the new gene

Climb every mountain? Some surprising results

- The power of selection: what is the fixation probability for a new mutation?
- If no selection, the ***pr* of loss in a *single generation* is 1/*e* or 0.3679**
- In particular: suppose new mutation has 1% selection advantage as heterozygote – this is a *huge* difference
- Yet this will have only a 2% chance of ultimate fixation, starting from 1 copy (in a *finite* population a Poisson # of offspring, mean $1+s/2$, the Pr of extinction in a *single generation* is $e^{-1}(1-s/2)$, e.g., **0.3642** for $s = 0.01$)
- Specifically, to be 99% certain a new mutation will fix, for $s = 0.001$, we need about 4605 allele copies (independent of population size N !!)
- Also very possible for a *deleterious* mutation to fix, if $2Ns$ is close to 1
- Why? Intuition: look at the shape of the selection curve – flat at the start, strongest at the middle
- To understand this, we'll have to dig into how variation changes from generation to generation, in finite populations

The fate of *selected* mutations



Regime 1: very low copy #

Regime 2: Frequency matters

$2Ns$ (compare to Nu factor)

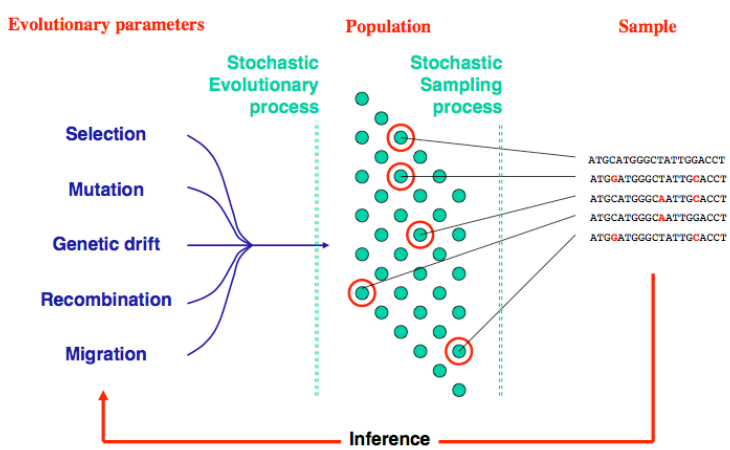
Fixation probability of a (neutral) allele is *proportional* to its initial frequency

All variation is ultimately lost, so eventually 1 allele is ancestor of all alleles
There are $2N$ alleles
So the chance that any one of them is ancestor of all is $1/2N$

If there are i copies, the ultimate chance of fixation (removal of all variation) is $i/2N$

(Simple argument because all alleles are equivalent – there is no natural selection)

Population genetic inference



What are we missing? History.

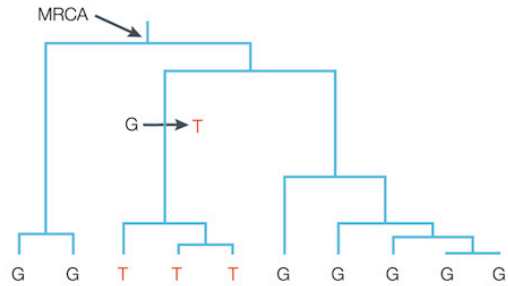


Figure 1 | **The source of genetic variation.** Polymorphism at a particular site results from mutations (shown here as G→T) along branches of the genealogical tree, which connects sampled copies of the site to their most recent common ancestor (MRCA).

The 3 mutations are *not* independent – increasing sample size n does *not* have the usual effect of improving accuracy of estimates! (In fact, it's only marginally effective)

$$\approx \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)G - 2uG$$

$$H = 1 - G, \text{ so } H' \approx \left(1 - \frac{1}{2N}\right)H + 2u(1 - H)$$

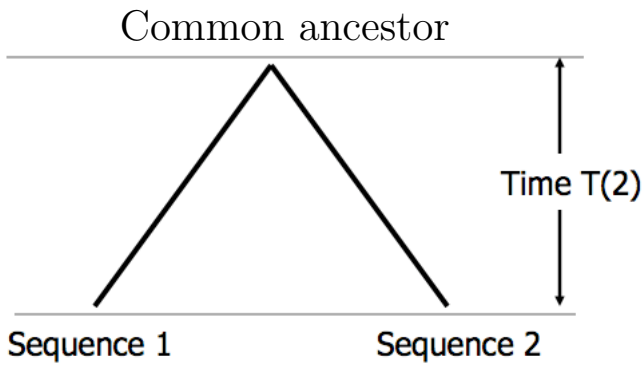
$$\Delta H \approx -\frac{1}{2N}H + 2u(1 - H)$$

$\Delta H = 0$ at equillibrium, so

Heterozygosity= $\hat{H} = \dfrac{4Nu}{1 + 4Nu}$
(AKA gene diversity)

$4Nu = \theta$ basic level of variation

The coalescent:
The *cause* of the decline in variation is that all lineages eventually coalesce…



Notation: T_i= time to collapse of i genes, sequences,…
This stochastic process is called <u>the coalescent</u>

Coalescent can be used for...

... simulation
... hypothesis testing
... estimation

Random genealogical trees. The trees were generated using the same model — the standard coalescent for sample of size ten. Therefore, the variation among the trees reflects chance alone.

Looking backwards: the coalescent

A *coalescent* is the lineage of alleles in a sample traced backward in time to their common ancestor allele

More useful for inference: we see a certain pattern of data, want to understand the processes that produced that data

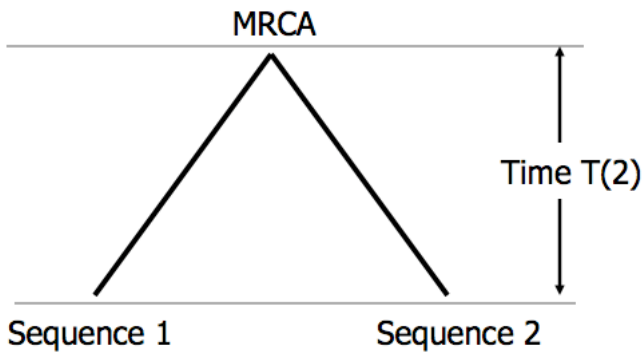
NB, we cannot actually know the coalescent (but who cares?)

Provides intuition on patterns of variation

Provides analytical solutions

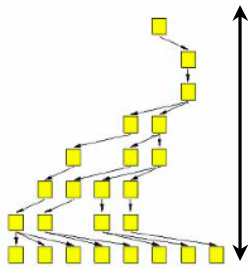
Key: We need only model genealogy of samples: we don't need to worry about parts of population that did not leave descendants (as long as mutations are neutral)

What is time to most recent common ancestor? (MRCA)?



Notation: T_i= time to collapse of i genes, sequences,…

In other words…



On average, depth $2N$ before collapse to 1 ancestor

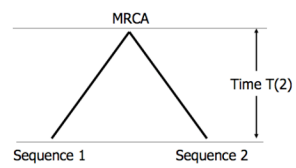
Can we prove this and use it?
If it's true, then we can use this to get expected sequence diversity, estimates of # of segregating sites, heterozygosity, and much, much more…

Pr that two genes differ (*ie*, *H* as before…)

$$H = \frac{P(\text{mutation})}{P(\text{mutation})+P(\text{coalescence})} = \frac{2u}{2u+\frac{1}{2N}} = \frac{4Nu}{4Nu+1}$$

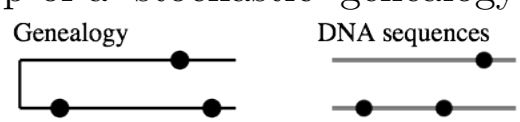
Q: where did *2u* come from?

For example, if $u = 10^{-6}$, then in a population of 10^6, mean heterozygosity expected is 0.8



This is a <u>lot easier</u> to compute than before!!!

We superimpose (neutral) mutations on top of a 'stochastic' genealogy tree



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

This product is our θ

Can we estimate it?

<u>Note</u> that *each* mutation in a <u>coalescent lineage</u> produces a distinct *segregating site* (Why?)

Why can we superimpose these 2 stochastic effects?

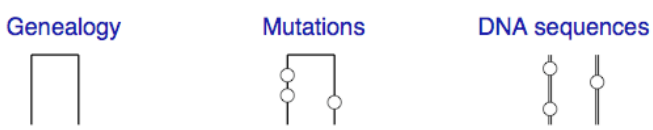
Because mutations don't affect reproduction (population size)

Basic idea

- More parents, slower rate to coalesce
- Neutral mutations don't affect reproduction (N) so can be superimposed afterwards on the gene tree

Now we can get the basic 'infinite site' result for expected # diffs in DNA seqs:

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



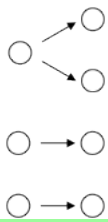
- Expected number of differences between a pair of sequences

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

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

$$\theta = 4Nu$$

Expected time to coalescence



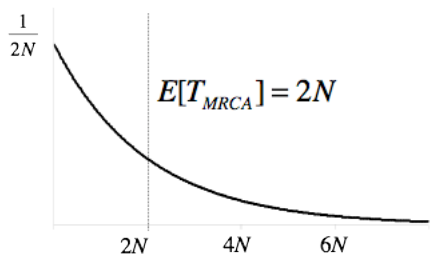
Probability from same parent (coalescence) $= \frac{1}{2N}$

Probability from different parents $= 1 - \frac{1}{2N}$

Probability of coalescence t generations ago $= \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$

Not coalesced for first t-1 generations

Coalesce in next generation

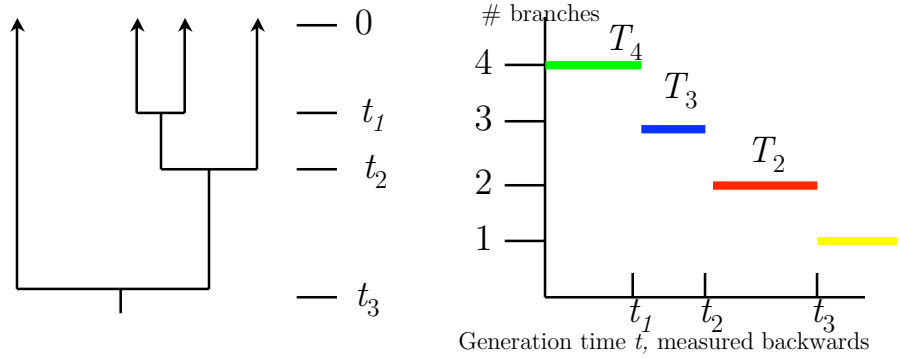


63% of outcomes have $T_{MRCA} < 2N$

Using the coalescent as a 'history model,' expectations can be derived either in a discrete time model (Fisher-Wright) or in a continuous time model

The discrete model yields a 'geometric' probability distribution
The continuous time model yields an 'exponential' probability distribution of 'waiting times' until each coalescence

An example coalescent for <u>four</u> alleles



<u>Total time in coalescent</u> $T_C = 4t_1 + 3(t_2 - t_1) + 2(t_3 - t_2) = 4T_4 + 3T_3 + 2T_2$

# of expected mutations is uT_C

What is the expected value of T_C?

Discrete time argument to find expected coalescent time, for n alleles

Allele 1 has ancestor in 1st ancestral generation

Allele 2 will be <u>different</u> from 1 with probability

$$1 - 1/2N = (2N-1)/2N$$

Allele 3 will be different from first 2, assuming alleles 1 and 2 are distinct, with probability:

$$(2N-2)/2N$$

So total probability that the <u>first three</u> alleles do not share an ancestor is:

$$(2N-1)/2N \text{ x } (2N-2)/2N$$

Probability all n alleles <u>do not</u> share an ancestor (*no* coalescence) is (dropping N^{-2} and higher terms):

$$(1-\frac{1}{2N})(1-\frac{2}{2N})\cdots(1-\frac{n-1}{2N}) \approx 1-\frac{1}{2N}-\frac{2}{2N}-\cdots\frac{n-1}{2N}$$

$$\Pr \text{ no coalescence } \approx 1 - \frac{1}{2N} - \frac{2}{2N} - \cdots \frac{n-1}{2N};$$

$$\Pr \text{ coalescence in any particular generation}$$

$$\approx \frac{1+2+\cdots+(n-1)}{2N} = \frac{n(n-1)}{4N}$$

So: Time to 1st coalescence is <u>geometrically</u> distributed with pr success of $n(n–1)/4N$

Mean of geometric distribution, is this reciprocal of success:

$$\mathrm{E}[T_n] = 4N/n(n–1)$$

So,

$$\mathrm{E}[T_2] = 4N/2 = 2N$$

$$\mathrm{E}[T_i] = 4N/i(i–1)$$

(coalescence time from i alleles to i-1)

Note: we *do not really care* about the trees – they are a 'nuisance' parameter

Here's another way to look at it: when there are 4 alleles, we have to pick 2 of them to 'coalesce' or merge… so there are 4 choose 2 ways of doing this, out of $2N$ *possible* alleles. This gives the Pr of Coalescent event, as follows.

The *time* to the next Coalescent Event is the reciprocal of this numbe so:

Probability of Coalescent Event

$$P(4) \approx \binom{4}{2} \Big/ 2N$$

Time to Next Coalescent Event

$$T(4) \approx 2N \Big/ \binom{4}{2}$$

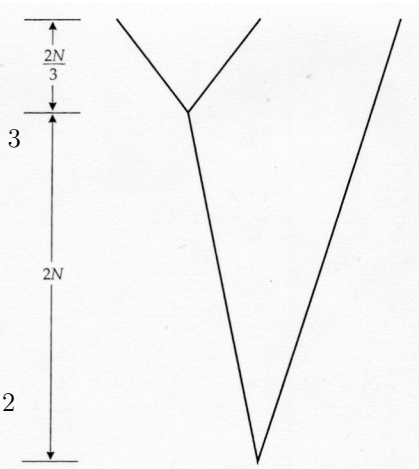
Note typical shape and amount of time at tips of tree!

$$\frac{2N}{\binom{3}{2}}$$

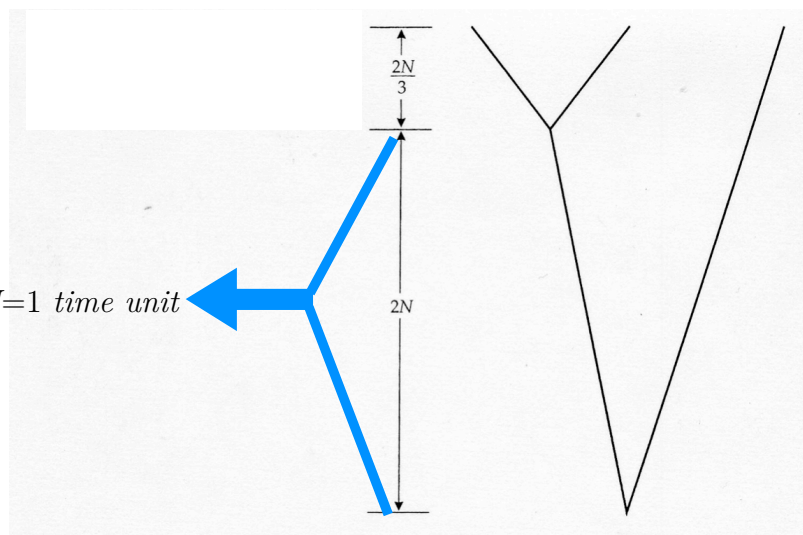
Shorter waiting time because only 3 choose 2 genes/sequences could 'collide'

$$\frac{2N}{\binom{2}{2}}$$

Longer waiting time because only 2 genes/sequences left to 'collide'

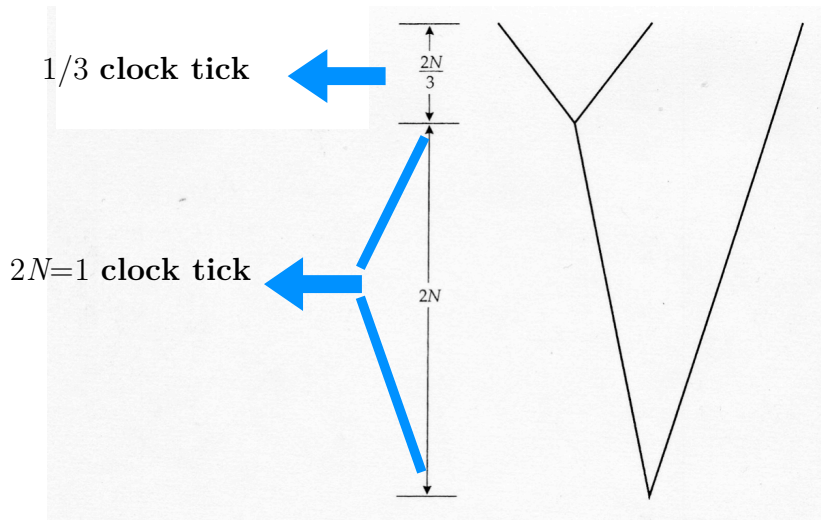


Rescaling time in terms of generational units



$2N$=1 *time unit*

Rescaling time in terms of generational units



Now we can actually get some results…!

Total time in all branches of a coalescent is:

$$T_C = \sum_{i=2}^{n} iT_i$$

i is just the # of 'mergers', ie 1 less than # of alleles at tips

So expected Total time in *all* branches is:

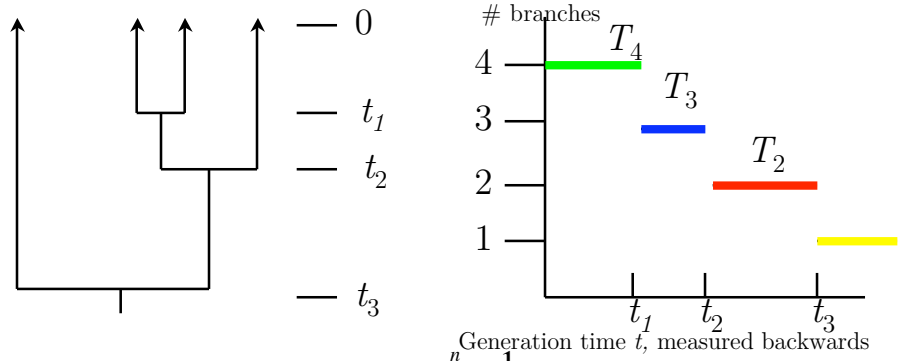
$$E[T_C] = \sum_{i=2}^{n} iE[T_i] = 4N\sum_{i=2}^{n}\frac{1}{i-1}$$

Expected # segregating sites is neutral mutation rate, u times the expected time in coalescent, therefore:

$$E[S_N] = uE[T_C] = \theta\sum_{i=2}^{n}\frac{1}{i-1}$$

$$\hat{\theta} = \frac{S_n}{1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n-1}}$$

Application to our example coalescent for <u>four</u> alleles



$$E[T_C] = 4N\sum_{i=2}^{n}\frac{1}{i-1}$$

<u>Total time</u> $T_C = 4N(1+1/2+1/3)=44N/6$
# of expected mutations is uT_C or $\theta(11/6)$ or 1.83 θ in a sample of 4 alleles, which is also the expected # of segregating sites

Application to Kreitman SNP data

<u># segregating sites:</u> 14
<u>Sample size:</u> n=11

$$\hat{\theta} = \frac{S_n}{1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n-1}} = \frac{11}{2.93} = 4.78\ (4Nu) \text{ for locus}$$

$$\hat{\theta} \text{ for nucleotide site} = \frac{4.78}{768} = 0.0062$$

What about sample size question?
Well, note:

$$E[S_N] = \theta\sum_{i=2}^{n}\frac{1}{i-1}, \text{ and } \sum_{i=1}^{n}\frac{1}{i} \approx \ln(n),$$

so # segregating sites increases with log of sample size

Another estimator for theta

Use E[π], # pairwise differences between 2 sequences (In a sample of size n, there are $n(n-1)$ pairwise comparisons.)

This is $2u$E[t], where E[t] is mean time back to common ancestor of a random pair of alleles, i.e., $2N$, so E[π]=θ

Let's apply this to an actual example, to see how π and θ might be used…

Summary statistics

- Good properties of summary statistics
  - Include most (all) information in the data
  - Different statistics should use different information
  - Expectations and variances should have simple relationship to model parameters

Statistic	Symbol	Expectation
Average pairwise diversity	$\pi = \frac{2}{n(n-1)} \sum_{ij} \pi_{ij}$	θ
Number of segregating sites	S	$\theta \sum_{i=1}^{n-1} 1/i$
Number of haplotypes	K	$\theta \sum_{i=0}^{n-1} 1/(i+\theta)$ (no recombination)

Example – control region of human mtDNA

Jorde et al (ref) published sequence data from the control region of human mitochondrial DNA. The example described here uses 430 nucleotide positions from HVS1 (the first hypervariable region). Jorde et al sequenced DNAs from all three major human racial groups, but this example will deal only with the 77 Asian and 72 African sequences. In these data:

	Asian	African
S	82	63
$\sum_{i=1}^{n-1} 1/i$	4.915	4.847
$\hat{\theta}_S$ (per sequence)	16.685	12.998
π (per sequence)	6.231	9.208

To compare statistics referring to sequences of different lengths, it is often convenient to divide by the number of sites, which produces:

	Asian	African
$\hat{\theta}_S$ (per site)	0.039	0.030
π (per site)	0.014	0.021

The key question (as usual): Why the differences between these two supposedly equivalent estimates??

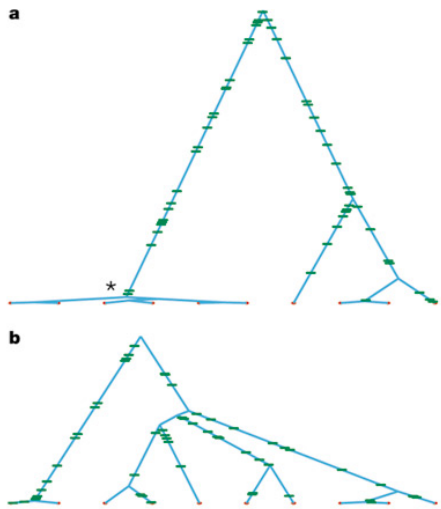
?? Sampling error??

?? Natural selection?? In fact, we can use the <u>difference</u> between these estimators to test for this (Tajima's D)

?? Variation in population size/demographics?? We've assumed constant N. Need to incorporate changing N, migration, etc.

?? Failure of mutation model?? We've assumed mutation never strikes the same nt position twice

Q: How do we get sampling error? A: coalescent simulation



How you do a coalescent simulation

When n = 4

Probability of Coalescent Event

$$P(4) \approx \binom{4}{2} \Big/ 2N$$

Time to Next Coalescent Event

$$T(4) \approx 2N \Big/ \binom{4}{2}$$

Sample time from exponential distribution

Pick two sequences at random to coalesce

Now we could use this spectrum to test our hypotheses about the model assumptions we made

Frequency Spectrum

- Constant size population
- Exponentially growing population

- Most variants are rare
  - For n = 100, ~44% of variants occur < 5/100.
  - For n = 10, ~35% of variants observed once.

Deviations from Neutral Spectrum

- When would you expect deviations from the spectra we described?

- What would you expect for …
  - A rapidly growing population?
  - A population whose size is decreasing?

- Why?

Intuition behind the continuous time model: life-span of a cup

Intuition: if pr breaking is h per day, and expected life-span is T days; show that T is $1/h$ $(= 1/2N)$

Same as 'coalesence' between 2 genes

Cup either breaks 1st day w/ pr h or doesn't with pr $1-h$; gene either coalesces or doesn't. If it breaks 1st day, mean life-span is 1

For surviving cups, life-span doesn't depend on how old it is, so if a cup has already lived a day, expected life-span is now $1+T$. So:

$$T = h + (1-h)(1+T) = 1/h$$

A bit more formally…

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

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

P_{NC} for t generations: $(P_{NC})^t = \left(1 - \frac{1}{2N}\right)^t$

P_{NC} for t generations and then coalescing in $t+1$:

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

Continuous time

If $2N$ large, > 100, use Taylor series expansion for e:

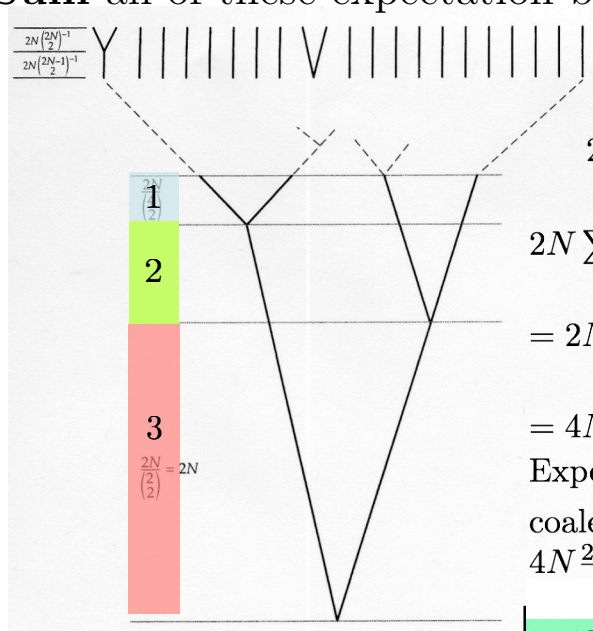
$e^{-1/2N} \approx \left(1 - 1/2N\right)$ so

$$P_{C,t+1} = \frac{1}{2N} e^{-\frac{t}{2N}}$$

exponential distribution for large t,

so $P[x] = 1/b \bullet e^{-xb}$ with mean b, variance b^2

Sum all of these expectation bars...



9 Coalescent diagram for all allele copies in a population.

$$2N \sum_{k=2}^{2N} \frac{1}{\binom{k}{2}}$$

$$2N \sum_{i=2}^{2N} \frac{1}{\binom{i}{2}}$$

$$= 2N \sum_{i=2}^{2N} \frac{2(i-2)!}{i!}$$

$$= 4N \sum_{i=2}^{2N} \frac{2}{i(i-1)}$$

Expected time to coalescence of all alleles= $4N\frac{2N-1}{2N}$

$$\boxed{\approx 4N}$$

Summary: the coalescent models the geneology of a sample of n individuals as a random bifurcating tree
The n-1 coalescent times $T(n)$, $T(n\text{-}1)$, ..., $T(1)$ are mutually independent, exponentially distributed random variables

<u>*Rate*</u> of coalescence for two lineages is (scaled) at 1
<u>*Total rate*</u>, for k lineages is 'k choose 2'

Basic references:

[1] Richard R. Hudson. Gene genealogies and the coalescent process. In Douglas Futuyma and Janis Antonovics, editors, *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44. Oxford University Press, Oxford, 1990.

[2] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

Summary equations

Coalescence Times (in 2N units)

$$E(T_j) = 1/\binom{j}{2}$$

Total Length (in 2N units)

$$E(T_{tot}) = \sum_{i=1}^{n-1} \frac{2}{i}$$

Number of Mutations

$$E(S) = 4N\mu \sum_{i=1}^{n-1} 1/i = \theta \sum_{i=1}^{n-1} 1/i$$

Extensions

- Add migration
- Population size fluxes ('bottlenecks')
- Estimation methods – based on likelihoods

Let's deal with population size issue: *effective* population size

Suppose population size fluctuates. For instance, in one generation, population size is N_1 with probability r, the next it is N_2 with probability $1-r$

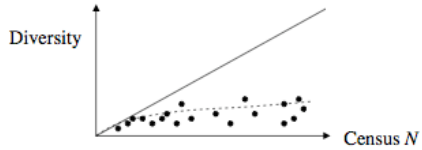
Can we patch up the formula?

General answer: Yes, we replace N with N_e – the <u>effective</u>

population size

Let's see what this means in flutating population size case

- Levels of polymorphism vary less between species than the census population size



- The rate of genetic drift varies due to
  - Inbreeding, skewed sex ratios, fluctuating population size, variation in family size

- Many biologically realistic complications can be modelled by a coalescent process with a smaller EFFECTIVE population size

$$N \rightarrow N_e \qquad E[\pi] = 4N_e u \qquad \theta = 4N_e u$$

Effective population size must be used to 'patch' the Wright-Fisher model

Variance for N_1 is $p(1-p)/2N_1$ with probability r
Variance for N_2 is $p(1-p)/2N_2$ with probability $1-r$
Average these 2 populations together, to get mean variance, 'solve' for N_e

$$Var[p'] = p(1-p)\left(\frac{r}{2N_1} + \frac{1-r}{2N_2}\right) \text{ or}$$

$$N_e = \frac{1}{r\frac{1}{N_1} + (1-r)\frac{1}{N_2}}$$

i.e., the harmonic mean of the population sizes (the reciprocal of the average of the reciprocals)

Always smaller than the mean

Much more sensitive to small numbers

Effective population size & bottlenecks

Example: if population size is 1000 w/ pr 0.9 and 100 w/ pr 0.1, arithmetic mean is 901, but the harmonic mean is $(0.9 \times 1/1000 + 0.1 \times 1/10)^{-1} = 91.4$, an order of magnitude less!

Thus, if we have a population (like humans, cheetahs) going through a 'squeeze', this *changes* the population sizes, hence θ

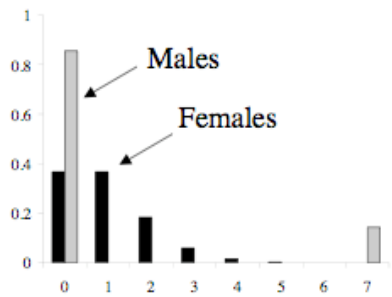
Suppose we have an *arbitrary* distribution of offspring numbers?

Fluctuating population size

$$N_e = \frac{1}{\frac{1}{t}\sum \frac{1}{2N_i}}$$

- Suppose population sizes: 11, 21, 1000, 21, 4000, 45, 6000, 12
- Arithmetic Mean (11+21+1000+21+4000+45+6000+12)/8 = 1389
- Harmonic Mean = 27
- Harmonic Mean is smaller (small values have more important effects!)

Different #s males and females



$$N_f = \tfrac{N}{2}$$

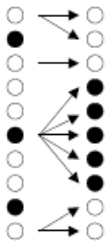
$$N_m \approx \tfrac{N}{12}$$

$$N_e = \frac{4N_m N_f}{N_m + N_f}$$

$$\approx \frac{N}{3.5}$$

Changing population sizes: the <u>effective</u> population size, N_e

Varying offspring #, breeding success, overlapping generations…



Pr{2 alleles from same parent}

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

$$\approx \frac{\sigma_k^2}{2N} \quad \text{If population size constant}$$

$$\boxed{\Rightarrow N_e = \frac{N}{\sigma_k^2}}$$

Where's Darwin????