Lecture 12.11.2013

COALESCENCE THEORY AND SELECTION TESTS

● Population genetics theory gives the basics for understanding how a population evolves under a given set of conditions. Evolution is a forward process: the genetic composition, allele and genotype frequencies change with time.

● Hardy-Weinberg-model, the basic null model, states that no change unless evolutionary factors - selection, genetic drift, mutation, gene flow from other populations – are in action.

● *Prospective population genetics theory* dominated for decades, after the seminal work of Sewall Wright, R.A. Fisher, J.B.S. Haldane, and Motoo Kimura. Although all this work is important and provides strong theoretical framework for understanding populations, current data analysis needs another viewpoint, too.

● In practice, in real situations for the researcher, the characteristics of a natural population (or a human population), are examined by taking samples from the population. Interesting biological questions that arise from a sample are mostly *retrospective*, such as the history of the population that gave rise to the sample, or the evolutionary mechanisms responsible for the characteristics observed.

● The accumulation of DNA sequence data since the 1980s has transformed the mainstream of population genetics research from prospective to retrospective, from demonstration of principles to inference of events that happened in the past.

INRODUCTION AND BACKGROUND

● From this necessity, inferring the past from a sample taken from a present population, a new approach arose: tracing backwards in time to identify events that occurred since the most recent common ancestor of the sample - coalescent theory.

● British mathematician, sir J.F.C. Kingman, published an important paper in 1982 (attached in course webpage).

● Coalescent theory is useful because

● it is sample-based theory and what we study are samples, not the entire population

● of its by-products: development of of effective algorithms for simulating population samples under various ppulation genetics models, allowing various aspects of a model to be examined numerically

● it is particularly suitable for DNA sequence samples which contain off-loadable information about the past.

● Especially interesting is detecting the role of (Darwinian) selection, at least for two reasons

● Stemming from from a natural curiosity about (our) evolutionary past and the basic mechanisms that govern molecular evolution.

● The realization that inferences about selection can provide important functional information. For example, genes that are targeted by selection acting on segregating mutations are more likely to be associated with disease.

● In general, positions in the genome that are under selection must be of functional importance, otherwise selection could not be operating.

GENETIC DIVERSITY, INFINITE-ALLELES MODEL

● Definition of a measure for genetic diversity, i.e. genetic variation (or variability).

● In the simple model case, two alleles (*A* and *a*) segregating at a locus, the expected (HW) proportion of heterozygotes, heterozygosity (*H*), is this measure: 2*pq.* Homozygosity = 1 – *H.*

● The heterozygosity of a gene, i.e. a function of the number of alleles and their relative frequencies, for example:

● A gene coding for a protein of 300 amino acids has a coding sequence 900 nucleotides. Each nucleotide site could be occupied by either A, T, G or C and thus the total number of possible alleles is 4^{900}.

● Let´s make an assumption that every new mutation creates an allele that does not already exist in the population. This is called the infinite-alleles model of mutation, which is – though being simple (and unrealistic) a useful standard.

● In this model two alleles that are identical in sequence are also identical by descent.

● Cf. page 9 in slides ”*Modelling mutations ….*”: Each allele was assigned by a unique label: α_1 , α_2 , α_3 , …, α_{2N} (interest not at their their status as *A* or *a*), each with a frequency of 1/(2*N*), random sampling from the gamete pool → genotypes in generation *t* + 1 . By chance, the two alleles forming a genotype may be replicates of the same allele in the previous generation, for example $\alpha_i\,\alpha_i$ or they may come from different alleles in the previous generation, for example $\alpha_i\,\alpha_j$

AUTOZYGOSITY - ALLOZYGOSITY

● The alleles in the genotype $\alpha_i\,\alpha_i$ are identical by descent because they descend from a single ancestral allele by DNA replication in a previous generation.
● The alleles in the genotype $\alpha_i\,\alpha_j$ might also be identical by descent: subscripts *i* and *j* imply only that did not derive by DNA replication in the immediately preceding generation, but if they derived by DNA replication in some earlier generation, they are identical by descent.

● Autozygosity: A term for alleles which are identical by descent.

● Allozygosity: A term for alleles which are not identical by descent

● In the infinite-alleles model, in which each mutation produces a new allele not previously present in the population, all homozygous genotypes must have alleles that are autozygous.

● To measure the homozygosity, calculating autozygosity is needed.

● Let´s define F_t as the probability that that, in generation *t*, two alleles randomly chosen from the population are identical by descent (autozygous).

● In the following we use notations $\alpha_i\,\alpha_i$ and $\alpha_i\,\alpha_j$ genotypes in generation *t* to derive an expression for F_t in terms of $F_{t\text{-}1}$, *N* and mutation rate *μ.*

INFINITE- ALLELES MODEL

● Consider the genotype $\alpha_i \alpha_i$. What is the probability that this genotype has alleles that are identical by descent?

● The alleles are identical by descent provided that neither allele has mutated in the course of one generation, and so the probability of identity by descent in this case is $(1 - \mu)^2$.

● The same question considering genotype $\alpha_i \alpha_j$.

● These alleles are identical by descent only if two randomly chosen alleles in generation $t - 1$ are identical by descent and if neither allele mutated in the course of one generation. The probability of identity by descent is $F_{t-1} (1 - \mu)^2$.

● Because each of the labelled α´s has the same frequency in the gamete pool, namely $1/(2N)$, the probability of a combination like $\alpha_i \alpha_i$ is $1/(2N)$ and the probability of a combination like $\alpha_i \alpha_j$ is $1 - 1/(2N)$.

● Collecting the above pieces together, the recurrence equation for F_t is

$$F_t = [1 / (2N)](1 - \mu)^2 + [1 - 1/(2N)](1 - \mu)^2 F_{t-1} \qquad (1)$$

● At equilibrium the value of F: the increase in autozygosity from random genetic drift in any generation is offset by the decrease in autozygosity from new mutations: $F_t = F_{t-1} = \hat{F}$

$$\hat{F} = 1 / (1 + 4N\mu) \qquad (2)$$

Negligibly small terms μ^2 and μ / N ignored.

INFINITE- ALLELES MODEL

● The number of alleles, resulting from mutation pressure, increases until F satisfies equation (2), which is the equilibrium value of autozygosity, the probability of identity by descent. Because of the assumption in the infinite-alleles models, that each allele in the population *arises only once,* all genotypes that are homozygotes must also be autozygous: *equation gives also the equilibrium value of the proportion of homozygous genotypes.*

● Above the N (which captures/depicts the amount of genetic drift), of course, refers to effective population size, N_e (see page 12 in slides ”*Modelling mutations…”*

● In population genetics the usual symbol for $4 N_e \mu$ is Θ.

● A genotype that is not homozygous is heterozygous:
the proportion of heterozygous genotypes in a population is

$$1 - \hat{F} = \Theta / (1 - \Theta) \qquad (3)$$

● Equation (24) gives an infinite-alleles model equlibrium which is actually and ”equilibrium”, a dynamic state, *steady state*, in which allele frequencies are always changing, new m utations continue to come into the population, alleles previously present are lost, and alleles that might at one time have been fixed are subject to eventual loss. The population remains at a steady state in the sense that the number of alleles and the homozygosity (autozygosity in the infinite-alleles model) remain stationary.

● If the number of alleles and the level of autozygosity are in steady state, then it is reasonable to assume that there is also a steady-state distribution of allele frequencies.

INFINITE ALLELES-MODEL → EWENS SAMPLING FORMULA

Allele-frequency spectrum in a population

- The joint distribution of allele frequencies, steady state, the most common allele has a frequency of p_1, the next most common p_2 , etc.
- The identity of the most common allele will change with time, i.e. the allele with fr. p_1 is not the same allele all the time.
- In a steady-state population not all alleles are equally frequent, and F is greater than it would be if all alleles were equally frequent.

● Consider now the steady-state allele-frequency spectrum from the point of view of a practical experiment: a sample is taken from a population.

- Let the sample be n genes, and suppose there are k different alleles in this sample. For example, a sample of size n=20 might consist of k=10 unique alleles, with one allele present six times in the sample, another allele four times, two alleles twice, each, and six alleles once, each.

● Regarding the practical achievement population genetics is the paper of Warren Ewens (1972, *The sampling theory of selectively neutral alleles,* Theor. Pop. Biol. 3:87-112). Ewens showed that the expected number k of alleles in a sample of size n is a function of Θ (derivation rather complicated)

$$E(k) = 1 + \Theta/(\Theta+1) + \Theta/(\Theta+2) + \ldots + \Theta/(\Theta+n-1) \tag{4}$$

● This is a kind of statistical tool for evaluating expected (neutral allele spectrum) in a real sample vs. observed allele configuration in a certain real sample from a population. The first statistical test, Ewens-Watterson test, was based on this.

INFINITE-SITES MODEL

● The commonly used statistical tests are based on another model, *the infinite-sites model*, which was developed by Motoo Kimura (1969, 1971). The very famous test, *Tajima D*, was proposed by Fumio Tajima (1989, *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* Genetics 123:585-595). For introducing Tajima D, we first have a look at the infinite-sites model:

● In a long sequence of nucleotides, if the mutation rate is sufficiently low, most sites will be monomorphic, and all polymorphic sites will be segregating for two nucleotides. If the DNA sequence is sufficiently long and the frequency of polymorphic sites low, then most of the time new mutations will occur at sites that were previously monomorphic.

An example sample
of four seqs, 16 sites.

Two types of information:

Segregating sites, S
S = 8 (sites 1,2,5,6,9,10,13,14)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
b	A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
c	G	A	A	A	C	T	T	T	A	G	G	G	T	C	C	C
d	A	G	A	A	T	C	T	T	G	A	G	G	C	T	C	C

Nucleotide mismatches, π
Among the four seqs a-d, there are 6 pairwise comparisons (a-b, a-c, a-d, b-c, b-d, c-d). Each of these combinations compares 16 nucleotide sites, and among 6 parwise comparisons, the number of mismatches is 0, 4, 4, 4, 4, 8. The total number of pairwise mismatches is thus 24 among a total of 6 pairwise comparisons: π = 24/6 = 4.

INFINITE-SITES MODEL

- The properties of the infinite-sites model is worked out with the concepts of segregating sites and nucleotide mismatches.

- Consider a sample of only two sequences. In this case, the number of segregating sites S and the average number of nucleotide mismatches π are identical, because there is only one parwise comparison. For a sample of size 2, the probability that the number of segregating sites equals any number i is

$$\Pr(S=i) = 1/\ (1+\Theta)\ [\Theta\ /\ (1+\Theta)]^{i} \qquad (5)$$

μ ($\Theta = N_e\,\mu$) is the mutation rate across the entire nucleotide sequence.
Formally μ can be considered as the sum of the per-nucleotide mutations rate
across all the nucleotide sites in the sequence.

- A particular case of equation (5) gives the probabaility that two sequences have no mismatches (i=0, i.e. they are identical).

$$\Pr(S=0) = 1\ /\ (1+\Theta) \qquad (6)$$

- Note that this is the same as equation (2) autozygosity in inifinite-alleles model. So, with sample size 2, in both models the probability that the sequences are identical is also the probability of autozygosity.

INFINITE SITES-MODEL

- From equation (5) it can be shown that that the mean and variance in the number of segregating sites, S, are given by $E(S) = \Theta$ and $V(S) = \Theta + \Theta^2$. For sample size *2,* the average number of pairwise mismatches, π, is equal to the number of segregating sites, and so $E(\pi) = \Theta$. The simplifying assumption for variance is that nucleotide sites are completely linked (no recombination). If there is recombination, variance is reduced. Because of this theoretical prediction, the relationship between the mean and variance, in practical data-analysis situations, has been used to make inference about (intragenic) recombination.

- These sampling properties of the infinite-sites model, with neutral evolution, without recombination (the assumptions), were worked out in 1970´s by Geoff Watterson, who derived the expected number of segregating sites in a sample of size n sequences:

$$E(S) = \Theta \sum_{i=1}^{n-1} 1/i \qquad (7)$$

$$V(S) = \Theta \sum_{i=1}^{n-1} 1/i \ + \Theta^2 \sum_{i=1}^{n-1} 1/i^{\,2} \qquad (8)$$

the average number of pairwise mismatches:

$$E(\pi) = \Theta \qquad (9)$$

$$V(\pi) = [(n+1)/(3(n-1))]\Theta + [2(n^2+n+3)/(9n(n-1))]\Theta^2 \qquad (10)$$

INFINITE SITES-MODEL → TAJIMA D STATISTICS

● The expected mean and variance equations are very important as they are are basics for *practical statistical analyses.*

● Corrections to eliminate the dependence on sequence length, L: the expected averages in equations (7) and (9) are divided by L and the variances in equations (8) and (10) are divided by L^2 .
These corrected values are called:

(number of segregating sites →)	nucleotide polymorphism
(number of pairwise mismatches →)	nucleotide diversity

● Let´s define $a = 1+1/2 + 1/3 + \ldots+1/(n-1)$ (the sum in equation (7)) (11)

Then, equation (7) yields the estimate $\Theta = S/a$ (12)
and equation (9) provides a method for estimating Θ based on the average number of pairwise mismatches π, and in this case the estimate is $\Theta = \pi$ (13)

● Tajima´s proposition: The difference between the estimates of Θ in equations (12) and (13) could be used as a test of goodness of fit to the model. This test is extremely widely used.

● The rationale is that that the number of segregating sites and the average number of pairwise mismatches differ because:

● the former is indifferent to the relative frequencies of the polymorphic nucleotides at a given site

● The two values lead to consistent estimates for Θ anyway, *unless some evolutionary process causes a discrepancy from the assumptions of the model.*

TAJIMA D STATISTICS

● If the model assumptions hold, or any discrepancies are too small to invalidate equations (12) and (13), then $\boldsymbol{\pi} - S/a = 0$

● Consider the example in page 8.

● $S = 8$ and $\pi = 4$

● $n = 4$ so that $a = 1 + ½ + 1/3 = 1.833$

● The estimate of Θ from equation (12) is therefore 8/1.833 = 4.36 and from (13) 4.00.

● In this example $\pi - S/a = 4.00 – 4.36 = -0.36$

● As the sample size is very small, any formal statistical test is not reasonable (i.e. whether -0.36 is ”significantly different from 0”. The very small discrepancy from 0 suggests no significant excess of rare alleles.

● In practice, the ”statistical significance” by simulations: for each simulated sample $\pi - S/a$ is calculated and the null disribution of the test statistics, for a given case, is produced – assuming neutrality. If the observed value falls in the upper or lower 5% of the null distribution, then the P-value for the test is regarded as significant (P<0.05).

Tajima´s D statistics is based on the normalized version of $\pi - S/a$ where the maginutude of the difference is expressed as a multiple of the standard deviation fo the difference

$$D = \pi - S/a \ / \ [\sqrt{V}(\pi - S/a)] \qquad (14)$$

TAJIMA D STATISTICS

- Tajima D differs from 0 when:

  - The frequencies of polymorphic nucleotides are too nearly equal => The average number of pairwise differences over its neutral expectation => $\pi - S/a$ is positive.
    This might indicate either

    - some type of balancing selection, or

    - the sample is a mixture of two different populations which differ; this might result from a recent admixture of populations

  - The frequencies of the polymorphic variants are too unequal, with an excess frequency of the most common type and too many rare types.
    This pattern results in a decrease in the proportion of pairwise differences => $\pi - S/a$ is negative.

  - One possible reason for an exess of rare alleles is

    - selection against genotypes carrying deleterious alleles

    - However, departures from the infinite-sites model do not necessarily imply that natural selection is operating. For example, a population that is growing will also feature an excess rare alleles and a negative value of $\pi - S/a$

GENE GENEALOGIES AND THE COALESCENCE MODEL

- Current populations with polymorphisms are products of past events and population genetic analyses often model the branching of gene lineages to predict the time to the most recent common ancestor. Recap the concepts genetic drift and effective population size (see lecture slides “*Modelling mutations….”*)

- Tracing the pattern of ancestry for allele copies in a pedigree provides a means to understand the present patterns in those allele copies. Next page shows and example of a simple pedigree:

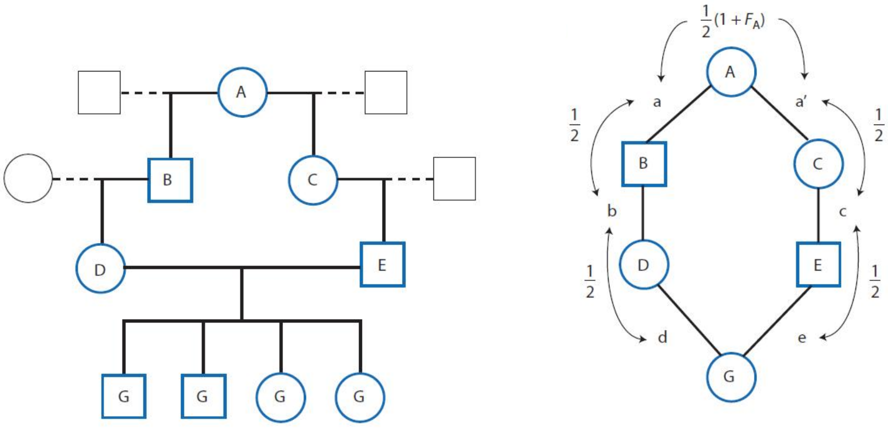
  - Equivalence of homozygosity in the present and the probability that two allele copies descended from a single ancestor in the past.

  - Given the known individuals at each generation in that pedigree, we traced ancestor–descendant relationships forward in time to predict autozygosity in the most recent generation.

  - Thus, that pedigree is an example of using a prospective or time-forward model, using knowledge of ancestors back in time and basic probability to work forward in time to predict the autozygosity at the most recent point in time.

This chapter in lecture slides is based on M Hamilton, Population genetics, 2009. Wiley-Blackwell (text here is almost a copy from this book).

GENE GENEALOGIES AND THE COALESCENCE MODEL



Average relatedness and autozygosity as the probability that two alleles at one locus are identical by descent.

(a) A pedigree where individual A has progeny that are half-siblings (B and C). B and C then produce progeny D and E, which in turn produce offspring G.

(b) Only the paths of relatedness where alleles could be inherited from A, with curved arrows to indicate the probability that gametes carry alleles identical by descent. Upper-case letters for individuals represent diploid genotypes and lower-case letters indicate allele copies within the gametes produced by the genotypes. The probability that A transmits a copy of the same allele to B and C depends on the degree of inbreeding for individual A, or F_A

GENE GENEALOGIES AND THE COALESCENCE MODEL

● Another type of analysis of ancestor–descendant relationships is possible based on a retrospective or time-backward model.

● Imagine that we have a sample of individuals taken in the present time, analogous to individual G in the pedigree (previous page), but there is no knowledge of their parents or grandparents or any of their genealogical relationships.

● Would it be possible to learn something about the past population genetic events that lead up to that sample of individuals?

● Yes, if we have models of ancestor–descendant relationships (genealogy) that allow us to predict identity by descent in the past based only on knowledge of the present. With such models, we look at patterns among the individuals available to us in the present and try to reconstruct versions of population genetics events (e.g.drift, selection) in the past that could have lead to the individuals in the present.

● These models are referred to collectively as coalescent theory since the perspective of the models is to predict the probability of possible patterns of genealogical branching working back in time from the present to the point of a single common ancestor in the past.

● When two lineages trace back in time to a single ancestral lineage it is said to be a coalescent event, hence the term coalescent theory

COALESCENCE THEORY

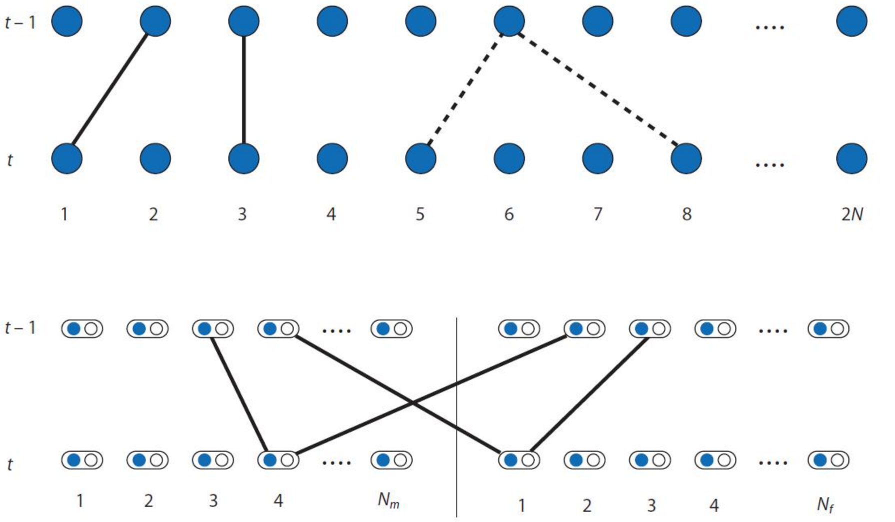
● A central concept in coalescent theory is connecting a group of lineages in the present back through time to a single ancestor in the past. This single ancestor is the first ancestor (going backward in time) of all the lineages in a sample of lineages in the present time and is referred to as the most recent common ancestor, MRCA.

● Recall genetic drift: a time-forward model that predicted that a sample of alleles (or lineages) eventually arrive to fixation or loss. Fixation is reached by random sampling that expands the numbers of a given lineage or allele in the population. The lineage that reaches fixation can be traced back to a single ancestor at some point in the past. In the process of reaching fixation, a population loses all lineages except one, the one that was fixed by genetic drift.

● This same genetic drift process can be viewed from a time-backward perspective. A sample of lineages in the present must eventually be the product of a single ancestral lineage at some point back in the past that happened to become more frequent under random sampling. The coalescent model turns the random sampling process around, asking: what is the probability that two lineages in the present can be traced back to a single lineage in the previous generation? Answering this question relies on the same probability tools that were used earlier to describe the process of genetic drift.

A metaphor: Imagine a sealed box full of bugs. Each bug moves around the box at random. Whenever two bugs meet by chance, one of them (picked at random) completely eats the other one in an instant. When a bug is eaten the population of bugs decreases by one and the remaining bugs continue to move about the box at random. The time that elapses between bug meetings tends to get longer as the number of bugs in the box gets smaller. This is because chance meetings between bugs depend on the density of bugs in the box. Eventually, the entire box that was full of bugs initially will wind up holding only a single bug after some time has passed. Each bug is analogous to a lineage and one bug eating another is analogous to a coalescent event. The very last bug is analogous to the lineage that is the most recent common ancestor.

COALESCENCE THEORY - ILLUSTRATION IN TERMS OF RANDOM SAMPLING CONCEPTS



Haploid and diploid reproduction in the context of coalescent events. In a haploid population, the probability of coalescence is 1/2N (dashed lines) whereas the probability that two lineages do not have a common ancestor in the previous generation is 1 – 1/(2N) (solid lines). In a diploid population, the two gene or allele copies in one individual in the present time have one ancestor in the female population (N_f) and one ancestor in the male population *(*N_m*).* Coalescent events in the diploid population arise when the gene copies in males and females are identical by descent. The haploid model with 2N *l*ineages is routinely used to approximate the diploid model with $N = N_f + N_m$ diploid individuals.

COALESCENCE THEORY

● So, using rules of random sampling based around the genetic drift model , a prediction can be developed for the number of generations back in time until two lineages “find” their MRCA or coalesce to a single lineage.

● Consider a random sample of two of the 2N total lineages in the present generation.
Given that one of these two sampled lineages finds its ancestor in the previous generation, what is the probability that the other lineage also shares that same common ancestor such that a coalescent event occurs?

● Given that one of the lineages has a given common ancestor, for coalescence to occur the other lineage must have the same ancestor among the 2N possible ancestors in the previous generation.

● Thus the probability of coalescence is 1/2N for two lineages whereas the probability that two lineages do not have a common ancestor in the previous generation is 1 – 1/(2N).

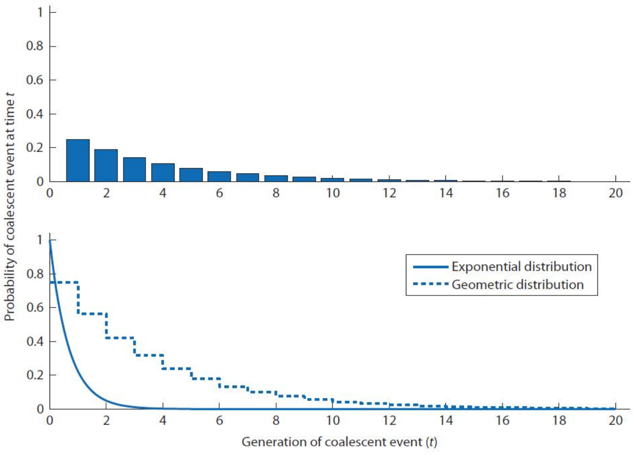
● In the diploid case, (each offspring composed of one allele copy inherited from a female parent and another allele copy from a male parent) a time-backward view: reproduction where one allele copy finds its ancestor in the male population of the last generation while the other allele copy finds its ancestor in the female population of the last generation. For a given male or female parent, each of their two allele copies has a probability of 1/2 of being the ancestral copy. As long as the number of males and females in a diploid population is equal and the haploid and diploid population sizes are large, the predictions of the coalescent model are very similar for haploid and diploid populations containing an identical total number of gene copies. The haploid model is more straightforward and so it is used in what follows.

COALESCENCE THEORY

● Like Markov chains, the probability of coalescence displays the Markov property since it is an independent event that depends only on the state of the population at the point of time of interest. Because of this, the basic probabilities of coalescence and non-coalescence between two generations can be used to describe the probability of coalescence over an arbitrary number of generations. If two randomly sampled lineages do not coalesce for $t – 1$ generations, then the probability that they do coalesce to their common ancestor in generation t is

$$[1 – 1/(2N)]^{t\text{-}1} \ [1/(2N)] \qquad (15)$$

● Example: In a population of 2N=10 the chance that two randomly sampled lineages coalesce in four generations is the product of the probability of three generations not coalescing $(1 – 1/10)^3$ = 0.729 and the chance of coalescing between any two generations (1/10), which gives a probability of coalescence 0f 0.0729. The distribution of probabilities of a coalescent event occurring for two lineages in each of 30 generations for the case of 2N=10 is show here.
The bottom figure shows the approximations of These coalescence probabilities based on Geometric and exponential distributions with a probability of ”success” of ¼.



COALESCENCE THEORY

● In practice, the probabilities of coalescence are approximated using an exponential function. To recap: The exact probability of coalescence for a pair of lineages is $1/2N$ and the probability of not coalescing is $1 - 1/(2N)$ in each generation.

● The exponential approximation $1 - e^{-\frac{1}{2N}t}$ gives the cumulative probability of a pair of lineages coalescing at or before generation t. This probability *is* symbolized as $P(T_C \le t)$ where

T_C is the generation of coalescence and
t is the maximum time to coalescence being considered.

● Example: The probability of coalescence at or before four generations have passed in a population of $2N = 10{,}000$.
The exact probability is the sum of the probabilities of coalescence in each generation, $P(T_C \le 4) = P(T_C = 1) + P(T_C = 2) + P(T_C = 3) + P(T_C = 4)$.
Substituting in expressions for the exact probability of coalescence at each of these four time points gives $P(TC \le 4) = (1 - 1/10000)^0\ 1/10000 + (1 - 1/10000)^1\ 1/10000 + (1 - 1/10000)^2\ 1/10000 + (1 - 1/10000)^3\ 1/10000 = 0.0004$

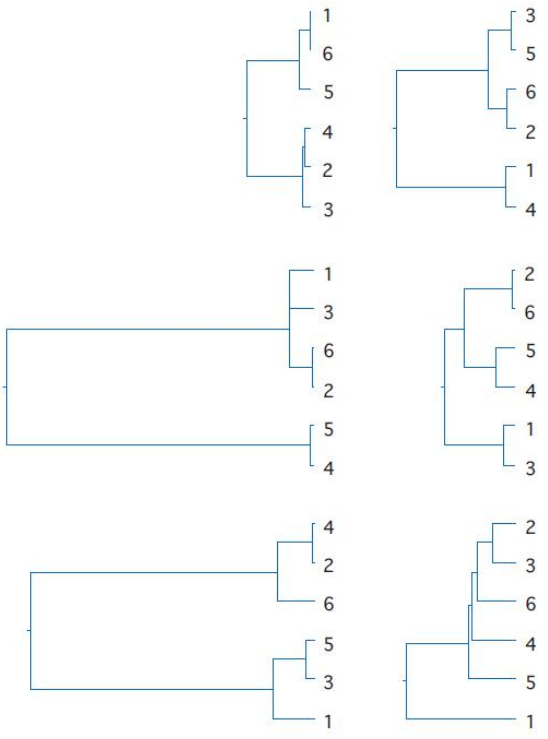
● Using the exponential approximation, gives 0.00039992 as the chance that a pair of lineages experiences a coalescence at or before 4 generations elapses. Quite good agreement with exact and prob and approximation. Approximating probabilities of coalescence with the exponential distribution makes computing more practical and also yields several generalizations about the coalescence process.

COALESCENCE THEORY – WAITING TIME

● As the probability of coalescence for a pair of lineages is $1/(2N)$, then the average time that elapses until coalescence is $2N$ when approximated by the exponential distribution.

● The average time to a coalescence is called the waiting time.

● The range of individual coalescence times around that average is quite large. Based on the exponential distribution, the variance in the waiting time is $4N^2$ so that range of coalescence times around the mean grows rapidly as the size of the population increases. Thus, the length of branches connecting lineages to their ancestors will be highly variable about their mean value, like in the example which shows six independent realizations of the coalescent tree for six lineages.



● It is possible to determine the average time for more than two lineages to find their MRCA. Suppose we want to determine the waiting time for *k* lineages where *k* is less than or equal to the total number of lineages sampled from a population of 2*N.*

● Let´s consider the case of *k = 3* lineages. When no coalescence events occur, one lineage finds its ancestor among any of the 2*N* individuals in the previous generation*.* That means the next lineage must find its ancestor among 2*N* – 1 individuals in the previous generation and the final lineage must find its ancestor among 2*N* – *2* possible parents. Thus the probability of non-coalescence is

$$\prod_{x=0}^{k-1}\left(1-\frac{x}{2N}\right) \qquad (16)$$

● If the number of lineages sampled is much smaller than the total number of lineages in the population (2*N*) then the probability of non-coalescence for *k* lineages can be approximated by

$$1-\left(\frac{k(k-1)}{2}\right)\left(\frac{1}{2N}\right) \qquad (17)$$

where *k*(*k*-1)/2 enumerates the different ways to uniquely sample pairs of lineages from a total of *k* lineages.

● The probability of a coalescence for any one of the unique pairs of the *k* lineage is then

$$\left(\frac{k(k-1)}{2}\right)\left(\frac{1}{2N}\right) \qquad (18)$$

● Now we bring these probabilities together like in equation (35) and obtain the probability that *k* lineages experience a single coalescent event *t* generations ago

$$\left(1-\left(\frac{k(k-1)}{2}\right)\left(\frac{1}{2N}\right)\right)^{t-1}\left(\frac{k(k-1)}{2}\right)\left(\frac{1}{2N}\right) \qquad (19)$$

● Since this probability also follows an exponential distribution, the average time to coalescence for *k* lineages in a population of 2*N* is

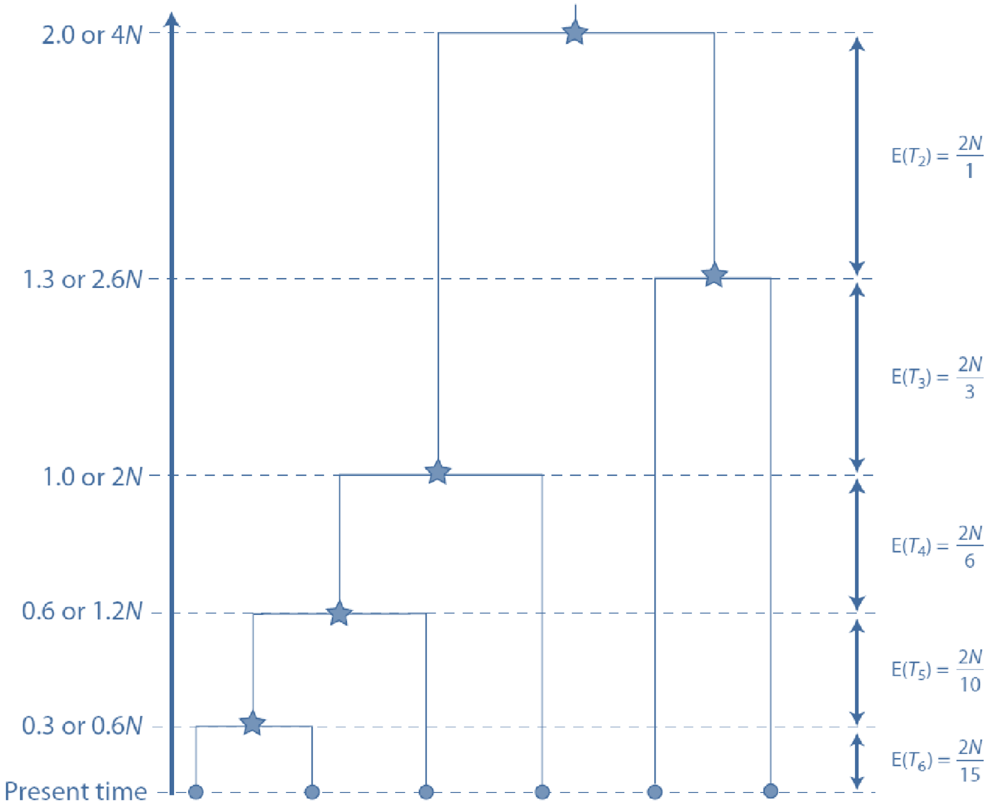
$$\frac{2N}{\frac{k(k-1)}{2}} \qquad (20)$$

● For example, if *k*=3 and 2*N*=10, the average time to coalescence is 3 1/3 generations
This is one third of the average waiting time for two lineages since each of the three unique pairs of lineages (1–2, 1–3, and 2–3) can independently experience coalescence.

● Figure next page shows the average coalescence times for six lineages based on this same logic. The general pattern is that coalescence times decrease when more lineages are present since there are a larger number of lineage pairs that can independently coalesce.

● In figure (next page) *E* refers to expected and *T* refers to time to coalescence so that *E*(T_n) is the expected or average time to coalescence for *n* lineages. The basic patterns seen in all coalescent trees apply to populations of all sizes, although the absolute time for coalescent events does depend on *N.* Values on the left are one realization of coalescent waiting times.
Genealogy is not drawn to scale.

COALESCENCE THEORY - THE AVERAGE TIME TO FIND MRCA



COALESCENCE THEORY

● When approximating the probabilities of coalescent events with the exponential distribution, it is standard practice to put coalescence times on a continuous scale of units of 2*N* generations. To see how this continuous time scale operates, let *j* be time measured as a real number (e.g. 1.0, 1.1, 1.2, 1.3 . . . *j)* in generations.

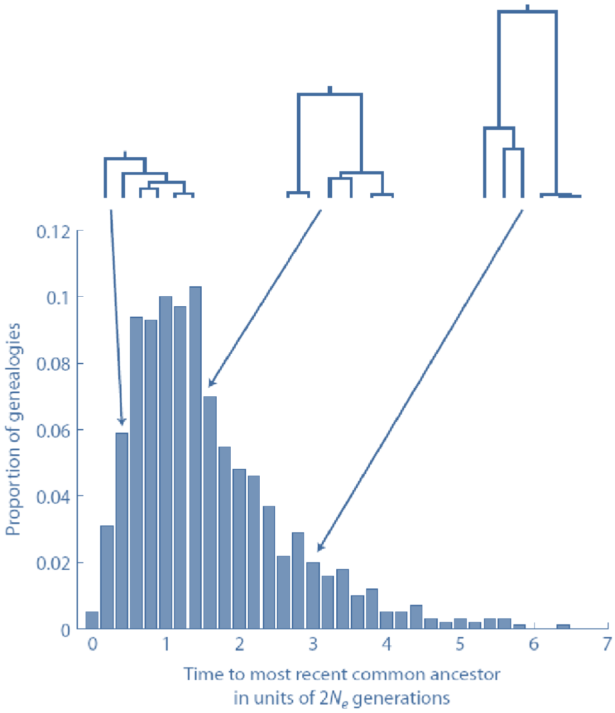
● The time to coalescent events *t can* then be expressed as *t = j/(2N).* As an example, imagine that a coalescence event occurred at *t = 1.4* on the continuous time scale. That coalescence event could also be thought of as occurring (1.4)(2*N)* = 2.8*N* generations in the past *(*see the previous figure). If the population size was 2*N* = 100 lineages, then that coalescent event was (1.4)(100) = 140 generations in the past. However, if the population size was 2*N* = *20* lineages, then that coalescent event was (1.4)(20) = 28 generations in the past.

● Population size serves to scale the time required for coalescent events to occur. Coalescent events occur more rapidly in small populations compared to bigger populations, a conclusion analogous to that of genetic drift effects.

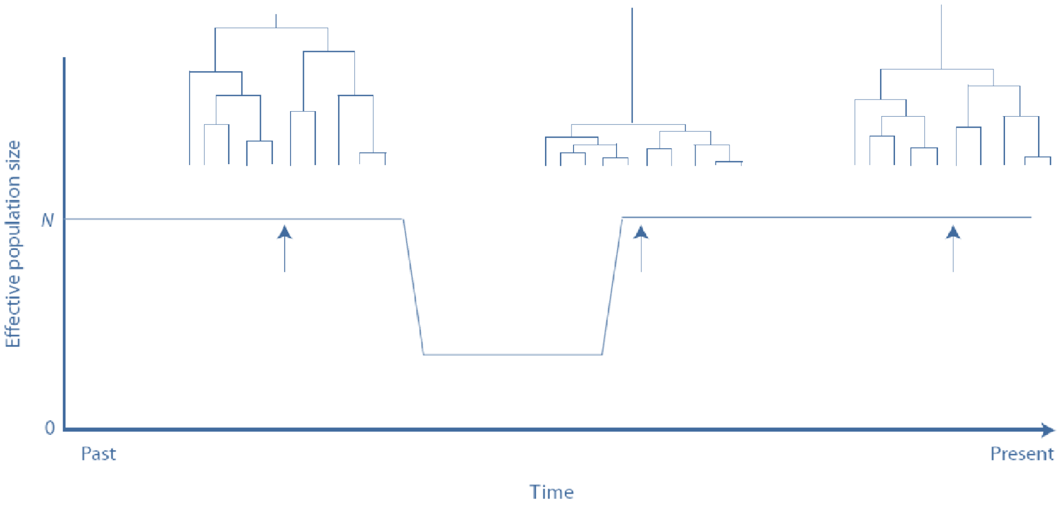
The height of a coalescence tree

● The height of a tree for *k* sampled lineages is the sum of the coalescence waiting times as coalescent events reduce the number of lineages from *k* to *k*-1 to *k*-2 down to one. The formulae are not presented fron now on. We turn to pick up some examples in order to reach the practical value of this theory.

● The figure illustrates the variance in the total height of genealogies by displaying the time to MRCA for 1000 replicate genealogies each starting with *k* = *6.* The range of time to MRCA is large and the distribution has a very long tail representing a small proportion of genealogies that take a very long time for all coalescence events to occur.

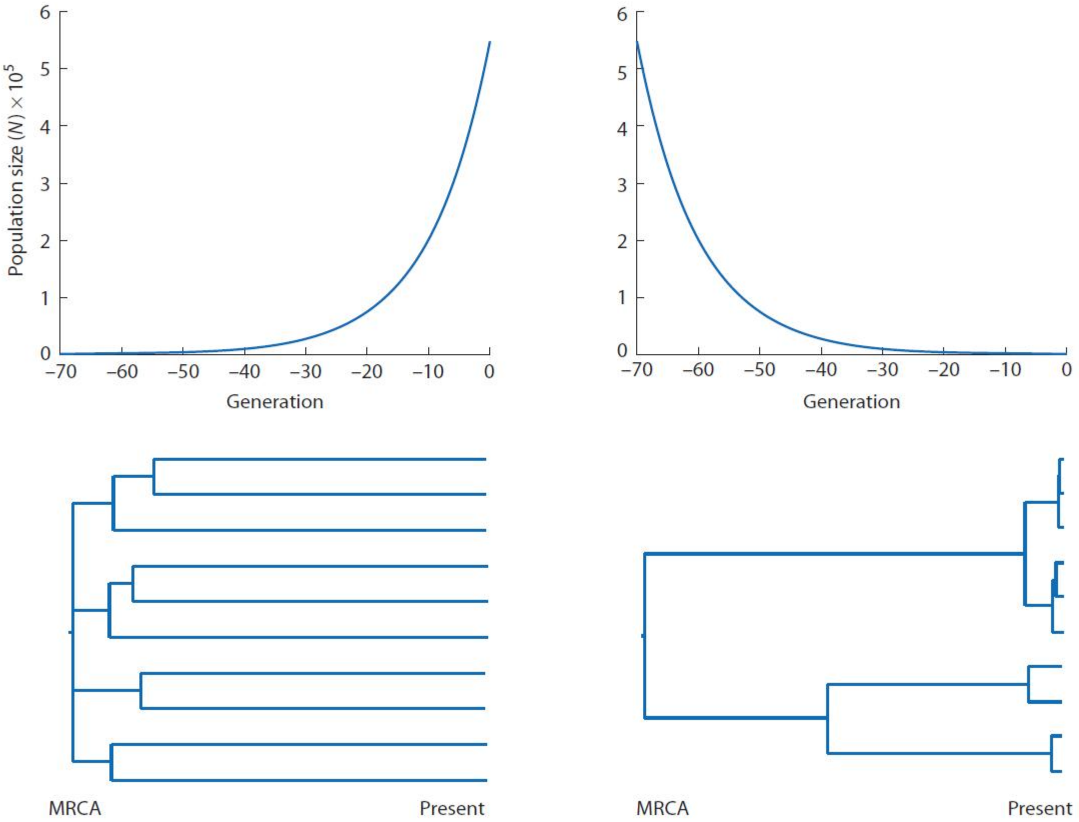


COALESCENCE THEORY - POPULATION BOTTLENECK



● During the bottleneck the chance that two randomly sampled gene copies are derived from one copy in the previous generation (1/2*N*) increases. This can also be thought of as a reduction in the overall height of a genealogical tree caused by the bottleneck since lineages that find their ancestors during the bottleneck lead to short branches. The overall effect of a bottleneck on coalescence among gene copies sampled in the present depends on the reduction in the effective population size and the duration. The arrows indicate the point in time when gene copies were sampled from the population.

COALESCENCE THEORY - POPULATION SIZE CHANGES



COALESCENCE THEORY

- Figure in previous page: The two genealogies illustrate examples of waiting times that might be seen under strong exponential population growth (left) and shrinkage (right). With strong exponential population growth coalescent times are longest in the present when the population is the largest, leading to genealogies characterized by long branches near the present and very short branches in the past around the time of the MCRA. With exponential population shrinkage, coalescence times are greatest in the past near the MRCA when the population was larger and shortest near the present when the population is at its smallest size.

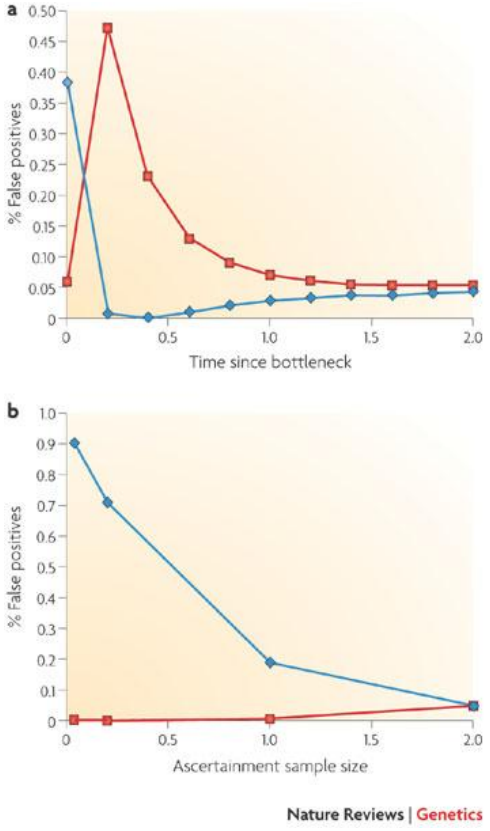
.

COALESCENCE THEORY - TAJIMA D

Effects of demography and ascertainment bias on tests of selection.

a)The false positive rate of Tajima's D in the presence of a population bottleneck.
A sample of 50 chromosomes was simulated using coalescent simulations, and the time since a population bottleneck that reduced the population size tenfold was varied. The duration of the bottleneck was 0.1 $2N_e$ generations, and time in the figure is measured in $2N_e$ generations, where N_e is the effective population size of a diploid population. Each simulated data set had 20 segregating sites. The proportion of time the test rejects at the 5% significance level in a one-sided test based on negative values (shown in red) and positive values (shown in blue) is shown.

b) The false-positive rate of Tajima's D in the presence of an ascertainment bias. Simulation conditions are as described in panel a, but the size of the ascertainment sample (expressed as a proportion of the final sample) used for SNP discovery is varied.



From: Nielsen *et al*. 2007. Recent and ongoing selection in the human genome. Nature Reviews Genetics 8: 857-868

COALESCENCE THEORY – TAJIMA D

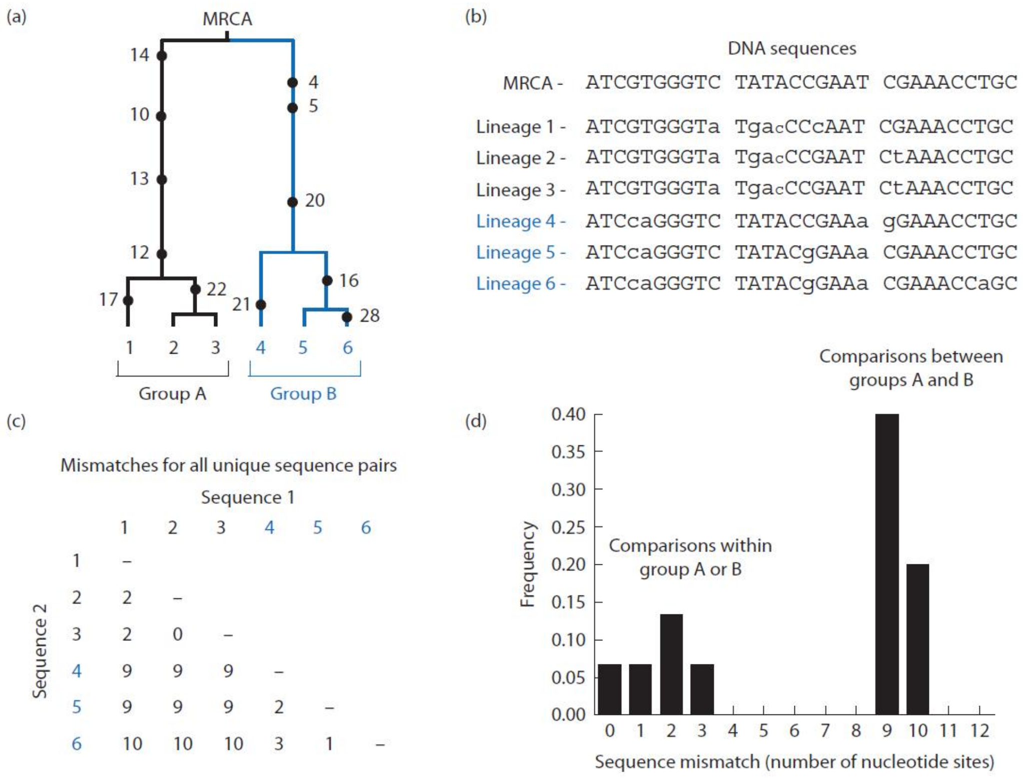
● Differences in the shape of genealogies are the basis of Tajima's *D* test. In the standard coalescent model of genealogical branching the probability of coalescence is constant per lineage over time.

● The standard coalescent therefore gives expected branch lengths when all alleles are selectively neutral and the effective population size is constant (center). Changes in the effective population size over time (population growth, population bottlenecks) change the probability of coalescence over time as well. Natural selection also alters the probability of coalescence based on the fitness of alleles each lineage bears.

● Changes in the effective population size and natural selection alter the expected time to coalescence and therefore the expected branch lengths in a genealogical tree. If the chance of coalescence is greater in the present than in the past (right), most coalescent events occur near the present and internal branches are long in comparison with external branches. If the chance of coalescence is smaller in the present than in the past (left), most coalescent events occurred in the past and external branches are long in comparison with internal branches.

● Since the chance of a mutation is constant over time, lineages with longer branches are expected to experience more mutations.

COALESCENCE THEORY - MISMATCH DISTRIBUTIONS



DNA sequences

MRCA -	ATCGTGGGTC TATACCGAAT CGAAACCTGC
Lineage 1 -	ATCGTGGGTa TgacCCcAAT CGAAACCTGC
Lineage 2 -	ATCGTGGGTa TgacCCGAAT CtAAACCTGC
Lineage 3 -	ATCGTGGGTa TgacCCGAAT CtAAACCTGC
Lineage 4 -	ATCcaGGGTC TATACCGAAa gGAAACCTGC
Lineage 5 -	ATCcaGGGTC TATACgGAAa CGAAACCTGC
Lineage 6 -	ATCcaGGGTC TATACgGAAa CGAAACCaGC

Mismatches for all unique sequence pairs

Sequence 1 / Sequence 2

	1	2	3	4	5	6
1	–					
2	2	–				
3	2	0	–			
4	9	9	9	–		
5	9	9	9	2	–	
6	10	10	10	3	1	–

COALESCENCE THEORY – MISMATCH DISTRIBUTIONS

- Figure in previous page: The basis of the mismatch distribution by using coalescence theory

(a) A neutral genealogy that bears multiple mutation events. Each mutation event is represented by a circle and the number of the random nucleotide site that mutated assuming the infinite sites mutation model. The six lineages in the present can be separated into two groups (called A and B) based on their ancestral lineage when there were only two lineages in the population.

(b) The DNA sequences for each lineage are shown based on the 30 base-pair sequence assigned to the most recent common ancestor (MRCA) with mutations shown in lower-case letters.

(c) The number of nucleotide sites that are different or mismatched between pairs of DNA sequences.

(d) The mismatch distribution shown is a histogram of the mismatches for the 15 pairs of DNA sequences compared. Neutral genealogies from populations with constant N_e through time tend to show bimodal mismatch distributions. The cluster of observations with few mismatches results from sequence comparisons between recently related lineages (comparisons within group A or group B). In contrast, sequences from distantly related lineages that do not share the same ancestor when $k = 2$ (comparisons between groups A and B) tend to have more mismatches. .

This high level of mismatch occurs because sequences from distantly related lineages are separated by much more time since they shared a common ancestral lineage, leading to many more mutational changes that independently altered each DNA sequence.

COALESCENCE THEORY – MISMATCH DISTRIBUTIONS

● Another way to think of the situation is that closely related lineages differ only by a few young mutations while distantly related lineages differ by more mutations, many of which are old and have been resident in the population for a long time.

● The mismatch distribution has distinct patterns depending on the demographic history of the population. Mismatch distributions from populations that have experienced a constant N_e over time tend to have two clusters of values in the mismatch distribution.

.

● Such a bimodal distribution is the characteristic signature of genealogies in populations with a relatively constant N_e in the past. The bimodal pattern is caused by roughly equal times to coalescence of all internal and external branches.

● In contrast, populations that had rapidly growing or shrinking N_e in the past tend to have distinct mismatch distributions. In populations that have rapidly growing N_e, most coalescence events happen early in the genealogy near the MRCA since the probability of coalescence decreases toward the present This leads to long external branches that each experience many unique mutations. The mismatch distribution then has a high frequency of sequence pairs with a high degree of mismatch and few sequence pairs with a low degree of mismatch. Alternatively, populations that experienced continual declines in N_e have genealogies where most coalescence events happen near the present because the probability of coalescence increases toward the present. In a shrinking population, the mismatch distribution tends to have a high frequency of sequence pairs with low mismatch counts

COALESCENCE THEORY – SELECTIVE SWEEP

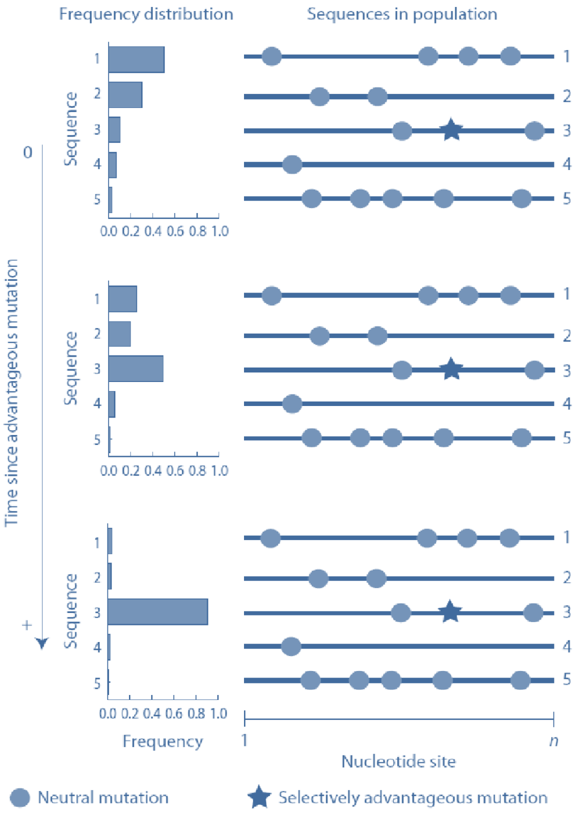
Selective sweep, the impact of natural selection on an advantageous mutation as well as on associated nucleotide sites.

● Imagine a single population that contains five distinct DNA sequences (without recombination).

● Each DNA sequence is distinguished by a number of neutral mutations and has a frequency given by the histogram on the left. Initially, the population has polymorphism since the population is composed of intermediate frequencies of each DNA sequence.

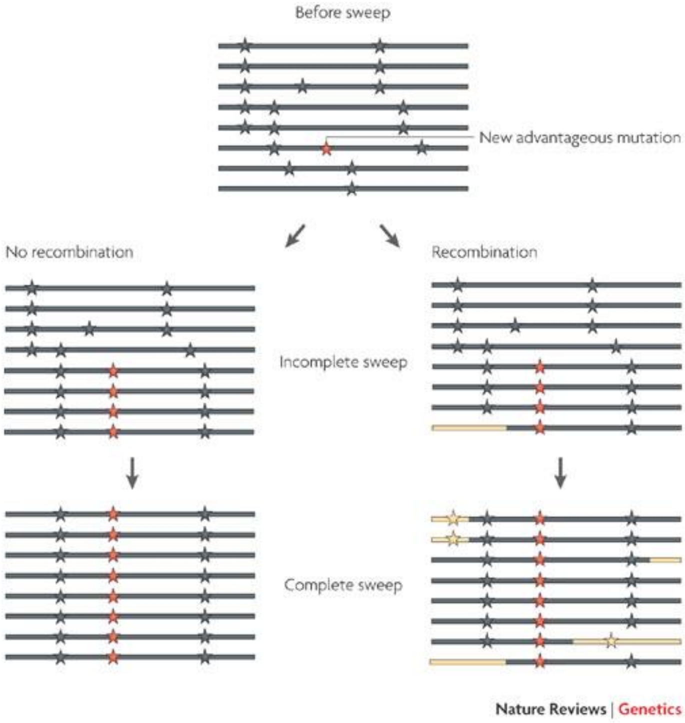
● At time 0, the third DNA sequence experiences a mutation that is strongly advantageous, indicated by the star. Natural selection acts to increase the frequency of the advantageous mutation over time, until the population approaches fixation for the third DNA sequence.

● Once selection has swept the advantageous mutation to near fixation the population has very little polymorphism. This is because only those original neutral mutations that were linked to the advantageous allele on the same DNA sequence remain in the population. Thus, positive selection on one site also sweeps away polymorphism at linked nucleotide sites if gametic disequilibrium is maintained. The figure assumes that positive natural selection is strong and increases the frequency of the third DNA sequence rapidly such that no new mutations appear in the population.



SELECTIVE SWEEPS

- ● The lines indicate individual DNA sequences or haplotypes, and derived SNP alleles are depicted as stars.
- ● A new advantageous mutation (indicated by a red star) appears initially on one haplotype. In the absence of recombination, all neutral SNP alleles on the chromosome in which the advantageous mutation first occurs will also reach a frequency of 100% as the advantageous mutation become fixed in the population.
- ● Likewise, SNP-alleles that do not occur on this chromosome will be lost, so that all variability has been eliminated in the region in which the selective sweep occurred.
- ● However, new haplotypes can emerge through recombination, allowing some of the neutral mutations that are linked to the advantageous mutation to segregate after a completed selective sweep.
- ● As the rate of recombination depends on the physical distance among sites, the effect of a selective sweep on variation in the genomic regions around it diminishes with distance from the site that is under selection.
- ● Chromosomal segments that are linked to advantageous mutations through recombination during the selective sweep are coloured yellow.
- ● Data that are sampled during the selective sweep at a time point when the new mutation has not yet reached a frequency of 100% represent an incomplete selective sweep.



From: Nielsen *et al*. 2007. Recent and ongoing selection in the human genome. Nature Reviews Genetics 8: 857-868

SELECTIVE SWEEPS

● As regards the human genome, “hunting” selective sweeps is very popular. Much interest has focused on identifying incomplete selective sweeps, which are seen when positively selected mutations are currently on the rise in the human populations but have not yet reached a frequency of 100%. The pattern that is left by such mutations is distinctive, involving some locally identical haplotypes that segregate at moderate or high frequencies, whereas the remaining haplotypes show normal levels of variability.

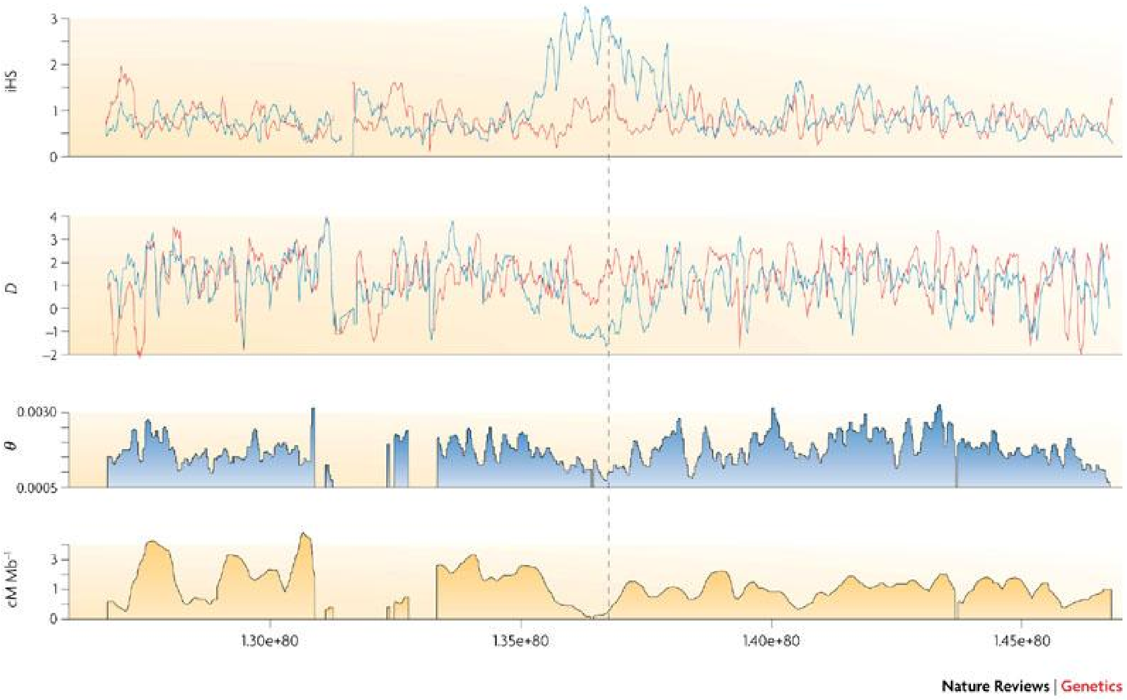
● One of the most famous examples of an incomplete sweep is that at the lactase (LCT) locus in European populations. Variants in this gene influence whether the ability to produce lactase, which enables the digestion of milk, persists into adulthood. Lactase persistence is thought to have increased in frequency as a result of positive selection during the past 10,000 years after the emergence of dairy farming.

● Figure next page: The LCT region shows a characteristic signature of an incomplete selective sweep. There is a haplotype of high frequency with strongly increased homozygosity as illustrated by the iHS (integrated haplotype score) statistic.

● There is a skew in the frequency spectrum as illustrated by the negative values of Tajima's D, and a reduction in variability as shown by the estimate of the population genetic parameter. Characteristically of many regions that show statistical evidence for an incomplete selective sweep, there is also a reduction in the local recombination rate (cM Mb^{-1}). For the top two panels, the red lines represent the Asian and the blue lines represents the CEPH HapMap samples.

From: Nielsen *et al*. 2007. Recent and ongoing selection in the human genome. Nature Reviews Genetics 8: 857-868

LACTASE REGION, INCOMPLETE SELECTIVE SWEEP



iHS
3
2
1
0
D
4
3
2
1
0
−1
−2
θ
0.0030
0.0005
cM Mb^{-1}
3
1
0
1.30e+80
1.35e+80
1.40e+80
1.45e+80
Nature Reviews | Genetics