Contents

2 Mutation

2.1 Introduction

Mutation

HWE Assumption. There is no mutation.

Definition: *mutation*

> A ***mutation*** is a heritable change in the genome, i.e. a change occuring in the reproductive cells of an organism.

Mutation provides the raw material for evolution. We wouldn't have alleles A and a if it wasn't for mutation.

All mutations are ultimately changes at the nucleotide level. The vast majority of mutations are deleterious. These mutations are present in populations because they arise by accident during genome copying during meiosis.

The human mutation rate per site per generation is 10^{-9} to 10^{-10}. If we treat sites as independent, then a gene locus of, say, 1000 nucleotides will have a mutation rate per gene per replication cycle of 10^{-6} to 10^{-7}.

Overview of mutation: `http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hmg.chapter.1049`

Types of Mutations

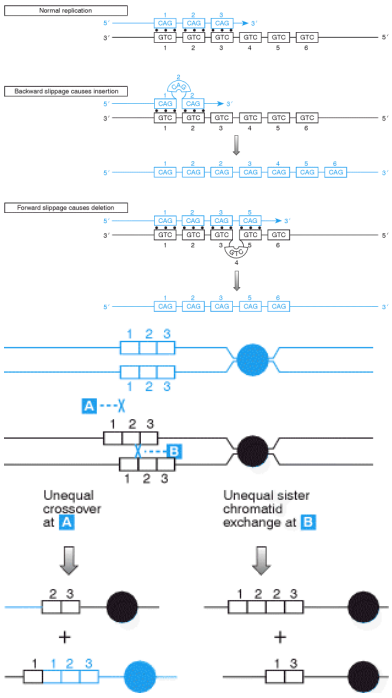
- large-scale chromosome abnormalities: caused by misrepair of broken chromosomes, improper recombination, or malsegregation
- simple mutations: local mutations, e.g. during a lifetime each gene will receive about $10^8 - 10^{10}$ mutations
  - ***base substitution***: replacement of a single base
  - ***deletion***: one or more nucleotides removed from a sequence
  - ***insertion***: one or more nucleotides added into a sequence, sometimes from another locus, in which case there are two types:
    - ***copy transposition***. sequence from one locus copied to another locus
    - ***non-copy transposition***. sequence is transposed from one locus to another, no copy left
- mutations involving exchanges: in particular, tandemly repetitive DNA is prone to insertion/deletion polymorphism

Variable Number of Repeats (VNTR) Loci

Definition: *VNTR polymorphism*

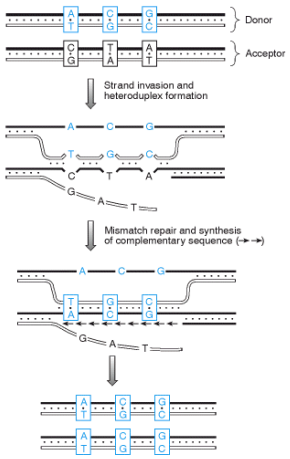
> A ***VNTR polymorphism*** is allelic variation caused by different number of repeats of very short (***microsatellite***), intermediate length (***minisatellite***), or long sequences at a locus.

There are many mechanisms that lead to polymorphism at such sites.



Microsatellites appear to experience strand-slippage during replication leading to loss or gain of repeats.

Unequal crossover results from recombination between non-allelic sequences and these events are facilitated by pairing of nonallelic repeats.



Gene conversion is a process by which a donor sequence donates and replaces the contents of an acceptor sequence.

2.2 Theory

Modeling Mutation

Consider a locus where there are 2 alleles possible A and a.

Suppose the mutation rate (per replication cycle per locus) for mutating $A \to a$ is u. Let v be the rate from $a \to A$.

Let $p_A(t)$ be the frequency of allele A in the tth generation.

In the next generation, type A alleles will arise by faithful copy of type A alleles from the previous generation, or by mutation during copying of type a alleles from the previous generation. So,

$$\begin{aligned} p_A(t+1) &= (1-u)p_A(t) + v\,[1-p_A(t)] \\ \Delta p_A(t+1) &= p_A(t+1) - p_A(t) \\ &= -up_A(t) + v\,[1-p_A(t)]\,. \end{aligned}$$

A Stable Mutation Equilibrium

A mutation equilibrium occurs when $\Delta p_A(t) = 0$ and $p_A(t) = p_A$. Solving the equation for $\Delta p_A(t) = 0$

$$0 = -up_A + v\,[1-p_A]$$

yields

$$p_A = \frac{v}{u+v}$$

at equilibrium.

Exercise. The equilibrium is stable. To verify this, plug in $p_A = \frac{v}{u+v} - \delta$ in the equation for Δp_A. Will p_A increase or decrease in the next generation? Repeat with $p_A = \frac{v}{u+v} + \delta$.

Exercise. If A is the vastly dominant allele, show $\Delta p_A(t) \approx -u$. How much relative error is introduced in making this approximation? Relative error is the difference between the exact and approximate values divided by the exact value.

Rate of Approach to Equilibrium

Take the recurrence relation for $p_A(t)$ and subtract the equilibrium p_A

$$\begin{aligned} p_A(t+1) - p_A &= (1-u)p_A(t) + v\,[1-p_A(t)] - p_A \\ &= (1-u)p_A(t) + v\,[1-p_A(t)] - (1-u)p_A - v\,[1-p_A] \\ &= (1-u)\,[p_A(t) - p_A] + v\,[1-p_A(t) - 1 + p_A] \\ &= [1-u-v]\,[p_A(t) - p_A]\,, \end{aligned}$$

and we're overly familiar with this kind of equation

$$p_A(t) - p_A = (1 - u - v)^t (p_{A0} - p_A),$$

where p_{A0} is the initial frequency of type A alleles. **The approach to equilibrium is very slow since $1 - u - v \approx 1$.**

Exercise. By Taylor's series, we know $(1 - u - v)^t \approx e^{-(u+v)t}$. Use this approximation to compute the landmark times $t_{1/2}$, the time it takes to decrease the starting disequilibrium $p_{A0} - p_A$ by one-half.

Neglecting *Back Mutation*

You will commonly hear someone say or write, "and we neglected back mutation."

If A is the normal (wild type) form and a is the mutant form of the allele, then neglecting back mutation is equivalent to setting $v = 0$. Back mutation mutates the mutant form back to the wild type form of the allele.

It is biologically reasonable to neglect back mutation because often we are speaking of an allele variant of a protein. Either the protein works (normal/wild type) or it doesn't (mutant). There are many more ways to make a protein that doesn't work than one that does, so generally $u \gg v$.

However, when considering DNA sequences it is *not* reasonable to neglect back mutation. If $A \to C$ with probability u, then it is normally not all right to assume $C \to A$ is virtually impossible (i.e. $v = 0$).

Mutation with Multiple Alleles

For DNA sequences it is often the case that you are dealing with a large number n of possible alleles $A_1, A_2, \ldots, A_n$.

Let u_{ij} be the mutation rate from allele i to allele j. Then, the recursion equation for type i alleles is

$$p_i(t+1) = p_i(t) \left[1 - \sum_{j \neq i} u_{ij} \right] + \sum_{j \neq i} p_j(t) u_{ji}$$

Equations can be established for the equilibrium allele frequencies p_i, by setting $p_i(t+1) = p_i(t) = p_i$ in the equations. For given u_{ij}, the resulting linear system of equations that can be solved for p_i.

Equilibrium with Multiple Alleles

For the special case that $u_{ij} = u$ for all $i \neq j$, then

$$\begin{aligned} p_i &= p_i [1 - (n-1)u] + \sum_{j \neq i} u p_j \\ (n-1)p_i &= \sum_{j \neq i} p_j \\ np_i &= 1 \\ p_i &= \frac{1}{n}. \end{aligned}$$

So, when all mutations are equally likely then all alleles are equally prevalent at equilibrium.

This model of evolution at the DNA level is called the Jukes-Cantor model of nucleotide substitution. It implies that all nucleotides A, C, G, and T are equally likely at every position in the alignment, when mutational equilibrium has been achieved.

Mutation and Linkage Disequilibrium

Let the gamete frequencies for two loci be $p_{AB}, p_{Ab}, p_{aB}, p_{ab}$, where locus 1 has alleles A and a and locus 2 has alleles B and b.

Suppose the mutation rate $A \rightarrow a$ is u_1 and $a \rightarrow A$ is v_1 for locus 1. Similarly define u_2 and v_2 for locus 2.

Follow p_{AB} over time and assume linkage equilibrium at generation t

$$\begin{aligned} p_{AB}(t+1) &= (1-u_1)(1-u_2)p_{AB}(t) + (1-u_1)v_2 p_{Ab}(t) \\ &\quad + v_1(1-u_2)p_{aB}(t) + v_1 v_2 p_{ab}(t) \\ &= (1-u_1)(1-u_2)p_A(t)p_B(t) + (1-u_1)v_2 p_A(t)p_b(t) \\ &\quad + v_1(1-u_2)p_a(t)p_B(t) + v_1 v_2 p_a(t)p_b(t) \\ &= [(1-u_1)p_A(t) + v_1 p_b(t)]\,[(1-u_2)p_B(t) + v_2 p_b(t)] \\ &= p_A(t+1)p_B(t+1) \end{aligned}$$

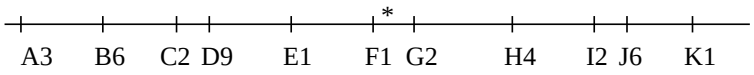
which is linkage equilibrium.

So, by the next generation, the two loci are still in linkage equilibrium.

We conclude that *mutation cannot create linkage disequilibrium.*

Unless....

If mutation is *so* rare that it becomes a random force, *then* mutation creates temporary linkage disequilibrium. Imagine a mutation that occurs on average once every million years. When it is first introduced, it will be introduced on a particular chromosome background.



Estimation of Mutation Rate

We'll talk about estimation of mutation rate ...

- ...when we discuss molecular evolution
- ...after we discuss finite population size

3 Genetic Drift

3.1 Introduction

Finite Population Size

HWE Assumption: population size is infinite

In finite populations, random changes in allele frequency result because of:

- variation in the number of offspring per parent.
- law of segregation in diploid species.

The random changes in allele frequency are called ***genetic drift***. Genetic drift is another force acting on genes in populations and it has two main consequences:

- it removes genetic variation at a rate inversely proportional to the population size.
- it affects the probability of survival of new mutations, in a manner approximately independent of the population size.

Neutral Theory

The concept that there exists a balance between genetic drift and mutation, with genetic drift removing variation and mutation restoring variation is fundamental in population genetics. The idea that much of the genetic variation present in populations is the consequence of the drift/mutation balance is called the ***Neutral Theory***.

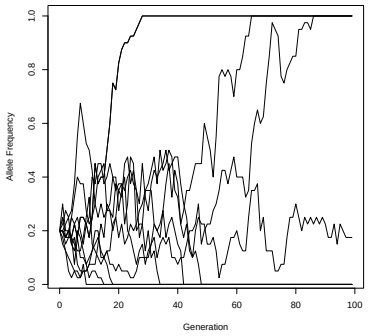
The theory has remained controversial since its inception because

- it is difficult to test
- it seems to negate the importance of natural selection, the core of Darwin's theory of evolution.

Simulating Populations

For a diploid population of size $N = 20$ and a starting allele A frequency of $p(0) = 0.2$, follow these steps to simulate generation $t + 1$,

- Set $K = 0$.
- Choose to copy an A allele with probability $p(t)$ and increment K, otherwise copy an a allele.
- Repeat $2 \times N$ times.
- Compute the generation $t + 1$ allele frequency $p(t + 1) = K/(2N)$.
- Repeat for 100 generations.

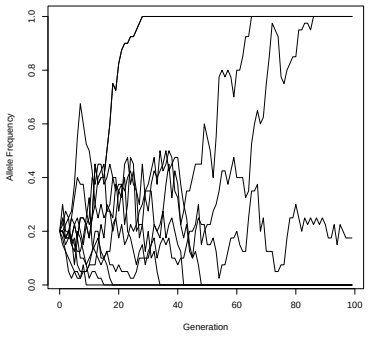


The result for 10 independent runs is shown in the plot.

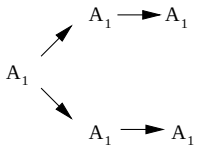
Observations About Simulations

There are a few things to observe about these allele frequency fluctuations

- There are random changes in allele frequency.
- All 10 populations display different allele frequencies over time, thus evolution can never be repeated.
- Alleles are lost from the population, 3 times a was lost, 6 times A was lost. Only one time did both alleles persist for 100 generations.
- The direction of random changes is neutral, i.e. not preferentially up or down.

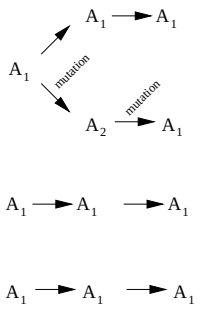


Identical by State/Descent



Definition: *Identical by descent (IBD)*

Two alleles are ***identical by descent (IBD)*** if they come from a common ancestor.



Definition: *Identical by state (IBS)*

> Two alleles are ***identical by state (IBS)*** if they do not share a common ancestor (Note: mutants are considered novel alleles and are not IBD with any other allele.)

Reminder: Inbreeding as a Measure of Nonrandom Mating

Recall that the model

$$\begin{aligned} P_{ii} &= p_i^2 + f p_i(1-p_i) \\ P_{ij} &= (1-f)2p_i p_j \quad i \neq j \end{aligned}$$

appeared as a general formulation of the nonrandom mating model, where the *inbreeding coefficient* $f \in [-1, 1]$ measured the degree of nonrandomness in mating.

An individual is inbred if she/he contains alleles at a locus that are IBD (identical by descent). An individual becomes inbred because his/her ancestors are related. For example, the homozygous descendents of selfers have IBD alleles. The only way to *not* be inbred is to have completely unrelated ancestors (or mutate a great deal).

Nonrandom Mating and Finite Population Sizes

But, if there were no relationships among your ancestors, then if you are generation t, you have 2 distinct ancestors in generation $t-1$ (your parents), $4 = 2^2$ distinct parents in generation $t-2$ (your grandparents), $8 = 2^3$ distinct parents in generation $t-3$ and so on. It doesn't take long before we're talking about terabytes of ancestors, more ancestors than the size of the population N.

Therefore, finite populations lead to inbred individuals and everyone is at least a little inbred *even if random mating is enforced*.

Inbred Populations

A population is inbred if the probability that two alleles selected without replacement are IBD is positive. As a population becomes inbred, so too the individuals.

In randomly segregating, randomly mating populations, the level of population inbreeding is equal to the level of individual inbreeding, i.e. the probability that two random alleles are IBD is the same as the probability that the two alleles in a random individual are IBD.

We will often refer generically to "inbreeding" without distinguishing whether we mean the individual or the population because of this equivalency.

Maintaining Diversity in Finite Populations

A diploid base population of size N will have $2N$ distinct alleles at each locus. What is the probability that *none* of these alleles is lost when passing alleles to the next generation?

Assume that each parent has precisely two offspring. Then the parent must pass one allele to one offspring (happens with probability 1) and the other allele to the other offspring (happens with probability $\frac{1}{2}$). So, the probability that *all* parents pass on both alleles is

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

a very small number! Therefore, it is very likely to lose one or a few alleles during each generation.

Please contrast this conclusion with the same situation under HWE. An allele that starts out at proportion $\frac{1}{2N}$ (i.e. it is present as one copy in a size N population), will persist in a HWE forever at the same allele frequency $\frac{1}{2N}$.

Genetic Drift & Inbreeding

In a finite population, the persistence of an allele is not guaranteed. We already know there is a very good chance that not all alleles will make it to the next generation. So, some will be lost. If the population size remains constant, the lost alleles are replaced with IBD copies of other alleles. These other alleles gain in frequency and the overall fraction of ibd alleles in the next generation will have increased.

So, small population size and inbreeding go hand-in-hand. Genetic drift is synonymous with increasing levels of inbreeding.

Genetic drift is a dominating force in small populations.

Allele Fixation

If we carry the above process of allele loss at each generation forward in time, it leads to the conclusion that ultimate there will remain only one allele in the population. While at first it is very easy to remove alleles, the numbers of the remaining alleles increase and it is less likely that they will be removed at each generation. However, there is always a positive, though small, chance that an allele goes extinct in each generation. Combine these probabilities over enough generations, and eventually all alleles but one will go extinct. In the end, genetic drift leads to the **fixation** of one allele at every position in a small population. The population will become completely inbred.

Clearly, something in our argument does not apply to real life since all populations are finite, but they are not entirely inbred. There are other forces (mutation, migration, selection) that can counteract the effects of genetic drift. We will discuss them and their relationship to genetic drift later.

3.2 Wright-Fisher Model - Accumulation of Inbreeding

3.2.1 Haploid population

Haploid Finite Population

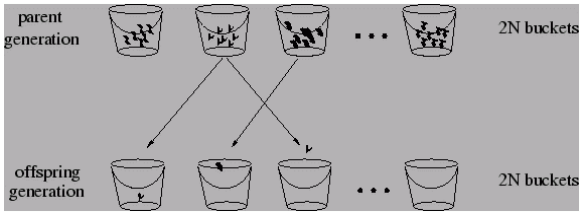
So far, our arguments have been intuitive. We now seek to develop a precise model describing the accumulation of inbreeding or the loss of diversity in finite populations.

Consider a haploid population of size N. If every individual copies itself to produce the next generation, there is no change due to genetic drift. However, if some individuals fail to copy themselves, while other copy themselves more than once, there will be random changes in allele frequencies and genetic drift will apply to the population.

Hence, genetic drift in haploid populations is determined by the variance in number of offspring. If there is no variability, there is no genetic drift.

There are many ways that the number of offspring could vary among individuals. The simplest and most mathematically elegant is to think backwards: Assume that each offspring randomly selects his/her parent.

Biological Plausibility



If (1) each parent produces a huge number of offspring initially, but (2) each parent produces exactly the same huge number of offspring, and (3) these offspring are killed off at random (without regard to genotype) until there are only N left, then the above model will apply quite well. Killing off at random can also be thought of as selecting a few lucky survivors (N to be precise) at random.

The key observation is that each time one of the lucky survivors is selected, the pool of offspring from that parent is not substantially decreased.

Population Inbreeding Recurrence Relation

Let f_t be the amount of inbreeding at generation t. Here, we mean population inbreeding, since there is no concept of inbred individuals.

We will develop a recurrence relation for f_t with time.

$$\begin{aligned} f_t &= P(\text{two individuals have ibd alleles}) \\ &= P(\text{ibd} \mid \text{two indivs. select same parent})P(\text{two indivs. select same parent}) \\ &\quad + P(\text{ibd} \mid \text{two indivs. select diff. parents})P(\text{two indivs. selected diff. parents}) \\ &= 1 \times \frac{1}{N} + f_{t-1}\left(1 - \frac{1}{N}\right). \end{aligned}$$

The solution is $f_t = \frac{\frac{1}{N}}{-\frac{1}{N}}\left(1 - \frac{1}{N}\right)^t - \frac{\frac{1}{N}}{-\frac{1}{N}} = 1 - \left(1 - \frac{1}{N}\right)^t$.

Interpretation of Solution

Looking at the solution

$$f_t = 1 - \left(1 - \frac{1}{N}\right)^t,$$

we see that in the limit $f_t \to 1$. And each generation, the probability of non-ibd alleles $h_t = 1 - f_t$ declines by another fraction $\frac{1}{N}$.

Note also that the above is the average behavior of many populations. Now random chance is very important force in these small populations. So the actual level of inbreeding in a population at generation t will be something around f_t but may show substantial variation from population to population.

3.2.2 Diploid population

Diploid Equivalent

Now, consider a diploid population where all the parents produce huge pools of gametes that are thrown into a common gametic pool. Assume random union of gametes and assume hermaphroditic adults that can self-fertilize. Then, the offspring can be thought of as randomly selecting two individuals from the previous generation to be its parents.

In fact, this model is equivalent to the haploid model in the sense that each offspring can be thought of as randomly selecting *two* alleles from the preceding generation. As before, the probability that these two alleles are IBD is given by the recursion equation:

$$f_t = \frac{1}{2N} + f_{t-1}\left(1 - \frac{1}{2N}\right),$$

since now there are $2N$ gametes.

The solution as before is $f_t = 1 - \left(1 - \frac{1}{2N}\right)^t$ with $2N$ replacing N.

Rate of Loss of "Heterozygosity"

Inbreeding results in the increase of ibd alleles and obligate loss of heterozygosity (each new inbred individual replaces either a homozygote or a heterozygote, the latter replacement resulting in gradual loss of heterozygotes from the population). How fast does this occur?

Let $h_t = 1 - f_t$ be the probability that the two alleles at a locus are *not* ibd. Then,

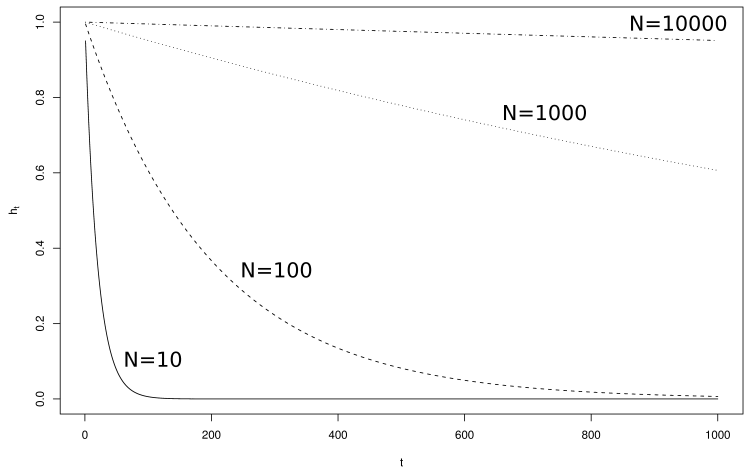
$$h_t = \left(1 - \frac{1}{2N}\right)^t.$$

Clearly, the smaller N, the faster the increase in heterozygosity.

And examining the update equation

$$f_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1},$$

we see in each generation a fraction $\frac{1}{2N}$ of all ibd alleles are *new* ibd alleles (i.e. *new* inbreeding).



Recovery from a Bottleneck

Sometimes populations are subject to bottlenecks (brief periods of time when their numbers drop to very low numbers where genetic drift dominates). An interesting question is whether restoring their numbers will reverse the "damage" done to during the bottleneck. To be specific, is inbreeding reduced after the population numbers are restored?

To answer this question, we will assume the population ends the bottleneck with inbreeding at $f_0 > 0$ and then subsequently the population immediately rises to infinite size $N = \infty$. What will happen to inbreeding?

We examine the update equation

$$f_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1} = f_{t-1} = \cdots = f_0,$$

so the level of inbreeding remains fixed at $f_0 > 0$ despite the huge increase in population size.

Time to Target Heterozygosity Loss

We can use the time-dependent equations to also find how much time is needed to lose a pre-defined amount of heterozygosity. For example, the half-life $t_{0.5}$ for heterozygosity, the time it takes to remove $\frac{1}{2}$ of the current heterozygosity h_0 is

$$\begin{aligned} h_{t_{0.5}} &= \left(1 - \frac{1}{2N}\right)^{t_{0.5}} = \frac{h_0}{2} \\ t_{0.5} &= \frac{\ln h_0 - \ln 2}{\ln\left(1 - \frac{1}{2N}\right)} \end{aligned}$$

For small x, $\ln(1-x) \approx -x$ by Taylor's series so for large N, $t_{0.5} \approx \frac{\ln h_0 - \ln 2}{-\frac{1}{2N}}$.

Or when the starting heterozygosity $h_0 = 1$, meaning no inbreeding, $t_{0.5} \approx 1.386N$.

(Note: For haploids, the half-life is $t_{0.5} \approx 0.693N$.)

Half-Life of Heterozygosity

$t_{0.5}$ can be thought of the half-life of heterozygosity. The diploid result

$$t_{0.5} = 2N \ln(2)$$

shows that the half-life is proportional to the population size. It takes longer to lose heterozygosity if the population is larger. For reasonably large populations, genetic drift is a slow process.

For example, a population size of 1 million that reproduces every 20 years would require 28 million years to lose half of its heterozygosity. The first monkey-like primates first appeared about 40 million years ago.

3.3 Wright-Fisher Model - Allele Frequency Variation

3.3.1 As a Markov Chain

Wright-Fisher - Haploid

We have described a formal model for the accumulation of inbreeding in finite populations. Using the same model, viewed from a different angle, we will now formalize the relationship observed earlier, that genetic drift relates to variability in allele frequencies. In essence, the more variability in allele frequency in time (or across multiple replicates), the *faster* inbreeding accumulates and the *faster* the population *loses* diversity (alleles).

In a population of size N, the frequency of allele A at generation t is

$$p_t \in \left\{ \frac{0}{N} = 0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N} = 1 \right\}.$$

Suppose that currently $p_t = \frac{i}{N}$, so there are i copies of allele A in the population. The next generation offspring will each independently select alleles at random from the current generation, and with probability p_t, each will select an A allele. If X_t is the number of allele A copies in the population at generation t, then

$$X_{t+1} \sim \text{Binomial}(N, p_t).$$

We can define transition probabilities $P_{ij}(t) = P\left(X_{t+1} = j \mid X_t = i\right)$, where

$$P_{ij}(t) = \binom{N}{j} p_t^j \left(1 - p_t\right)^{N-j}, \qquad \text{with } p_t = \frac{i}{N}.$$

Wright-Fisher - Diploid

For the simple diploid model, the offspring select their *alleles* randomly from the previous generation (without worrying to make sure that they obtained one from a father and the other from a female). Then, the transition probabilities are

$$P_{ij}(t) = \binom{2N}{j} p_t^j \left(1 - p_t\right)^{2N-j},$$

with $p_t = \frac{i}{2N}$.

With the simple diploid model we need not change the equations except to multiply all N's by 2.

We have defined a simple Markov chain.

Some Results from Markov Chain Theory

The *state* of this Markov chain is the number of A alleles in the population. At each generation, this state is updated according to the transition probabilities defined above.

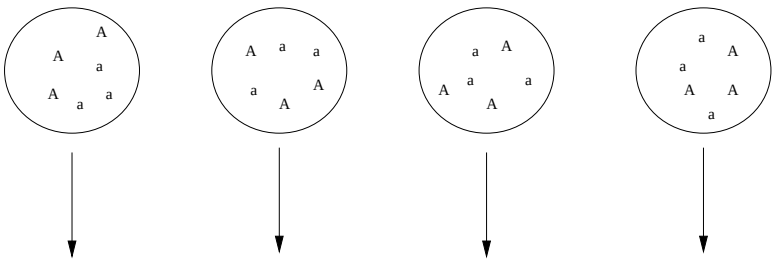
How the chain's state is updated depends *only* on the current state of the system (the current number of A alleles). How it will be updated does not depend on the whole history of this population. For this reason, the stochastic process is called Markovian.

In addition, we observe that there are two *absorbing states* (0 and 1) from which the chain cannot escape, meaning once that state is entered, the chain will never leave that state. This is of course true *only* in the absence of mutation and migration.

All other states are what are called *transient states* meaning that they will only be visited a finite number of times. *Eventually* the chain will bump into one of the absorbing states, preventing any more visits to the transient states.

Properties of this Process

While the Markov chain can provide us with many useful results about this process we are modeling it fails to give us a closed form solution for the distribution of future allele frequencies. In fact, there is no known solution to this complete specification of the process. Instead, we must rely on summary descriptions (e.g. mean and variance) to describe the random process of genetic drift.



3.3.2 Mean allele frequency

Mean of Genetic Drift

Recall X_t is the number of A alleles in the population at time t, and X_{t+1} is a binomially-distributed random variable with success probability $\frac{X_t}{2N}$. Then, by the properties of the binomial distribution

$$E\left(X_{t+1} \mid X_t\right) = 2N\frac{X_t}{2N} = X_t.$$

In general, because $E[E(X|Y)] = E[X]$, we have

$$E\left(X_{t+1}\right) = E\left(X_t\right) = E\left(X_{t-1}\right) = \cdots = X_0,$$

where X_0 is the initial number of A alleles in the population.

Divide through by $2N$, to obtain

$$E\left(p_{t+1}\right) = E\left(p_t\right) = E\left(p_{t-1}\right) = \cdots = p_0.$$

Fixation Probabilities

All populations ultimately become fixed.

Examine a collection of populations that all start with the same allele frequency p_0. At time t, let Y_{it} indicate whether the ith population has fixed allele A. We want to know what proportion of populations fixed allele A as $t \to \infty$, i.e.

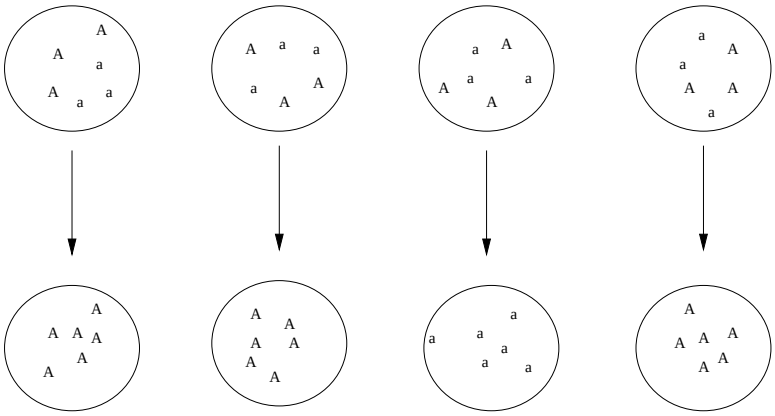
$$\lim_{t\to\infty} P(Y_{it} = 1) = \lim_{t\to\infty} E(Y_{it})$$

Because the expectation does not change in time, the overall frequency of A allele across all these populations must be p_0, therefore

$$p_0 = P(Y_{i0} = 1) = P(Y_{i1} = 1) = \cdots = P(Y_{it} = 1) = \cdots = \lim_{t\to\infty} P(Y_{it} = 1)$$

and we conclude that the probability that A fixes in the population is equal the initial starting frequency of allele A.

The more A allele present at the beginning, the more likely it will persist in a population (or it will persist in greater fraction of replicate populations).



3.3.3 Variance of Allele Frequency

Variance in Allele Frequency

Since $X_{t+1} \sim$ Binomial$(2N, p_t)$, the binomial variance provides

$$\text{Var}(X_{t+1} \mid p_t) = 2Np_t(1 - p_t).$$

Converting to a variance in allele proportions, we have

$$\text{Var}(p_{t+1} \mid p_t) = \frac{p_t(1-p_t)}{2N}.$$

This expression indicates that the variability in allele frequency introduced each generation is related to the current allele frequency and is mazimized when $p_t = \frac{1}{2}$. Furthermore, it decreases as population size N increases.

However, we really want the unconditional variance $\text{Var}(p_{t+1})$. We will derive a recurrence relation for it.

We can also think of the next generation count as $X_{t+1} = X_t + e_x$, or the next allele frequency as

$$p_{t+1} = p_t + e,$$

where e is some random deviation from the previous allele frequency. Notice, e has mean $\text{E}(e) = \text{E}(e_x) = 0$, since $\text{E}(p_{t+1}) = \text{E}(p_t)$.

In addition, all the variance of the update to $p_{t+1} = p_t + e$ is contained in the deviation e

$$\text{Var}(p_{t+1} \mid p_t) = \text{Var}(e) = \text{E}(e^2),$$

since p_t is non-random by the conditioning and $\text{E}(e) = 0$.

$$\begin{aligned}
\text{E}(p_{t+1}^2 \mid p_t) &= E\left[(p_t + e)^2\right] = E\left(p_t^2\right) + 2E\left(p_t e\right) + E\left(e^2\right) \\
&= p_t^2 + 2p_t\text{E}(e) + \text{E}(e^2) \\
&= p_t^2 + \frac{p_t(1-p_t)}{2N} = p_t^2\left(1 - \frac{1}{2N}\right) + \frac{p_t}{2N}.
\end{aligned}$$

However, we seek the unconditional probability

$$\text{E}(p_{t+1}^2) = \text{E}\left[\text{E}(p_{t+1}^2 \mid p_t)\right],$$

which is obtained by taking a second expectation over all possible p_t.

$$\text{E}(p_{t+1}^2) = \text{E}(p_t^2)\left(1 - \frac{1}{2N}\right) + \frac{\text{E}(p_t)}{2N}$$

but we know allele frequency is unchanging in expectation $\text{E}(p_t) = p_0$ and of course, by definition of variance

$$\text{E}(p_t^2) = \text{Var}(p_t) + p_0^2.$$

Thus,

$$\text{Var}(p_{t+1}) + p_0^2 = \left[\text{Var}(p_t) + p_0^2\right]\left(1 - \frac{1}{2N}\right) + \frac{p_0}{2N}.$$

The equation we have just derived can be rearranged into a recurrence relation for the allele frequency variance

$$\text{Var}(p_{t+1}) = \text{Var}(p_t)\left(1 - \frac{1}{2N}\right) + \frac{p_0(1-p_0)}{2N},$$

with initial condition

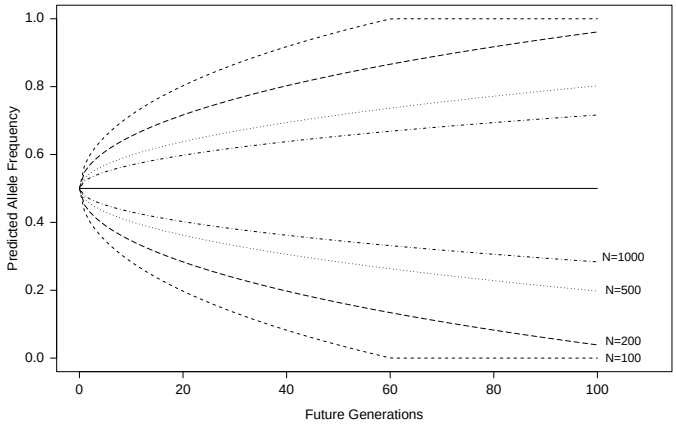
$$\text{Var}(p_0) = 0.$$

This is a standard recurrence relation you know how to solve. The solution is

$$\text{Var}(p_{t+1}) = p_0(1-p_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right].$$

As $t \to \infty$, the variance in allele frequencies approaches $p_0(1-p_0)$.

Predicted Allele Frequency and Variance



Genetic Drift and Inbreeding

Notice that

$$\frac{\text{Var}(p_{t+1})}{p_0(1-p_0)} = 1 - \left(1 - \frac{1}{2N}\right)^t = f_t.$$

This quantitates the relationship between allele frequency variation and inbreeding. The amount of inbreeding at generation t is direct proportional to the degree of variance in allele frequencies between replicate populations of the same process.

Total Variance

Recall we spoke of two sources of variance in genetic data: (1) sampling variance and (2) genetic variance, where the latter is caused by the fact that replicate runs of evolution under the same conditions do not produce identical populations. In infinite populations, the results *are* deterministic, and replicate runs produce the same results. It is only in finite populations where variation between replicate populations would be observed.

In any case, $\text{Var}(p_{t+1})$ is genetic variance.

3.4 Effective Population Size N_e for Non-Wright-Fisher Populations

Handling Assumptions - Effective Population Size

Of course we have made some rather strong assumptions in deriving these results. How can we model more realistic populations?

One way is to compare real populations to the *ideal* Wright-Fisher population we have been assuming. Specifically, we can find the size of the *ideal* population that would having the same level of inbreeding as we observe in a real population.

The (inbreeding) **effective population size** of a real population is the size of the ideal population (satisfying our hermaphroditic random mating model describe previously) that would have the same level of inbreeding as our real population.

The symbol used to represent (inbreeding) effective population size is

$$N_e.$$

Multiple Definitions of N_e

To match levels of inbreeding, we can choose N_e such that the fraction of new inbreeding produced each generation (recall this is $\frac{1}{2N_e}$ for the ideal Wright-Fisher model) matches the fraction of new inbreeding produced in the real population.

Most often this is how the matching between ideal and real populations is done, but there can be other ways to compare populations and make them "match." The concept of effective population size can then be difficult to interpret, especially if it is not clarified how the equivalency is made.

3.4.1 The Selfing Assumption

Disposing of Selfing Assumption

The assumption that might have you most worried is the assumption of hermaphroditic species. Very few species actually satisfy this assumption and it seems that it might cause some difficulties. To quantitate how big an effect distinct sexes might have on the results, we will explicitly model it in two phases. First we remove the assumption that parents can "self" (fertilize themselves). Next, we remove the assumption that anyone can mate with anyone, and force male plus female matings.

As before, let h_t be the probability that two alleles at a locus are non-ibd. Let k_t be the probability that two alleles selected at random from two randomly selected individuals are not-ibd. First note

$$h_{t+1} = k_t.$$

We'll derive

$$k_{t+1} = \frac{1}{2N} h_t + \left(1 - \frac{1}{N}\right) k_t.$$

Second Equation

k_{t+1} is the probability that two randomly selected alleles are not ibd in two random individuals. They can only be non-ibd if the two individuals selected these genes from different parents or they selected different alleles from the same parent.

- They select the same parent with probability $\frac{1}{N}$, then with probability $\frac{1}{2}$ they select separate alleles. With probability h_t, those two alleles will be non-ibd.
- They select different parents with probability $1 - \frac{1}{N}$. The probability that they select non-ibd from these two parents is k_t.

All together, the equation resulting is

$$k_{t+1} = \frac{1}{2N} h_t + \left(1 - \frac{1}{N}\right) k_t.$$

Solving Recurrence Relations

Using $h_t = k_{t-1}$, we have

$$k_{t+1} = \frac{1}{2N} k_{t-1} + \left(1 - \frac{1}{N}\right) k_t.$$

If we assume logarithmic linear grow, $k_{t+1} = \lambda k_t = \lambda^2 k_{t-1}$ or

$$\lambda^2 k_{t-1} = \frac{1}{2N} k_{t-1} + \left(1 - \frac{1}{N}\right) \lambda k_{t-1}.$$

The relevant solution is

$$\lambda = \frac{1 - \frac{1}{N} + \sqrt{1 + \frac{1}{N^2}}}{2},$$

which is just slightly different from $1 - \frac{1}{2N}$, the factor that applies in the Wright-Fisher model.

N_e for Non-Selfers

To compute the effective population size, we need the population size N_e that if evolving like the ideal population would have the same single-generation change in inbreeding as observed in this non-ideal population. In other words, we need

$$\frac{1 - \frac{1}{N} + \sqrt{1 + \frac{1}{N^2}}}{2} = 1 - \frac{1}{2N_e}.$$

It turns out that

$$N_e \approx N + \frac{1}{2},$$

so this population, where individuals must select two distinct parents, behaves like an ideal population where there is $\frac{1}{2}$ additional individual in the population. That $\frac{1}{2}$ extra individual will slow the effects of inbreeding very slightly. Essentially, there is very little impact of forcing selection of distinct parents.

3.4.2 The Hermaphrodite Assumption

Adding Distinct Sexes

Suppose there are N_f females and N_m males and each offspring must select one random father and one random mother from these pools. If the allele of interest is not sex-linked, then sex is assigned independent of the genotype.

We use the same quantities h_t and k_t. As before,

$$h_{t+1} = k_t.$$

Two genes from different individuals can only be non-ibd if they selected these alleles as two distinct alleles from

- the same female parent
- the same male parent, or
- entirely different parents.

Furthermore,

- Two individuals select the same female parent with probability $\frac{1}{N_f}$ or the same male parent with probability $\frac{1}{N_m}$.
- Once the same parent is selected, the individuals will select the same allele with probability $\frac{1}{2}$. They must do this twice.

$$\begin{aligned} k_{t+1} &= \frac{1}{4}\left[\frac{1}{2N_f}h_t + \left(1 - \frac{1}{N_f}\right)k_t\right] + \frac{1}{4}\left[\frac{1}{2N_m}h_t + \left(1 - \frac{1}{N_m}\right)k_t\right] + \frac{1}{2}k_t \\ &= \left(\frac{1}{8N_f} + \frac{1}{8N_m}\right)h_t + \left(1 - \frac{1}{4N_f} - \frac{1}{4N_m}\right)k_t \end{aligned}$$

Compare this equation to the one for distinct parents:

$$k_{t+1} = \frac{1}{2N}h_t + \left(1 - \frac{1}{N}\right)k_t.$$

N_e for Distinct Sexes

The size N^* of the distinct-parent population that would acquire inbreeding as fast as this distinct-sex population is given by

$$\frac{1}{N^*} = \frac{1}{4N_f} + \frac{1}{4N_m},$$

which yields

$$N^* = \frac{4N_f N_m}{N_f + N_m},$$

but then

$$N_e \approx N^* + \frac{1}{2} = \frac{4N_f N_m}{N_f + N_m} + \frac{1}{2}.$$

Notice, that when $N_f = N_m = \frac{N}{2}$, $N^* = N$, and there is no effect of distinct sexes beyond that of distinct parents when the sexes are balanced.

N_e for Distinct Sexes

Nevertheless, there can be a substantial impact when the sex ratio is not equal. For populations with $N = 100$ but unequal sex ratios, the effective population size is

N_f	N_m	N_e
1	99	4.46
5	95	19.5
10	90	36.5
25	75	75.5
50	50	100.5

3.4.3 Monogamy

N_e for Monogomous Species

Of course the human species favors monogamy for the reproduction process. How does monogamy effect N_e?

To model this, have the N adults split into $\frac{N}{2}$ pairs that last for life (we assume $N_f = N_m = \frac{N}{2}$ so that the sexes match up).

Then, when offspring pick their parents, they first pick a pair (family) and then start picking genes.

$$\begin{aligned} h_{t+1} &= k_t \\ k_{t+1} &= \frac{2}{N}\left(\frac{1}{2}\cdot\frac{1}{2}h_t + \frac{1}{2}k_t\right) + \left(1 - \frac{2}{N}\right)k_t \\ &= \frac{1}{2N}h_t + \left(1 - \frac{1}{2N}\right)k_t, \end{aligned}$$

and we conclude that monogamy when there are equal fractions of each sex does not change the rate of inbreeding fixation.

3.4.4 Varying Population Size

Varying Population Size

What if the population size N_t depends on time t? When the population size varies, the update equations reveal

$$h_t = \prod_{i=0}^{t-1}\left(1 - \frac{1}{2N_i}\right).$$

The equivalent ideal population would have

$$_eh_t = \left(1 - \frac{1}{2N_e}\right)^t.$$

Equating these such that the same level of inbreeding (or non-inbreeding actually) is achieved by generation t, we have

$$\left(1 - \frac{1}{2N_e}\right) = \prod_{i=0}^{t-1}\left(1 - \frac{1}{2N_i}\right)^{1/t}.$$

N_e for Varying Population Size

When all N_t are large despite the variation from generation to generation, the approximation

$$\left(1 - \frac{1}{2N_i}\right)^{1/t} \approx 1 - \frac{1}{2N_i t},$$

so we have

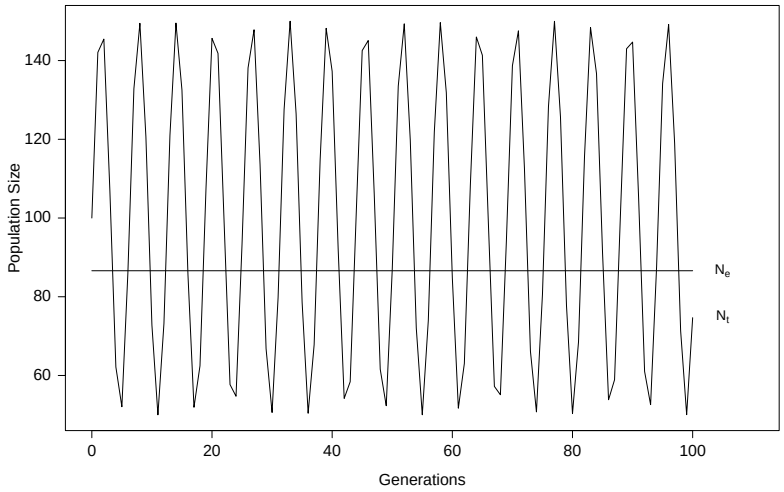
$$\left(1 - \frac{1}{2N_e}\right) = \prod_{i=0}^{t-1}\left(1 - \frac{1}{2N_i t}\right).$$

Solving for N_e, we have

$$N_e = \frac{t}{\sum_{i=0}^{t-1}\frac{1}{N_i}},$$

i.e. N_e is the **harmonic mean** of the population sizes across all those generations. Harmonic means tend to strongly weight low values over high values, so a few dips in the population size will strongly dip the corresponding effective population size N_e.

Oscillating Population Size



3.4.5 Varying family sizes

Random Variation in Offspring Numbers

We will now consider the effect of variability in the number of offspring.

All individuals in the ideal population produce an infinitely large number of offspring which are then selected randomly to proceed into the next generation.

This is clearly an unreasonable assumption because

- obviously, no one can produce infinitely many offspring, and
- the number of offspring produced per adult may vary (perhaps because of random local environments good or bad for raising offspring).

In relaxing this assumption, we will continue to assume infinitely many offspring are produced, but different "levels of infinity" apply to different parents.

Probability of Selecting Two Gametes from Same Parent

Suppose the fitness of the ith individual is w_i. These fitness vary from individual to individual *not* for genetic reasons but because of random environmental effects. In the infinite gamete pool from all adults, there will be a proportion $\frac{w_i}{\sum_i w_i}$ of gametes produced by individual i.

The probability that two gametes come from parent i is

$$\left(\frac{w_i}{\sum_i w_i}\right)^2.$$

The probability that any two gametes come from the same parent is the sum of the above over all possible parents

$$\sum_i \frac{w_i^2}{\left(\sum_j w_j\right)^2}.$$

New Inbreeding per Generation

The amount of new inbreeding introduced per round of replication is just the probability that *the same* two gametes are selected from the same parent

$$\frac{1}{2}\frac{\sum_i w_i^2}{\left(\sum_j w_j\right)^2}.$$

The fraction $\frac{1}{2}$ is the diploid factor, the probability that the *same allele* is selected both times from that parent.

The amount of new inbreeding introduced by each each round of replication of an ideal population is $\frac{1}{2N_e}$, so we conclude the effective population size for this real population with varying offspring numbers is

$$N_e = \frac{\left(\sum_j w_j\right)^2}{\sum_i w_i^2}.$$

N_e for Offspring Variation

We are now going to write this in terms of the variance in offspring numbers (or equivalently the variance in the individual fitnesses w_i).

$$\text{Var}(w_i) = V_w = \frac{1}{N}\sum_i w_i^2 - \bar{w}^2,$$

where $\bar{w} = \frac{1}{N}\sum_i w_i$ is the mean fitness across all individuals. Therefore,

$$\begin{aligned} N_e &= \frac{N^2\bar{w}^2}{(NV_w + N\bar{w}^2)} \\ &= \frac{N}{1 + V_w/\bar{w}^2} = \frac{N}{1 + C_w^2}, \end{aligned}$$

where $C_w^2 = \frac{V_w}{\bar{w}^2}$ is the squared coefficient of variation of fitness.

Interpretation

Notice $N_e < N$, so variation in offspring numbers reduces the effective population size below the census size and increases the rate of appearance of inbreeding. This should make intuitive sense. Some individuals will contribute more offspring to the next generation and others will fail to contribute at all. The results is more ibd alleles in the next generation.

The derivation requires that the fitnesses w_i are constant throughout the generations, well at least that the mean and variance are constant. So, since these fitness differences are caused by the environment, the environment must be effectively constant from generation to generation. There can be local variation, but the overall "average environment" and variability in the environment must remain constant.

Varying Surviving Gametes

It is also possible to write the effective population size in terms of the variation in the number of gametes n_i that survive into the next generation from individual i. Now, we are measuring absolute contribution to the next generation, rather than relative contribution (it might be easier to measure).

Then (details not shown),

$$N_e = \frac{4N - 2}{2 + V_n},$$

where V_n is the variance in n_i across the population.

When all parents contribute exactly 2 gametes to the next generation, then $V_n = 0$ and $N_e = 2N - 1$, so the real population acts like a much larger ideal population; inbreeding is slowed.

Wright-Fisher Variance V_n

Notice, the Wright-Fisher model also has an intrinsic variance in offspring number.

In Wright-Fisher, n_i is the number of successes in $2N$ draws of a Bernoulli random variable with probability of success $\frac{1}{N}$. The variance is

$$V_n = 2N\left(\frac{1}{N}\right)\left(1 - \frac{1}{N}\right) = 2 - \frac{2}{N}.$$

This variance is the smallest achievable variance when offspring randomly select gametes from parents. So, any other form of selecting gametes that is still random will decrease the effective population size.

On the other hand, if offspring do not select gametes at random the effective population size can be increased. If gametes are not selected at random, the above formula is useful for computing V_n and N_e. The derivation of N_e via V_w assumes random selection of gametes and cannot be used.

Summary

The effective population size is an efficient means of comparing populations by relating them all to a standard, ideal population. It allows us to quickly evaluate a population and mating scheme to determine how much inbreeding it will produce.

If $N_e > N_t$, where N_t is the census size of the real population at time t, then the real population has accumulated less inbreeding than we would have expected under ideal conditions.

If $N_e < N_t$, the the real population has been accumulating more inbreeding than the same size ideal population.