The following content is provided by MIT OpenCourseWare under a Creative Commons license. Additional information about our license and MIT OpenCourseWare in general is available at ocw.mit.edu.

PROFESSOR: Ready? OK. Well, welcome to the third lecture. Quick review of what we did last time-- first slide. We talked about purification, basically at every level of the hierarchy in complexity, from the elements all the way up to organisms, and in particular, the awesome purification one can get by serial dilution down to single molecules-- recombinant DNA molecules, in this case, embedded in E. coli cells-- and how this purification led to a revolution, first in biochemistry, then in recombinant DNA.

Molecular biology then led to the Genome Project and systems biology, which brings us to models of interconnections. And to start in with algorithms that are useful in systems biology, we start with one of the simplest and most robust ones, which is this genetic code, shown here in the lower right-hand corner.

But as one looks at the huge diversity in the tree of life, you find examples of exceptions to almost everything you can come up with, including the genetic code. And we talked about how to cherish the exceptions, as usual, here. And then, getting back to systems biology, in terms of how one creates qualitative and quantitative models from functional genomics data and establishes evidence.

Finally, we ended on mutations and selection, as we will do in all three lectures here. This lecture in particular will focus on mutations and selection.

So what is today's menu? We'll start with types of mutants and how they're represented for bioinformatic purposes. We'll talk about the three main methods by which mutations occur, drift, and select so that you can determine the frequency of different alleles in populations. In doing that, we'll rely on our friend from the first lecture, the binomial distribution, in the context of an exponentially growing population.

Then, give you some very practical training here on association studies, where we illustrate it with a very important example of HIV resistance and a very useful statistic, easily as useful as the binomial and the Gaussian-- the chi-squared statistic. And then, we'll continue to talk about association in the context of causative alleles and importance of getting haplotypes, and then technologies that are required that have been used to get the framework first genome and how one might change strategy in getting subsequent genomes in order to make cost-effective these very large association studies.

And finally, in the context of that, we'll talk about random and systematic errors in more detail so you get lots of examples and ways of dealing with it computationally.

So just to go from our brief discussion about our friend, the 100% DNA sequence identity, or amino acid sequence identity that you might find in the lens protein and enolase enzyme, we can see that, even at 100%, you can find differences in function. And so functional measures are a good adjunct to DNA or amino acids' identity.

At 99.9% identity, we're talking about the level of single-nucleotide polymorphisms that might exist between any two positions in your mother and your father's chromosomes in your body, or the differences between one of your genomes and one of my genomes. About once every kilobase, there will be a polymorphism, a difference. It's often a single nucleotide, an A for a G.

Then, as we go to 98% differences, then we're talking about the differences between one of our genomes and a chimpanzee genome-- in other words, a completely different genus and species. However, by the criterion we mentioned in a previous lecture about bacterial definitions of species, those would be almost identical, and you'd have to go to identities less than 70% in order to call it a new bacterial species. So you can see that this is a very soft number, very dependent upon the context and the branch of the tree of life that you're working with.

Then, we have sequence homology and very distant homologs which are only detectable, as we'll see when we get to the proteomics section of this course, by three-dimensional structures, not by sequence. And that switchover-- sequence homology will be the topic of next lecture, and the very distant one in 3D structures will be later, in proteomics. But that occurs at about 25% to 30%.

This is just a reminder and an introductory slide to the next slide. We have different phenotypic effects due to different types of mutations. And ignoring the phenotypes that we talked about before, you have the types of mutations. Classes are null mutations, dosage, conditional mutations, gain-of-function, altered ligand specificity.

Now, that's in broad terms, colloquially. But how is this achieved on a more molecular level, and how do we represent it compactly for bioinformatic discussion? So we have single substitutions. These can be a single base par, an AT base pair, to a CG base pair, or a GC base pair to a TA base pair, and so on.

There can be deletions and duplications. These can range. The deletions and duplications be as large as a chromosome or as small as a single base pair. You can delete that A rather than change it into a C, and that would be one base pair.

So when you delete an entire chromosome, that's called aneuploidy. The example in the previous slide was trisomy 21. If you remove that chromosome 21 instead of having three copies of it, it would be monosomy. And these are huge consequences even though they're fairly subtle changes in dosage.

Now, a special class of deletion and duplication occurs when you have a tandem repeat of a sequence. When you have anywhere from a single base repeat, AAAAA, or dinucleotide, trinucleotide, all the way up, those have a very high tendency for both forward and reverse mutation, both deletions and duplications, because at the very small level, you have polymerase slippage and other microscopic events. And then, at the larger level, you have some kind of recombination, often homologous recombination, that occurs that cause deletions and duplications of tandem repeats.

An inversion here. We're representing a complicated chemical event which involves double-stranded DNA going 5 prime to 3 prime on the top strand, from left to right, and 5 prime to 3 prime on the bottom strand, going from right to left. And you're taking a little chunk of that and flipping it so that you break some bonds and remake them. All the 3 prime is conserved, but you've done a reverse complement of the top strand and the bottom strand. In that reverse complement, you basically turn Cs into Gs and change the order.

This is sometimes abbreviated for simple genetic description. Let's say we inverted CDE. You might change the case or add primes to it, color it. In some way or another, you indicate that it's now a reverse complement, and of course, the gene order or the DNA segment order is reversed to indicate that inversion.

Translocations, insertions, and recombination have in common that you will make a break somewhere, and then you will bring in a new piece of DNA, just like the inversion, where you inverted in place. Here, it involves something acting at more of a distance. You'll break between B and C and X and Y, and you'll, in this case, do a reciprocal translocation where you conserve all the DNA, and A and B is now next to Y and Z.

An insertion is like that, but now the DNA need not be reciprocal in any sense. It can come in, more or less, from outer space. Here, it came from Greece. It can be a transposable element that came in from who knows where and inserted in between the A and B.

Recombination. Here, I illustrated this homologous recombination. And you can have an non-homologous combination, but this is by far the most common and interesting. Even some non-homologous recombination involves short regions of homology where you basically have either two sister chromosomes or homologs from mother and father or paralogs within the genome where you've got to duplicate a gene that might even be on the same chromosome arm.

And now, just as with the deletions and duplication, where you have tandem repeats, if you have repeats anywhere in the genome, you can take a small difference. We're emphasizing the similarities here, but they have a few small differences that allow you to track them. Those can be exchanged by single-strand or double-stranded DNA, various chemistries, and you can either do a nice reciprocal exchange here, where you preserve all the DNA and the little C gets replacing the big C by some kind of crossover between D and E, somewhere between C and F, or you can have a nonreciprocal exchange where you pick up a little bit of C and duplicate it, and now there's no big C left over and only little C. Called gene conversion.

But you get the idea. These are a fairly exhaustive list of the kinds of simple, elementary mutations you can have. Now, you can pile these in various combinations with each other, and over long periods of time, you can get completely new sequences. There's even ways that you can get de novo synthesis of DNA, such as mechanisms of terminal transferase, which is used-- yes?

AUDIENCE: What are the [INAUDIBLE]?

PROFESSOR: Oh, we're going to get to that a little bit later on, but I can give you a preview. In human genome, these point mutations occurred about 10 to the minus 8 per base pair per generation. Deletions and duplications, especially in tandem regions, can be as much as six orders of magnitude higher-frequency than that. So there's a huge variation from position to position and types of alleles that will determine the mutation rate, and you should be quite aware of that as you go through various computational exercises.

This is just a bit of commonly accepted nomenclature. Mutations and polymorphisms are basically same thing. There are differences between you and me. They become polymorphisms, meaning when they are common alleles, when their frequency is greater than 1% in a population. This is common in the human genome community. It may differ for other ones. But this is roughly where a mutation which is rare-- less than 1%-- becomes a polymorphism, which is frequent-- more than 1%.

As a counterpoint to that, I would say that there is a good chance that every possible single-nucleotide polymorphism that could exist does exist in a population as large as ours and mutation rates which are modest but still, over long periods of time, allow the accumulation of mutations such that, rather than having maybe 3 million common single-nucleotide polymorphisms, or SNPs, we might have 12 billion-- one at every position.

And in a later slide, we'll actually go through and calculate crudely why the frequency should be around 10 to the minus 5 and why there should be about 10,000 of us representing each of these so-called rare alleles. They're rare individually, but they're common as a group.

It will make a difference as to whether we're talking about whether these polymorphisms are linked to your favorite trait or whether they actually cause it, whether they're part of the cause. No particular one can be said to be 1-to-1 with a cause. It's all a collection of mutations and environment.

Now, haplotypes. What do we mean by haplotypes? We've introduced SNPs, single-nucleotide polymorphisms. If you have a SNP that, let's say, is involved in causing a APOE4 we'll introduce in just a moment, an allele that is associated with increased risk of Alzheimer's disease-- and let's say that protein has a known change in the protein three-dimensional structure, and you can call that the bad allele, or the associated allele.

Now if that protein were expressed at a low level-- for example, if you had a promoter mutation or enhancer mutation-- then that haplotype, that combination of promoter mutation and protein mutation, is a more significant predictor than either one of them separately. And the fact that they're on the same chromosome is important because if the promoter or enhancer that determines the level of the protein in cis, meaning on the same DNA, or in trans, makes a huge difference.

So that's what haplotyping is all about. It's determining what mutations are in cis on the same DNA in order to make associations biologically meaningful and, in terms of systems biology, interpreting them, in terms of what you know about regulatory elements and protein elements.

These haplotypes can be inferred indirectly from diploid data from how alleles segregate in small families or in siblings and so forth. Or, easier to think about and probably more accurate, in general, especially with limited data sets, is direct observation.

The most extreme direct observation is you pull out a DNA molecule that has, say, your promoter mutation in your putative protein mutation, or just a series of linked polymorphisms. By cloning out or otherwise physically isolating that DNA molecule, you can sequence it, and you can determine. By definition, if they're all in the same sequence, then they're on the same molecule.

But you have to be careful about the sequencing method that you're using there and the specific cloning and/or physical separation method, because there are certain methods where you can get a chimeric sequence either due to cloning of two species together or somehow disassembling them through bioinformatics.

You can also directly observe it when you go through the genetic processes of mitosis, which we've talked about before, which is that, as the cells divide, they split up the chromosomes, or meiosis, which is the process by which they get prepared for recombination, which we'll come back to in a little while.

So when you do want to do it by linkage, you want to do a direct observation. And the best way to follow the haplotype is, if there is a difference at every position in both the parents and the child inherits those differences, then it's called informative. If the parents happen to share a single-nucleotide polymorphism even though the child is a heterozygote for one of them, it could be that the parents have additional alleles that can confuse things, and it's not informative.

But the point is, if you have enough single-nucleotide polymorphisms, you can do a case-control study where you have lots of children that are affected for whatever trait you're interested in and, hopefully, a close-to-equal number which are in the control group which do not have it.

An example of that-- and I'm just trying to give you a flavor for this and where to look, rather than to completely empower you on this, because that would be an entire separate course-- but you can look for association. And you have to worry about things like structure and admixture, where you've had populations that have developed independently in different parts of the world and have been randomly mixing, which is part of the model in these separate populations. But then, you bring them together very recently, and now it's no longer fair to model it as if it were a uniformly mixing population.

And we'll refer throughout the course to the null hypothesis, which is the thing that you're trying to rule out, and the probability refers to the probability that you can reject this null hypothesis. And in this case, you're trying to reject that the allele frequencies in the candidate locus, whatever you picked, do not depend on the phenotype within the subpopulations. And that's the way that they deal with these case studies.

Now, what are some of the motivations for studying either individual polymorphisms or haplotypes-- combinations of polymorphisms can affect the activity of a protein? Now, I could use hundreds of different examples of well-established and useful examples.

But here's one that, hopefully, will hit a resonating chord in the sense that these are actually used now in certain clinical settings, and certainly in clinical research, to ask whether a patient population, either in the process of developing a new drug or using an established drug to keep the patient toxicity down and the effectiveness up.

And so in the far left-hand column, is the gene or enzyme affected? And then, in the middle is our examples of drugs which interact with this enzyme. And then, the quantitative effect is in the far right-hand side. For example, thiopurine methyltransferase is something which, if you have a large amount of the activity, whether you have, say, a very active enzyme and/or a very active promoter element that causes high levels of it, then these various chemotherapeutics that are used for fighting cancers-- the amount of them need to be adjusted.

So you have lots of the methyltransferase. That means you have to give lots of the drug, or else your drug study will fail, or your patient will succumb to cancer because it's ineffective. You haven't added enough. The thiopurine methyltransferase is overcoming the drug.

On the other hand, if you have very low levels of the modifying enzyme, you want to lower the dose, or else you'll have toxicity. So that's an example that's actually being used in clinical situations where you can use the information to adjust drug levels or to stratify your patient population so that you put patients into different classes or exclude them from the study altogether because you know that the drug will have some bad effect. And this, hopefully, decreases the costs.

On the downside to pharmaceutical development, if you do stratify your patient population and you make it through the drug study with that caveat, then the FDA will require that you put that proviso, and that makes the size of the population that will be buying the drug smaller, since the costs of developing the drug are fairly fixed. It decreases the profit.

Now, I pointed out that there may be a very large number of rare single-nucleotide polymorphisms. But in terms of common ones, we're getting pretty close to saturating the common ones. And there are databases of these just like there are databases of almost everything you can imagine, some of them better than others.

The common ones will, of course, be the ones that either are neutral, with respect to their phenotypic effect. That is to say, it doesn't really matter whether they're one common allele or the other one. The one allele might be at 30% and the other one at 70%, but they're both fairly neutral with respect to function.

Or it could be that they both provide different advantages in different scenarios. Or the heterozygote, where you have one over the other, provides some advantage, and that's kept it in the population. But it's unlikely that they're highly deleterious because the highly deleterious alleles are going to be rare. They're going to be selected against. And we're going to fully model that in just a moment.

Now, let's say that, somehow, anybody who wanted their genome could have their genome tomorrow. You could have your complete genome sequence. How would you, then, as computational biologists, prioritize the single-nucleotide polymorphisms you find in there, relative to the whole genome sequence, which is in GenBank? Now, this would be an excellent project for you to do for the term project.

But what you might say, first of all-- what single-nucleotide polymorphisms would you throw out, for example, or put low on your priority list? Or which ones would you put high on your priority list? Yeah?

AUDIENCE: Introns would be low-priority.

PROFESSOR: OK, introns.

AUDIENCE: Maybe this is a very simplistic thing to say, but I guess [INAUDIBLE] matter if they're different from one another. [INAUDIBLE].

PROFESSOR: That's fine. Everybody has their own pet part of the genome they don't like. Introns almost sunk the Genome Project. They said, why are we going to sequence the 98% of the genome that doesn't code for proteins? Fortunately, we went ahead and sequenced it anyway.

Another favorite thing that people mention is repetitive DNA. That was another part of the genome. But they didn't actually sequence it from Drosophila, the repetitive DNA. And it's considered also not protein coding. And I'm going to give you a couple of examples, as we go through here, to illustrate other points, but also to illustrate that repetitive regions that are not in protein-coding regions, whether introns or other non-coding regions, can be important.

And here's an example. This is one of the most repetitive elements in the human genome. It's called an Alu repeat. As those of you who have done bioinformatics before realize, it's the bane of our existence, in terms of assembling and searching and so forth.

But here's an example of a single base mutation in this repeat. There's about 500,000 copies of this in the human genome scattered about. It's called an intersperse repeat as a consequence. And this A-to-G transmission is found upstream from the myeloperoxidase enzyme-encoding gene.

So how do we find out whether this is important in any sense? First of all, the observation is that it is associated with several-fold less transcriptional activity. That particular position creates or destroys binding sites for these transcription elements, and that might be the reason that it has lower transcriptional activity.

And finally, it is over-represented in a particular type of cancer. And we're going to go through the ways that we take an observation, like a polymorphism, move to an association, like here, with cancer, then take it to a mechanism, such as here, with the transcriptional regulators.

I think that's what this is about. It's not sufficient to observe the polymorphism. You can't say, a priori, whether it's important or not, whether the Alu repeats are not conserved. That's another thing that people say. Throw out all the nonconserved single-nucleotide polymorphisms. It's not conserved. It's not present in mouse, for example. It's non-coding, and it's repetitive.

Now, in addition to the types of mutations, we have the modes of inheritance-- that is to say, the different ways in which a change, a polymorphism can be transmitted. And I include this to broaden your perspective. Rather than getting entirely fixated on the 3 million DNA SNPs, let's broaden the discussion here a little bit.

You can have not only DNA polymorphisms, but you can have RNA polymorphisms, which are heritable. For example-- and I use this an an extreme case-- RNAi-- 22 nucleotide or so. Probably a variety of mechanisms. Bits of RNA can be induced in various ways. And once induced, they can replicate, essentially, within a cell and between cells. They can spread throughout an organism and probably be propagated over generations between different generations of organisms. So this is heritable.

And you can consider it epigenetic or genetic polymorphism, depending on the nomenclature that you adopt. Even a protein conformation can be considered a polymorphism that is heritable. The central dogma tells us this protein is encoded by a nucleic acid, and there certainly is still the case-- even these heritable polymorphisms and proteins.

But it has a different conformation, and that conformation recruits other conformations and so is inherited not only within a cell, but between cells and between organisms and is the causation of things like mad cow disease and so on.

And finally, modifications of biopolymers, such as methylation, can occur. And this is not formally a DNA sequence change, but it's a heritable change that can have very significant effects on things like cancer and gene expression in general.

Now, this is a broadening of the definition of polymorphisms. And now let's talk thoroughly about the ways that it can be inherited as horizontal or vertical. Horizontal typically means between species, but in a certain sense, it could reflect some of the things that are going on here with RNA and protein inheritance in the sense it is being horizontally transmitted between different cells within the same organism.

But generally, it's a mechanism that does not involve mitosis or meiosis of the nuclear or organellar genomes. And the natural processes are transduction and transformation, being distinguished by-- transduction typically involves a viral or protein coat for the nucleic acid, and transformation involves something closer to naked nucleic acids. Transgenic is a completely laboratory-based version of these two more natural methods.

I think vertical we've already talked about. This is what we normally think of inheritance. Horizontal, we saw in the tree of life, is very common, even in the late branches and the early branches. Vertical, though, is what we normally think about when we're doing crosses in the laboratory. Some of these are maternally inherited, like mitochondria and chloroplasts, but it's still the same kind of process-- mitosis, segregation of DNA.

So now we've got types of mutations we want to talk about, mutation drift and selection, as the main source of the frequencies that we find in populations. We want to know, where do allele frequencies come from? And I will maintain that, generally speaking, in almost all living systems, whether they're cells from organisms that are mutating or whole organisms, whether they do recombination or not, they will do mutation selection and drift.

Now, to develop a fairly rigorous model here, yet simple, we have some assumptions. We always have assumptions in models. If people tell you there are no assumptions, then you need to dig a little further. The assumptions that we'll make here for a little while-- and then, I'll give you a nice example to undermine them all-- but it's constant population size n.

You have random mating. Remember, we were talking about admixture before. Every member of the population can randomly mate. They're non-overlapping generations. This is a convenience.

We are not making any assumptions about the population allele frequencies being at equilibrium. Those of you have taken biology courses with-- Hardy-Weinberg makes that assumption. Here, this is a much more general model. It includes non-equilibrium, and equilibrium can be a special case. So this is relatively minor non-assumption. But we are assuming that we do not have an infinite number of alleles, nor an infinite number of population size.

So now ignore everything on the slide but the upper left-hand corner of slide 15 here. This should look familiar. This should look somewhat like the logistic map where we had the incremental, slow exponential increase of one allele in a population over another, or one organism over another, based on the different selection coefficients of those organisms.

And this could be a very small difference. Say, a 1% increase per generation will dominate after a thousand generations or so, quite definitely. And so that's what you're seeing-- the exponential curve-- and then it plateaus as it gets to 100% allele frequency. That's the full range on the vertical axis, is 0% to 100% allele frequency. And generation goes from 0 to 1,400.

Now, what it actually represents in this particular case is closely related, a little more complicated than just one allele replacing another, because here we have diploids-- that is to say, not a haploid that just has one allele, like in the bacterial species that we've implicitly been talking about. But you can have, here, now, three combinations of alleles. You can have, say, the reference genotype of capital A, capital A, which we'd just call 100% fitness.

So we have fitness and selection coefficient, which are interchangeable terms that are-- very trivial mathematical relationship between them. So we'll use w and s in different contexts here. It just has population [INAUDIBLE]. Then, you can have big A, big A, little a, little a, and big A, little a as the heterozygote.

And we're assuming an additive model here, where you get a little more selection with one allele of little a, and then two alleles of little a results in 2s-- 1 plus 2s. And this has a very similar curve to the logistic map, if you had simple allele replacement.

And what you tend to have in this population-- a thousand generations, in the big scheme of things, may seem like a lot to you. But in the big scheme, it's a very short period of time, and so you tend to have alleles at frequency of 0% and 100%. On the other hand, if you have overdominant mode, where the heterozygote has the highest fitness of the three possibilities, where 1 plus s is larger than 1 or 1 plus t, then you aim for equilibrium.

Whether you're starting at close to 0% or close to 100% the allele frequency will converge on some equilibrium point-- in this case, somewhere around 0.6 for one allele and 0.4 for the other. And it could be anywhere. It depends on the relative fitnesses, s and t.

This is just a reminder slide combining two slides from before. That slide connects the logistic map from lecture 1 to the selection coefficients we've been talking about in lecture 2 to the diploids that we're mainly talking about in this lecture, because humans are diploids. And here, just connect this to the fact that these selection coefficients, s, or the fitness, w, is relevant to different environments and the different times that organisms spend in those different environments. And all mutants are tagged by their DNA, and they're pooled, and they're selected, and you can read them out in a variety of methods.

So now let's dig down into where the allele frequencies come from, based on mutation or migration, which we'll lump together here as M, selection, and drift. Now, the mutation will have a forward rate constant and a reverse. This should remind you of the conversation we just had about the different kinds of alleles-- the duplications and deletions-- how they can have different rates.

If they're tandem duplication, then that has a great tendency to delete. If it deletes down to a single repeat, there's now no longer any repeat, and so the chance of generating the exact duplicate is low. So the frequency of forward and reverse mutation is represented by f and r, respectively.

Now, what we'll see-- we're going to walk through this. Starting with a frequency, t sub i, where i is the number of mutants in a population size n. So here, down at the bottom, is i mutants in a population sized n. And we'll see that applying the mutation, applying the selection, and applying the drift are all applications of binomial distributions when we're talking about this discrete population of n individuals.

Remember the three bell-shaped curves [INAUDIBLE]? Curves that can be bell-shaped. Here, it's clearly a discrete population because we have n individuals, taking i mutants at a time. So the binomials that we'll be using-- all three processes have the same form, where you have a combination of some population, n, and a subpopulation, i. And of course, the remainder is n minus i.

And the different combinations are times some probability, because probability is the last parameter in the binomial here. A binomial is a function of n, i, and p-- the population size, subpopulation size, and some frequency. It can be either a forward or reverse mutation frequency. It can be a selection probability, or it can be a drift probability. And we'll see how each of these work out.

So we start with a frequency. You can have i ranges from 0 to If i is 0, then the frequency is 0 over n, or 0. If i is n, it's n over n, or frequency is 100%. Just that same vertical axis we had on all previous slides.

The starting frequencies have some distribution, t sub i, for i going from 0 to n. And now we want to derive a new vector of frequencies, which would be m. And all we do is we apply the binomial distribution for the forward process or forward mutation. And then, once we're done with that, we now use the m's and adjust it.

So now, give it a chance to do the reverse mutation, because you'll generate this binomial distribution forwards, and then you'll give them a chance to revert. Then, you apply selection. A new binomial.

Now, here, the probability of a transition from your mutants, starting with t-- then, you go to m. Now, to get to s, the transition probability's a function of this fitness. Remember, w and s, fitness and selection.

Here, the fitness determines the probability of a transition in a binomial distribution. And there are two slightly different equations that you use, depending on whether the fitness is greater than 1 or less than 1, whether it has a tendency to increase with time, due to selection, or to decrease that allele with time, as a function of selection. And that's what these two cases are here.

And then, finally, after you've applied forward and inverse mutation and selection, whether it's more fit or less fit, then you apply drift. Now, drift just means that-- think of it. If you have a small population of individuals-- let's say they're a set of colored balls in a jar in front of us, here. It's a small number, and I pull a handful out, because in each generation, you're going to duplicate the population.

But we remember the assumption of constant population size. If I pull out the new generation and forget about the rest of the duplicated, they could all be the same color. And that's the chance that you could drift to fixation, where one of them now dominates, not because it's superior, from a selection standpoint, not because it's been mutated in a directed way to that point, but just because of random drift that you have a constant population.

And so you can see how that would depend on how many are in the jar. If I take half of them out of the jar, and the jar only has five in it, then it's a much higher chance that we'll go to fixation quickly than if the jar has millions in it. If I take half a million out, it's very likely that it will still be more or less the same ratios.

And this is exactly how it plays out when you look at random genetic drift. It's very dependent on population size. So here are simulations going out to, say, 150 generations on the horizontal axis and the allele frequency, as usual, going from 0% to 100%.

And we start with a population with 50/50 ratio of two alleles. And what you see is, if the population, the number of individuals, is only 25, then you quickly fix. Maybe 30 generations, you can-- and this is anecdotal-- get fixation.

As it goes up to a population size of 2,500, you can see it, for all intents and purposes, is flat. It will eventually fix. I assure you, this simulation-- if you run it long enough, one of the two alleles will go to zero, and the other one will go to 100%. And you do another simulation, and it might be the other one, because this is random drift, not selection.

So you can see the final frequencies are going to be some complicated function of mutation rates and selection rates and drift rates, which, in that last one, is a function of population size. When the population size is very, very large, you can see that's going to be constant enough. That's why you can see very subtle differences in selection coefficient.

Now, we're going to come back to the mutation, selection, and drift in the context of human genetics in just a moment. But first, I want to tell you the last component of population genetics that we'll be talking about, which is recombination. Now, this doesn't occur in every biological system. I made the argument before that the mutation, selection, and drift occurs in every biological system, from cells to organisms of all types.

In those biological systems where DNA can be exchanged by transduction, transformation, meiotic fusion, and so forth, then you can get recombination. And these two figures illustrate that. On the left-hand side of slide 19 is the non-recombination scenario, and on the right-hand side is the sexual, or recombination-mediated, change in gene combinations.

So let's look at this. Time is going horizontally. And you can see, at the very left-hand side, the beginning of each of these scenarios, you get a certain rate of occurrence of mutations. This is based on the forward mutation rate in our previous equations. And they occur very early on. A, B, and C all occur.

But they tend to die out because of drift. And the population size is small enough that half of them die out and one of them fixes. And A is destined to fix, but it takes some time for it to fix. It starts out at a frequency close to 0, and by the time it's 50% to 100%, it's ready to start picking up a second mutation at the same frequency they were occurring before.

And now you can pick up the B mutation while the C dies off. It comes and goes due to drift. And then, finally, once AB fixes, then you can pick up the C, and you get ABC. That was a long, slow serial process.

On the right-hand side, we see what can happen in the case of exchange of genetic material in a really mixing population now that A, B, and C all occur at the same frequency as on the left. But now, because they're exchanging information very early on, before B has a chance to die off due to drift, it combines its DNA with A, and you get AB.

And some of the small A population is just destined to be fixed anyway, but happens to combine with C. And again, before drift can wipe them out, the very small selection that you have that couldn't overcome drift in the left-hand panel now fixes them all, the really favorable combination of A, B, and C, very early on.

And so there's a whole series of arguments and counterarguments in the population genetics literature about why is there sex. Of course, we all know have our own reasons. But in here, they say there's a huge cost. The counterargument is there's a huge cost of having two genders-- maybe as much as 50%, possibly higher-- because they have different morphologies and different capabilities and so on.

But then, there's this benefit. This is the risk, and the benefit is the earlier recombination. But then there's counterarguments and so forth that we won't go into.

Yeah?

AUDIENCE: Going back to drift. The drift is actually random. What is it that causes either allele to fix? If you have two different alleles [INAUDIBLE] zero for one allele, why wouldn't it just [INAUDIBLE] come back?

PROFESSOR: Well, it can. You can see here that it's starting to head so that it's fixing on one allele, and it changes direction, and it fixes the other one. So it can go all the way down to close to zero and then bounce back up.

This is just some typical simulations here. If you do enough of these, you'll find every possible behavior.

AUDIENCE: So what you're saying-- basically, eventually, if you just give it enough time, it will fix.

PROFESSOR: It will fix. You can't necessarily predict which one will it fix on. If there's no selection, half the time, it'll fix on one, half on the other if they start out at the frequency of 50/50.

But you have to think of it in the context of mutation and selection, too, because they're all acting there. And when you say that something is selectively neutral, what you really mean-- everything has some very, very tiny selective coefficient. But if it's tiny enough and the population size is small enough, then drift will blind you, and you can't see it. But if you get really big populations, then you'll get small drift, and then you'll see more subtle selection.

The human population is very large. Some of the oceanic species are truly enormous. So you need to think about that as a possibility.

So now let's talk about common diseases. The question is, are common diseases really-- to what extent are they caused by common variants? Clearly, some of them are caused by common variants. And we've said, well, common variants really shouldn't be deleterious, because they would be wiped out by selection.

They could be very, very mildly deleterious such that drift will cause them to get fixated or persist, but they can't be really noticeably. Selection coefficients of 10 to the minus 4-- very, very tiny effects-- can be wiped out in normal-sized populations.

So here are three examples of common variants that almost certainly are associated with common diseases. So why are these special cases, in my opinion, rather than the general case? APOE4 I earlier alluded to is associated with Alzheimer's dementia. It's involved in lipid transport and metabolism. And this particular allele, the E4 allele, is present at 20% in humans. About 80% is the other allele, the second most common allele is the E3 allele.

And so you might say, well, this bad allele, this E4, is the one that should be present in the common ancestors of humans. The E3, the good one, should be present in the common ancestors, and this E4 is a recent aberration. It's somehow getting into human population.

But actually, the E4 is the ancestral one, presumably due to some difference in diet or some other selective effects. The E4, which is currently bad for us, we think, was really good for some of our related species. And so we need to think very carefully before we do anything drastic about eliminating this allele from, say, the human population.

Hemoglobin sickle cell, the sickle cell allele, is probably the oldest and most famous of the molecularly characterized alleles. Zuckerland and Pauling made this famous many decades ago. And it exists in 17% of the human population, and this is responsible for oxygen transport in red blood cells. And you saw in one of the earlier slides today those sickle-shaped cells, which have a huge effect on the hemodynamics of the red blood cell.

Well, another red blood cell component, an enzyme, G6PD, which is involved in maintaining the redox function in the cell, is as high as 40%. The mutant, whichever one you want-- they're so close to 50%. They're just two alleles, and one of them is deleterious, in a certain sense, a biochemical sense. But both of these have a heterozygote advantage in being malaria-resistent. And so probably that's the reason that it's common in the population.

And this is probably proving to be the rule that these are examples of that convergence that we saw in the case of balanced polymorphisms where the heterozygote has some advantage, and so you get a balanced point rather than one or the other dominating through drift or selection.

And the third example, which we will develop in much more detail, is CCR5. There's a deletion of 32 base pairs which confers resistance to one of the greatest plagues in the history of the human race, which is the AIDS virus. And its frequency is 9% in Caucasians.

I think we wish that it were 100% when we worry about HIV, but we need to wonder why it's not 100%. There may be some other reason for it being a nondeleted version in so many humans, and we need to understand that. And as far as I know, we don't understand that.

Now, this is the last slide before the break. I promised you that we would take that simple mathematical treatment of mutation selection and drift and show that it actually has some impact, that it's actually used in human genetics. Now, I must warn you that this is not a consensus view. This is a view that I find appealing, in slide 21, that Jon Pritchard has authored. And he titled it in a provocative way-- *Are Rare Variants Responsible for Susceptibility to Complex Diseases?*

And this is a quote. It's customary in theoretical work relating to complex diseases that the allele frequencies are treated as parameters of the model. And typically in models, you'll have derived values and parameters, which are input, the things that the user is expected to provide. And you don't want to have to be guessing at allele frequencies and having those parameters.

So what's new here is that, using an evolutionary process, which includes-- you guessed it-- selection, mutation, and genetic drift, we can learn or, as I say, model the underlying allele frequencies. They can be derived rather than being required as an input.

And this is illustrated here with a model. This is entirely theoretical, but the parameters that are used are based on some genetic studies that have been done, for example, on autism. And so let me just define some of the terms here. This risk ratio is related, in a certain sense, to the selection and fitness coefficients we were talking about before.

Now, selection and fitness coefficients refer to reproductive fitness. And in human genetics, we have broader interests into all sorts of things that either don't affect reproductive fitness, or we don't know how it affects reproduction. But it affects some kind of medical or just some trait of interest.

In this case, the risk ratio replaces the selection coefficient. And this is basically saying, my brother or sister has autism-- if they did-- then I would have a 75-fold higher chance of having it than someone selected randomly from the same population as I come from. So the 70-fold increase is a very high heritability for this particular example.

Also in this example, through some genetics we won't discuss, it seemed reasonable that the number of loci involved might be a large number, on the order of 100 or so. And you can think of a lot of common diseases. When you start listing the well-characterized ones, we start listing the number of genes that either are known or are plausible, like cancer or walking down the street, or whatever. These really complicated traits involve a very large number of cell types, a large number of cell components, and hence a large number of genes.

So here, they've assumed a model with 100 loci, 100 different genes. And they have multiplicative effects of the polymorphisms in those genes between loci. That is to say, you need to have gene 1 working and gene 2 and gene 3, so that's multiplicative. But then, the polymorphisms you introduced into a gene-- you could think there are a lot of genes involved, and there are a lot of different positions within the gene that could be involved, and each of those is additive.

You have a little reduction in activity due to first mutation, and then a second mutation, a third. Each of those is this or this or this or this, and so that's an additive effect. And so they have various historical justifications for having additive penetrance within the gene and multiplicative across different loci, different genes in the genome.

Now, for these 100 loci, there will be a top five that affect the relative risk the most, and that's what's been plotted here, where we have-- the zero curve, you see, is in the absence of any of these multiplicative effects. It's as if you had no loci that were affecting the frequency of susceptibility alleles.

And so, as in most unselective populations, you'll have most of the alleles being either at zero frequency or 100%. Basically, the zero representing the absence of the other the alternatives. The 100% is there.

But if you have these top five loci out of 100 that contribute to the risk ratio, then those are represented by these four curves. The frequency of susceptibility alleles, which range along the horizontal axis now, from 0 to 100%, have what he calls a probability density. It's clearly not a probability density, because it's not going to integrate to 1, but here, it's the probability, the histogram of risk ratios.

So let's take a break, and we'll drill down on the association study in a very interesting case. Any problem sets to hand in can put them here during the break.

AUDIENCE: So I understand about modeling mutation selection.