An Introduction to Evolutionary Biology

Prof. Sutirth Dey

Biology Department, Population Biology Lab

Indian Institute of Science Education and Research (IISER) Pune

Week 5 Lecture 27

Effective population size and evolution of mutation rates

Hi, so in our last discussion, we looked at a few examples of drift, and I said that if we are to go beyond this topic, We need to figure out some ways by which we can measure the effects of genetic drift. Now, the simplest way of getting a foothold into this is through the concept of heterozygosity. Now, what exactly is that? To understand that, let us go to our favorite simulator. Here we are, and we say start the tool and simulations; we go to replicated simulations. So, since we are trying to study drift, we will reduce the population size. Let us say we have a population size of 10, and let us say we have generations; we can make it 200, good enough.

Starting with an allele frequency of 0.5 is fine, and let us say we start with 10 populations and run the simulation. What do we see? As expected, we find that sometimes allele A1 gets fixed and sometimes allele A2 gets fixed. No surprises there; that is what you expect out of drift.

Now, the problem is that we do not know whether A1 will get fixed or whether A2 will get fixed; if you track any one of them, You know, either you track p or you track q; their frequencies sometimes go up and sometimes go down. However, there is one thing whose frequency is generally always going to go down. And that is the frequency of the heterozygotes, the heterozygosity. Why do I say so? Now, in this particular case where I am showing you the genotypic frequency, Unfortunately, they are giving me the

genotypic frequency only for the last simulation, right? But you can clearly see that the green one, this is the, you know, frequency of the heterozygotes; you can see that it is falling. Why will that happen? Very easily.

We know that the ultimate fate under drift is either that the allele will become fixed or it will be lost. In either case, p will either become very high, go to 1, or become very low, going to 0. Now, we know that the frequency of the heterozygotes is 2pq. And we also know that it gets maximized when p and q are close to each other. Actually, you know p and q given that both of them are between 0 and 1; the quantity gets maximized when p = q = 0.5. But the important part is that if either p or q goes down, then the heterozygosity The frequency of the heterozygotes also has to go down. So, that is a simple concept that whenever an allele is drifting, it is going towards fixation or loss. The frequency can go up or down, but the heterozygosity of the population will always go down. So, with this extremely simple concept, we can now go back to our PowerPoint, which is here. So, here I talked about the fact that when the population is moving toward fixation or loss of an allele, Then the heterozygosity is going to be reduced.

Now, there are multiple ways in which this can happen. Of course, the sampling size, due to small sampling error from a small population size or drift, is one reason. but other things can also happen. So, for example, suppose you have mating between close relatives, Then, as we are going to see in one of our subsequent discussions, that can also lead to. In general, there is an increase in the homozygosity in the population and therefore a reduction in the fraction of the heterozygotes.

Similarly, if you have selection happening, then that can also push one allele's frequency close to 1 or 0. As a result, the frequency of the heterozygotes in the population will again go down. Therefore, although drift by itself in the long run will reduce heterozygosity, reduced heterozygosity does not necessarily imply that drift is happening. There are also other things that can reduce heterozygosity. So, now the other major thing to keep in mind is what we mean by heterozygosity.

Now, I gave you a definition, which is the fraction of heterozygous individuals in the population, and that is a correct definition, but it turns out that Like many other concepts in ecology and evolution, the term heterozygosity is also used in multiple contexts. And therefore, you have to be a little careful when you read the word "heterozygosity" in the literature. Some of them might mean it in slightly different contexts. Often, it becomes clear just by reading the context what exactly they are talking about. Sometimes it does not; so, you have to be careful.

But anyway, in the context of our discussion for this course, We are going to stick to heterozygosity as a fraction of heterozygotes in the population. Now, here we have a problem. I have been saying that heterozygosity is going low, but how exactly am I defining low? I mean low and high; these are all relative terms, right? They have to be defined in the context of something else. So, compared to what? How low is low? So, it turns out that the comparison is always in the context of something known as an ideal population or an idealized population. Now, what do you mean by that exactly? So many places you know simply say all the population that is following.

All the assumptions of Hardy-Weinberg equilibrium define an ideal population. That is actually a wrong definition. So, an ideal or idealized population in the context of genetic drift is supposed to have the following properties. First, we assume that all individuals are hermaphrodites. In other words, all individuals are able to produce both male and female gametes, and mating is random which basically means that any male gamete can fertilize any female gamete. But in this particular case, self-fertilization is explicitly allowed. So, in real-life biology, there are many cases where self-fertilization never happens, but in the context of an idealized population, Remember, it is a theoretical construct; in that context, you assume that self-fertilization is allowed. And of course, you have no mutation, no migration, and no selection. This part is exactly like Hardy-Weinberg.

But then the last point—this is the crucial one—the population size remains constant across generations; it does not change; it is finite. This is very, very important. So, now that you have defined an idealized population like this, Obviously, some of you are

thinking, "Why, why exactly am I taking these assumptions and not other assumptions?" So, the reason you need precisely these assumptions has to do with how you make the derivation. We are not going to look at the derivation in the context of this particular course. Which is why it becomes a little difficult to see where these assumptions come from.

But many of you will probably end up taking an advanced course, and at that point, You will do the derivation for heterozygosity and, or rather, the change in heterozygosity over time. And at that point, it will become clear why precisely you need these assumptions and not others. So, anyway, now that we have defined the ideal population, we have established a benchmark against which to measure. We are going to compare; we can now ask the question: what is the size of the ideal population? How many individuals are there in the ideal population that undergo the same amount of drift? As the population I am interested in, the focal population, the given population. This size of the ideal population undergoes the same amount of drift as my population of interest.

This size is known as the effective size of the given population, which is very important. The effective size is a theoretical construct. It is the corresponding size of the ideal population that is undergoing change; it is a number of individuals in an ideal population. That is undergoing the same amount of drift, and in this particular case, we are going to say. It is losing heterozygosity at the same rate as the population of my interest.

Now, please appreciate that when it comes to understanding drift, you can think in terms of heterozygosity. But there are many other quantities in terms of which drift can also be defined. And since you can define drift in terms of other quantities, you can define your effective population size in terms of those quantities. So, there is no one effective population size in the literature; there are multiple kinds of effective population sizes in the literature. And depending on which definition you are looking at for the given population, you can have multiple effective sizes, right? I mean that should not be too tough because if you are thinking in terms of how fast it is losing heterozygosity, You think in terms of the heterozygosity-based effective population size.

If you think in terms of how fast its variance is changing, the variance of alleles, Then you think in terms of what is known as the variance effective population size, and so on and so forth. So, there are three or four types, and it depends on the quantity of drift. Whichever aspect of drift you are interested in, you end up getting a different, you know, estimate of your effective population size. And all these effective population size values can be very, very different from each other. Typically, though not always, the value of the effective population size is less than the census size.

However, in many cases, it can even be greater than the census size it depends on. Which concept are you using and under what context are you measuring it? Anyway, that is the underlying conceptualization and theorization of effective population size. But in order to actually measure it, you need to have some numbers, right? So, now I will talk about two or three different measures that are often used. So, for example, suppose you have a population—just one population. And you want to know, as of this moment, what its effective population size is.

You are not interested in knowing how it is losing heterozygosity or some other quantities. You want to know then and there what the effective population size is. So, if you have a scenario like this, then typically, for sexually reproducing organisms, the formula is given like this. Effective population size Ne = 4Nf*Nm/(Nf+Nm), where Nf is the number of females and Nm is the number of males. Now, if you just look at this expression, certain things should be very clear to you.

You can see that what you have in the numerator is a product of Nf and Nm. And what you have in the denominator is essentially the total population size, Nf + Nm. Now, for any given total population size, the closer the value of Nf is to the value of Nm, The larger the product is going to be, the larger the numerator will be, and therefore, the larger the Ne will be. In other words, if you have a population in which the number of males and females is hugely different, just a random example, Let us say there are 100 males and only 2 females, or there are, you know, 200 females and just 5 males. In all those

cases, the effective population size is actually going to be much lower than the census size simply because that product is going to become smaller. Whereas the same 100 individuals in the population, instead of having, say, 90+10, If you have, say, 50 + 50, then the product on the top becomes 50 * 50 = 2500, of course, multiplied by 4 and so on. But the product above in the numerator becomes large. Whereas, if it is, say, 90+10, it becomes 90*10=900. So, that is what I mean by saying that if the males and females in the population are in equal numbers, then, The effective population size is the highest, and any deviation from the 1:1 ratio leads to a reduction in the effective population size.

The second way of measuring effective population size that is very commonly used, when you are assuming that Males and females are in the same ratio of 1:1, or let us say you are thinking in terms of asexual populations in some cases. Then, across multiple generations, the effective population size is given by this formula: t/((1/N1)+(1/N2)+...+(1/Nt)), where t is time. So, N1 is the population size at generation 1 or time step 1, N2 is the population size at generation 2, and so on. So, you can see that this is the formula of the so-called harmonic mean of the population sizes over generations. So, it is a property of the harmonic mean that it is affected more by lower values of n than by higher values of n. And therefore, when I say, when are you going to use this kind of formula? You are going to use this kind of formula when, let us say, a population size is fluctuating over generations. You want to know what the one value is that I can say is the effective population size over the entire period of time.

And you can see that each time it is going down due to the property of the harmonic mean. The low values are going to have a disproportionate effect on the harmonic mean compared to the higher values, which is why. When you have a population whose size is fluctuating over time, it is the population crashes that end up affecting. The effective population size is more than the population booms. And finally, remember that we were talking about heterozygosity or loss of heterozygosity.

So, the formula for the loss of heterozygosity is given as Ht+1 = Ht * (1 - (1 / (2Ne))), where Ht is the heterozygosity at generation t. What fraction of the population is

heterozygous at generation t, and Ne is the effective population size? Again, if you look at this formula straight away, you can see certain things: Ht+1 = Ht * [multiplied by something]. So, heterozygosity at generation t+1 is the previous generation's heterozygosity multiplied by that thing, (1-(1/2Ne)). Now look at (1-(1/2Ne)). The larger the value of Ne, the smaller (1/2Ne) will be because Ne is in the denominator. And the smaller is going to be (1/2Ne); this (1 - (1/2Ne)) is going to be larger, right? I mean (1 - (1/2Ne)). will always obviously lie between 0 and 1. So, (1-(1/2Ne)) will become larger and therefore, Ht+1 will be. A bigger fraction of the previous generation's heterozygosity. In other words, the smaller the value of Ne is, The faster heterozygosity is lost, the larger the value of Ne; the slower heterozygosity is lost.

Now, I am not giving you the formula here, sorry, the derivation of the formula here, simply because That is not a part of a basic course, but anyone who is more mathematically inclined and would like to see for themselves. How this comes, just go to this particular link. This is the course note of somebody, and it is not a part of the course. But you can assure yourself that this comes from proper, you know, mathematical logic. This particular equation that you are seeing is known as Sewall Wright's equation of genetic drift the same Sewall Wright whose viability selection equation you know we looked at. So, with this understanding, we can now revisit our Buris experiment, which we saw in our previous discussion. So, remember in Buri's experiment, we started with 107 populations, all of which were explicitly made up of heterozygotes. And then over time, we saw that the genetic composition of the populations started to diverge from each other. And by generation 19, you had about 30 populations in which the allele, in this particular case bw75, was lost. Around 28 populations in which this bw75 allele was fixed. So, this is the classic case of drift. Now, obviously, when you have these two extremes in generation 19, when one allele is fixed in 30 populations, The other is fixed in 28 populations; then, obviously, the overall heterozygosity of the population has gone down drastically, right? But as you can see, that has not happened in one generation; it has happened over time. So, how exactly does this thing proceed? Here is the graph. So, here on the x-axis, we have generations 0 to 19, and on the y-axis, we have the average heterozygosity.

So, what is the average heterozygosity across 107 populations, and that data is plotted in the middle curve? In green, the one that is bouncing around, and you can see that I have also plotted two other lines. The first line is theoretical Ne = 16. What is that? That is simply this equation where I have fit Ne = 16. I have put Ht as the first generation's heterozygosity, and I have simply simulated it, which has led to this line. So, theoretically speaking, if my Ne were to be 16, then this is the way the heterozygosity should have come down.

Now, why did I choose N = 16? That is because, remember, in this particular population, Buri had kept the population size constant at 8 pairs: 8 males and 8 females. Throughout the experiment, which is why I am assuming that the census size equals the effective population size; hence, Ne = 16. And I have done the same thing for this blue line below, except that here I have ended up taking theoretical Ne=9. And now I want to compare the green line, which shows how heterozygosity actually went down, with the orange line and the blue line. And what do I see? Although initially, it looked as if the reduction in heterozygosity was similar to the Ne=16 curve, but very soon you will know it actually became much, much closer towards the later part. It completely, I mean not completely, to a great extent overlaps with the Ne=9 line, which essentially means That in the long run, this population is behaving as if it has only nine individuals of the idealized population. And that is why we say that the effective population size in the context of Buri's experiment was only 9, even though. the census population size was 16. So, this gives us some idea about how we can measure the effect of drift in real life data.

Now, with this information, we are going to ask the question that What exactly does drift do to a population when it is interacting with other evolutionary forces? Remember, in real life, drift never operates alone, nor does selection for that matter, typically. So, in real life, drift has to interact with selection. So, what really happens when this interaction occurs? And the example that we are going to look at will be in the context of the evolution of mutation rates. Now, as we all know, there is a tremendous variation in mutation rates across all life forms.

So, for example, if you look at simple bacteria like E. coli, they have a mutation rate per base substitution. Roughly of the order of 10^(-10); whereas, if you take something like Drosophila melanogaster, it is about 10^(-9). Homo sapiens, which is us, 10^(-8) and all the way up to the viruses that can go as high as 10^(-5) or 10^(-4). So, comparing bacteria to viruses, you have everything ranging from 10^(-4) to 10^(-10) that is about 6 orders of magnitude. That is a very very you know large amount of variation. Now, we have already discussed Most of the mutations that actually end up affecting the organism are either deleterious or lethal. So, if that is the case, then there should actually be a very strong selection against mutation rates. However, if that is the case, why do you expect to get or why do you get that tremendous variation that we actually saw? Why does the mutation rate not evolve to become 0? So, this was the very famous question asked by Sturtevant. So, this is the person who figured out that you can use the recombination rates to map the distances between loci on a chromosome.

He did that when he was a teenager and an undergraduate, which ultimately led to a Nobel Prize. And of course, you know he made many other contributions, but he also ended up very succinctly putting forth the question. Why does the mutation rate not evolve to 0? There can be multiple hypotheses in this context; we are going to deal with just 1 or 2. So, one possibility is that there exists a cost of fidelity with replication speed. Now, what do I mean? Now, we know that different organisms have cells that are replicating at very different rates.

Some bacteria probably replicate every 20 minutes or so on average. Whereas you know, humans and some other cells are replicating at a much slower pace. Now, we know that whenever a cell replicates, it needs to create copies of its DNA. Now, we also know that this copying process, or replication process, is actually error-prone. So, when the bases are being inserted by DNA polymerase, it sometimes ends up incorporating wrong bases.

But fortunately, there are multiple mechanisms by which these errors can be corrected. So, for example, DNA polymerase itself has the ability to go back, take out a wrong base,

and then go forward. So, the hypothesis is that if you have, you know, a speed requirement, you need to replicate fast; then, if you every time, you know. If the wrong base is incorporated, DNA polymerase has to go back and change it, which is obviously going to slow it down. In other words, that is going to negatively affect the replication speed, and therefore, the entire variation that we see in mutation rates.

So, different organisms replicate at different rates; therefore, they have different degrees of tolerance towards How much they can do in terms of error correction, and because of that variation, you get the entire variation in mutation rate. Now, what is the problem with this logic? The problem with this logic is that if this were indeed correct, then you would expect All those guys who are replicating very, very fast are the ones who will be paying the cost of replication. In terms of speed, they are the ones who will have the maximum mutation rate. In other words, you expect the bacteria, for example, to have a very low mutation rate. Sorry, a very high mutation rate; whereas, in reality, you see exactly the opposite.

As I showed you in that particular chart, bacteria have much lower mutation rates than Say vertebrates; however, it is a vertebrate that replicates much, much lower than the bacteria. So, obviously, this is not going to work out. So, another hypothesis is that after all these error corrections, These are all physicochemical processes, and all physicochemical processes will have some upper limit. We can only correct mutations up to this limit and not further.

Not every mutation can be corrected. And because of selection, all organisms have actually ended up reaching the physicochemical limit of this accuracy. Of replication and repair; therefore, no further changes can be made. Great. But here we have a problem. The problem is, as far as we understand, the chemistry of replication and all these processes are remarkably conserved across all taxa.

So, if that is the case, if these limits are indeed due to physicochemical reasons, then they should be similar across all taxa. If they are similar across all taxa, you expect a certain

mutation rate, but you expect that mutation rate to be the same everywhere. You actually do not expect to see the huge amount of variation that I ended up showing you. So, that tells us that perhaps this hypothesis is not a good explanation for why there exists so much variation. So, as I said, there are many other hypotheses, and one can, you know, examine them if one is in an advanced course.

So, we will not go into those; we will go straight into a hypothesis which, as of this moment, seems like the most credible one. And this is the so-called drift barrier hypothesis. So, what does it say? So, it says that selection is continuously trying to reduce the mutation rates in populations. However, the effectiveness of selection depends on the amount of drift that the population is facing. If you have high drift, then selection's ability to increase or decrease the frequency of some alleles will be compromised.

And therefore, we also know that the amount of drift a population is experiencing is inversely proportional. To its effective population size, the so-called Ne. Therefore, the whole idea is that species that have lower Ne, They will end up evolving a higher mutation rate because selection is not very effective on them. Whereas those with higher Ne will end up having a lower mutation rate. Now, understand that effective population size is not a property of the species by itself.

Effective population size, I mean, part of it is a property; it is a property of the physiology, but part of it is also very ecological. It depends on how many resources it gets, it depends on who else is there with the same resources, and so on and so forth. Therefore, this hypothesis is saying that, due to ecological reasons, organisms will end up having very different kinds of Effective population sizes and those with higher effective population sizes will have lower mutation rates and vice versa. So, just to show you the same thing graphically. So, here this is just a cartoon; we have effective population size on the x-axis and a trait like mutation rate on the y-axis.

Now, selection is trying to bring the mutation rate down, and selection is more effective when the effective population size is high. So, the length of the green arrows shows you

the effectiveness of selection, and it clearly indicates that. when effective population size is high, selection is more effective and that is why the trait goes down further, The mutation rate goes down further. On the other hand, the red bars represent the effect of genetic drift. When the effective population size is low, then the red bars are of no longer size, and that is because that is the point at which genetic drift has a much greater effect when the effective population size is low and vice versa. So, if this entire logic is correct and if this indeed is the reason for which you get so much variation in the mutation rate, Then the prediction is that you should get a negative correlation between mutation rates and effective population size. Is this correct? So, in order to figure this out, Michael Lynch and you know other authors. So, Michael Lynch is the person who is the primary mover behind this particular hypothesis. So, what they did was measure, or you know, from the literature, they figured out mutation rates in a large number of species.

I think they have about 123 species, and they also measured their effective population sizes or obtained the estimates. And when they plotted the effective population size on the x-axis and the mutation rate on the y-axis, you can see a Very nice strong negative correlation with a pretty high R^2; $R^2 = 0.86$, and as I said, this is data from 123 species. Ranging from bacteria to archaea to unicellular eukaryotes to invertebrates, vertebrates, and vascular plants.

So, that is a very broad coverage, and you can see that it is a very strong relationship. Now, please appreciate that I am just giving you the gist of the whole thing. The drift barrier hypothesis is a very sophisticated hypothesis. The mathematics of it is extremely well done. by Lynch and his colleagues, and there are, you know, many, many papers and a lot of discussion on the entire thing. I am not going into the nitty-gritty, gory details of that; I am just giving you the overall picture.

The overall picture is very simple. The overall picture says that the ability of natural selection to refine a phenotype is ultimately limited by the noise created by. Random genetic drift, which is itself a consequence of the finite population size that the particular species is experiencing. So, that is the crux of the whole matter. Now, note something

here: in this definition, you do not even see the mutation rate.

So, the basic principle is generic; it is applicable to any kind of phenotypic trait. The trait in the context of which Lynch et al. have studied it in great detail is the mutation rate. But that does not necessarily mean that it would apply to other places. So, one of the things that What people are trying to do is see whether, in the context of other traits, this is also going to be operative or not. So, this is one of those eco-evolutionary theories where ecological and evolutionary forces are coming together to tell us.

What is the ultimate outcome? So, we are going to stop here with this. This is where our examination ends. Drift and its effects will end, and in the subsequent discussion, we are going to talk about. What are some of the other forces that can also affect evolution, and do you know some examples thereof? So, see you in the next discussion. Bye.