Course Name: I Think Biology

Professor Name: Dr. Sravanti Uppaluri

Department Name: Biology

Institute Name: Azim Premji University

Week:5

Lecture:26

W5L26_Genetics - III

Hello, and welcome to the third lecture on genetics in this unit. In the last lecture, we talked about the improved cost and accessibility of sequencing. So we went from needing billions of dollars to sequence a human genome, where now we can actually sequence a human genome under $1,000. And in fact, we're hoping that we can do this in $100. So $100 is around 8,000 rupees. And as of 2022, such as last year, hundreds of thousands of people have had their genome sequenced for research or medical applications.

And genome scale data such as SNPs, or also called “Snips", which we will talk about later in this lecture, have been collected for tens of millions of people. So you can see that now the reach of sequencing and naturally the consequences of such technology really has changed the way human society might work. So we'll explore these ideas over the course of this lecture.

So first of all, how easy is it to get your genome sequenced? So we talked about the Human Genome Project last time and discussed how the project was conducted.

And we know that the Human Genome Project was conducted by taking blood samples from donors. But nowadays, all you really need to do, there are actually many private companies that send you a tube, and you can simply spit into that tube. And in a few weeks later, they send you back potential family trees, your ancestry, do you have ancestors who come from completely different countries, say, Portugal, or Mexico or Africa, and they may even tell you about genetic predispositions to diseases. What else is possible?

So when a woman is expecting prenatal genetic screening as possible, and you can detect chromosomal abnormalities, but there is also the possibility of detecting other kinds of genetic disorders. Sequencing has already been regularly put to use as a tool in cancer treatment to verify whether or not you have specific genetic predispositions to specific types of cancer. And treatments have also been tailored to specific kinds of genetic predispositions. In the clinic, genome editing using technology such as CRISPR is also about to change medicine. So genome editing has not directly been, let's say approved, as of 2023 in most countries but this might change very soon. And so targeting gene editing, targeted gene editing could be a reality very soon for us.

So when you sequence a genome, how much data does your entire DNA actually hold? And how can we get a tangible feel for what this means?

So a single human cell, and note that we have millions and billions of cells in our body right, carries the equivalent of 1.5 gigabytes of computer data. So that's about the size of a phone, right. Most, many sorts of up-to-date phones contain that much data. And that's what a single human cell carries. So you can imagine that this is a whole lot of, whole lot of data that you have to sort of process once you get a genome sequenced. And relative to other organisms, the assumption that humans being very complex, or whatever we think of as being complex, would have larger genomes is not actually true, right?

So E.coli have, of course, only five megabases in your genome of this genome size. And you can see that this increases progressively as shown on this slide. From E. coli to the Drosophila, which is 175 megabase pairs, humans jump to 3.1. But there are other organisms, such as the Spruce tree, and Axolotls that have an order of magnitude larger genome sizes. So it's not necessarily the case that a genome size corresponds to what we typically imagine as being complexity, right? Not only that, the genome size is really big, but the number of genes an organism actually has, so protein coding genes, is actually a lot lower than we initially expected.

So we do have a lot of non-coding DNA as well in our genomes. So once you get these gigabytes of data, how do you really analyze it? And that's really the job of bioinformatician. But what is really this field of bioinformatics? And what kinds of skills does a bioinformatician need? Well, it turns out that bioinformatician actually does a whole lot of things, right? You need computer science, because you need to be able to handle a lot of data.

So you need to be able to code, to be able to manipulate this data. You need engineering, you need mathematics, because there's a lot of data you need to be able to think about, how to analyze it, how to make comparisons. And again, look at statistics. And of course, you need to understand the biology, right. So really bioinformatics actually is a large field that is very sort of a multidisciplinary. So biologists, the typical geneticist is no longer somebody who's working in a lab, say looking at fruit flies, or looking at human cells under the microscope.

Really a geneticist could be doing any number of the things that are shown on the screen as well. So let's take an example. Let's try to look at an example of the kinds of things a bioinformatician might be looking at. So, you know, one of the most important things in genetics is to try to study genetic variation, right.

So how different are we from each other? How do we differ from each other? And one of the ways in which this can be done is by looking at Single Nucleotide Polymorphisms, this is sometimes pronounced as SNPs, right. And what this is, is you can see an individual, an example is illustrated here, individual one and individual two. And the genome looks very similar.

This is just an example, right. And you can see that in this particular region, right, two alleles are possible. So either A or G, right. A on one strand, and G or G on the same strand. And on the reverse strand, it would be either T or C, right. Because A is complementary to T and G is complementary to C. So really these two alleles are possible, and that's what we refer to as the SNP, right. So these are single nucleotide region where any one, so in this case, the SNP has two possibilities, either A or G, and that's on the forward strand, right. If you look at the reverse strand, it's either T or C. So we can then calculate the frequency of a particular SNP.

So this SNP that you see here, where it can be A or G, can be quantified in this way. So each of us has two copies of this SNP. And let's say that you have a population, and within this population, individuals can either have GG, AG, AA, or GA, right. And in this example population, there are 10 individuals, and because there are two locations for this allele, there are 20 in total, and there are eight As in this entire population. So the frequency of this, a frequency of A is 40%, right or 0.4. So you can also refer to this frequency as P is equal to 0.4. On the other hand, this is for A, for G, right. Let's give it a frequency Q that is 0.6, right. Which is actually 1 minus 0.4. Okay.

So now imagine that you have to find SNPs across the whole genome, where you don't have just an illustrative set of base sequences, but rather you have 3 billion base pairs. So you can imagine that this isn't something that can be done by hand, right. So you really need a computer to be able to do it. And that's essentially what a bioinformatician, a large part of what a bioinformatician does is to look for these kinds of variations, right. And it turns out actually that looking at using the simple example that we just talked about, we can actually relate this to the Hardy-Weinberg equilibrium.

This is a very simple model that actually helps us predict the frequency of these alleles, right? So the probability of allele A, this is what we just talked about in the previous slide. Suppose it's P, and the probability of allele G, suppose it's Q, right. And suppose you have a population of parents that have any of these combinations, right. As I said, A-G, A-A, A-G, G-G, etc.

And so between the two parents, right, the probability of the mother giving an A corresponds to P, the father giving an A corresponds to P, the father giving an A also corresponds to P. And so the probability of this individual having the genotype A-A is P into P, which is where you get the P squared from. Likewise, if we have the probability of the G allele, we represent it with a Q, right. For somebody to be, to have a combination A-G, either from mom and dad, where the mom donates the G and the dad donates the A, or vice versa, in each case it is PQ, right. So this is P into Q, and P into Q. And then the same logic applies to a kid obtaining, or a progeny obtaining, G-G as the genotype. That's Q squared, so it's Q into Q.

Yeah. And so what this, what the Hardy-Weinberg model tells you really is that it predicts the ratio of these genotypes within a population. Yeah. And that is really the idea. So whenever you have a population and you're looking at SNPs that have two positions, you can predict actually, given that parents actually, the choice of mate doesn't depend on the, this particular allele, you will get this probability distribution for the different 2 alleles, P squared to PQ for A-G and Q squared for G-G. So this kind of a model is probably the simplest to understand, but actually bioinformaticians use many such models to try to predict and to try to predict the, say, genetic variation within a population.

And these models are based on certain assumptions, right. And so if these, if the measured variation doesn't match what they see, doesn't match what they predict based on their model, then there's something that they have to go back and look at. And then they try to discover why it doesn't match the model. So this is really the idea of taking data, trying to see if it matches the model, going back and verifying whether the data matches the model. If it doesn't, what do you do etc. Right. So this is actually the kind of mathematical theoretical approach that bioinformaticians might take.

So SNPs are not the only kinds of variations that you see across genomes in a human population. There are small scale variations. The SNP is the first one that you see listed here, that's what we explored. There are also deletions. So where basically a few nucleotides are, a short length of a set of nucleotides are completely deleted.

That's what's illustrated here. You might also see short tandem repeats and individuals might have different lengths of these short tandem repeats. That's what STR stands for over here. Yeah. Then there are intermediate to large kinds of variations.

So here you see again, there are these tandem repeats that could be much longer. Large scale variations could be in the form of deletions, duplications, inversions. So in the case of inversions, you see that the new nucleotides in red are actually flipped. So they're inverted. And then you may also see complex structural variations where these arise from a combination of any of the small scale, intermediate and large scale deletions that we talked about.

So in order to be able to explore these kinds of variations, you really have to be able to explore and identify, I should say, these kinds of variations. You really need to be able to manipulate these long sets of data. okay.

So another thing to think about is what is the effect of these variations on function? All humans have more or less the same set of genes, but suppose there are, there is a deletion in an individual's gene. Once you identify that this deletion exists and it is different from the rest of the population, you can try to look on, look at what is the phenotypical effect of this deletion.

So suppose this deletion occurred within a protein coding area within the genome,right. Then you can look at whether a specific, that corresponding protein is functional or not. And what is then the, is there, let's say, the emergence of a disease in that individual? And can you correlate this deletion back to that disease? It could also be in, let's say you could also have duplications or inversions in non-coding regions. And so you can, we've had a lecture on non-coding RNA and you know that non-coding RNA have all kinds of regulatory functions, right. And so is this non-coding RNA still able to carry out its function in regulating gene expression? And if not, what are the consequences and so on? So you can try to correlate changes or let's say variations in DNA to function, right. It could be both in the form of protein, but also as RNA itself.

All right. So the other thing that having access to the kinds of variation that you see allows you to do is also study evolution, right. So this is a phylogenetic tree. A phylogenetic tree is basically a method through which we can look at the evolutionary trajectory of different organisms. And what this means is, if you look at this specific phylogenetic tree, you can see that at some point, right, there was a divergence at this. So you can look at from left to right as sort of time, right. There was divergence where Orangutans started to evolve. And this divergence then led to the sort of, let's say precursors of Gorillas, Bonobos, Chimpanzees, and humans, right. And humans diverged from Chimpanzees and Bonobos five to seven million years ago, right. So, you know, on the slide where it says humans are closely related to chimpanzees and bonobos, what this really means is that we have the least amount of variation in our genomes when we compare the human genome to chimpanzees and bonobos as compared to when we, as compared to gorillas or orangutans, right. So we are the least related to Orangutans among the great apes.

So not only can we look at our sort of evolutionary trajectories across the animal kingdom, the great ape specifically, we can also look at archaic humans, right? So there were of course other species of humans around before the current age, before modern humans. So two examples are Neanderthals and Denisovans. So you can see here in this map, right, basically all humans originated from Africa and then migrated outwards towards Eurasia. And you can see that the Neanderthals, right, you can see the circles, you can see their paths, right. Basically their geographical distributions. And then the Neanderthals you can see in the, sorry, the Denisovans you can see within the triangles.

So what you, how do we make, you know, how do we make a map of this sort? What's happened is that people have found artifacts, like really ancient artifacts, things like bones or hair and other kinds of material that still can contains genomes or at least some parts of genomes that could be sequenced and can then allow us to detect whether this DNA material belong to modern day humans, to homo sapiens or to Neanderthals or Denisovans, both of which are other species of essentially archaic humans, right. And not only that, we can also tell by looking at genomes and genome data, we can look at whether these individuals have come about through species mixing, right. So when homo sapiens mixed with Neanderthals or mated with Denisovans, right. So we can actually see that we have actually found remains of ancient humans who actually have been the product of mating between Neanderthals and Denisovans, for example.

And so this allows us to look at migration patterns and really understand our origins in ways that we never could before. We can also try to look at the diversity across modern day humans, right. So, you know, as I just said in my previous slide, that humans, all humans that we know of originated from Africa. And the circles on this map, the blue circle and the red circle, actually represent the genetic diversity that we see in populations in Eurasia and in Africa. So in Africa, it turns out that there is actually, even in present day Africans, there's a lot more genetic diversity than there is in the rest of the world.

And how could this be possible if we all originated from the same population? So, you know, one of the explanations is that potentially the people that migrated out of Africa were only a small portion, right, of the African population. They migrated out, and there was what you call a bottleneck effect. In other words, since it was a sub-sample of the genetic diversity that was already present in Africa that actually migrated out, that population never reached the same amount of genetic diversity that the African population has. So this kind of work has been continued by things like the 1000 Genome Project.

So in the 1000 Genome Project, 1000 genomes are being sequenced of people across the world, right. And so one really sort of interesting outcome of this project, where 184 Africans and 100, sorry, I think I have this number of 186 Africans and 184 Eurasians, the genomes have been sequenced and compared.

And it turns out that there isn't a single locus where we have a 100% difference across the populations, right. And so this already tells us something, right, that our differences, the differences that we see, whether they're the physical or physiological differences that we see across human populations, don't actually arise from any single locus, right, but rather from an interaction of many different kinds of, say, genetic interactions. All right, so not only can we look at evolution, then, across the animal kingdom, the previous phylogenetic tree that I showed you was one with the great apes.

Now we can also look at human evolution, right. So on the slide, you see what I already said before, where humans have diverged from the other apes 5 to 7 million years ago, right. But it was around 800,000 years ago, so not that long ago, that we diverged, and that there was further speciation, right, between Neanderthals and Denisovans around 600,000 years ago. So it's really interesting that we're able to use this molecular scale genomic data to actually trace back our history, right. So genetics has a lot of implications for how we see ourselves in our place in the world. And so we've been, as I said, we've been able to trace migration patterns, right, and we've been able to then trace also, you know, our evolution across time.

And that's what this, this is actually another form of a phylogenetic tree that shows you how Neanderthals and Denisovans, how they evolved and when they evolved, right, and what our common ancestors were. So we went from Homo erectus to Homo heidelbergensis, and then Neanderthals and Denisovans.

Okay, so once we are able to extract genomic data and actually trace back our ancestry, right, so typically, when we think about the word ancestry, we think about our grandparents, our great grandparents, perhaps, you know, five, ten generations back at most. But this is allowing us to look back hundreds of thousands of years, right. So how does this knowledge of human history really inform us? What does it tell us? So there's, there are some ethical questions to be considered. The first is, you know, ethical questions don't always have to be bad, first of all.

But one thing we should ask ourselves is, you know, how does knowing this, does it in any way reinforce any biases that you already have, right? Does it reinforce stereotypes that we may have about certain populations, certain countries, the origin of certain people?

So these are kinds of things that you have to think about very carefully. And the kinds of questions that you ask, right, should also be very carefully chosen. I don't mean ethical questions. I mean, the scientific questions that you ask when you are exploring genetic data, right, have to be really carefully chosen. Because if you ask questions that may lead to further bias in a society, or that may lead to a certain group of people being stereotyped in some way, right. So this may lead to really dramatically different societal structures as well.

So we have to really ask ourselves as we explore genomic data, right, how we want to proceed with this? And this is not a question just for the scientists. It's also for the public at large, who is actually funding most of the research that's being done, right. So the public really should participate in thinking about how we want to explore not only knowledge of human history, but genetic data in general. Other applications in human genetics include things like newborn genetic screening. So some examples of genetic disorders are included on this slide.

So everything from cystic fibrosis to PKU, which you may have heard of, often are related to say, a single nucleotide mutation, right. So a single nucleotide has been changed. And that gives the newborn a very high chance of having a particular disease, right. So what do you do with the data from newborn genetic screening, right? What are the consequences of knowing about a baby's genetic status before birth, right? So these are again questions that parents may have, communities may have, societies may have, right. When do we decide that a pregnancy is worth carrying through and when is the pregnancy not? When should it be terminated? Or should it be terminated based on the knowledge of a baby's genetic status?

So again, questions that come up as we sort of move forward in our ability to understand genetics. I want to bring up one more example. And that is, but before I do that, I have to explain Cytoplasmic inheritance. So generally, when sperm and egg come together to fertilize the former zygote, the cytoplasm of the zygote comes primarily from the maternal gamete, right. So the cytoplasmic inheritance always comes from the mother.

And within the cytoplasm is where the mitochondria are present, right. So that's within the cytoplasm. So, as you know, the mitochondria also have, mitochondria also have their own genomes, right. So they carry genetic material, and this genetic material can be sequenced. And it is entirely possible that a mother may carry mitochondrial disorders, right, DNA disorders. And so this means that if a mother has a mitochondrial DNA disorder, then the child will be guaranteed to have that disorder as well, because the child will inherit all of the mitochondria, or most of the mitochondria from the mother. So what people have done through the process of IVF, in vitro fertilization, they've generated what are called in layman's terms, three parent babies.

And this is actually legal in many countries now, right, where the mother who has a mitochondrial genetic disorder donates her nucleus, I shouldn't use the word donate, the mother, the mother's egg, the nucleus is removed, and then embedded within a donor egg, right. And the donor's nucleus is removed. So what this means is now you have a healthy egg, where the mitochondria, the cytoplasmic inheritance comes from the donor, and the mother's nucleus is still present within the donor cytoplasm. So upon fertilization, the child gets the mitochondrial DNA from the donor, which is I should say a much smaller portion, proportion of the total DNA within the cell, and the nuclear DNA from the mother.

So this is an interesting example of three parent babies and our ability to detect, say, mitochondrial disorders has allowed for this process to come about, right? And actually, it turns out that mitochondrial DNA editing using technologies such as CRISPR have really been emerging lately. And we never know, it may soon be the case that a donor egg is no longer required, donor cytoplasm is no longer required, and we can directly go in and edit mitochondrial DNA.

So again, ethical questions come up, right. If DNA can be edited, what prevents designer babies? So what do I mean by designer babies? Designer babies are, you know, babies where you can actually go in and edit specific genes, right. So suppose I'm not worried about disease anymore, but I'm worried that, you know, parent is worried that their baby should be really tall. So can we go in and edit genes selectively? So the outcome is that their child is really tall.

So is this a question that we want to confront as a society? Is it bad? Is our designer babies bad? How far do we want to manipulate nature? And then of course the question comes up that, well, we are also part of nature, so whatever we do is natural, right? So these are really complex ethical questions and don't have simple answers to them, but there are questions that we should very carefully consider as we move forward again with genetic technology. So finally, I'd like to summarize this lecture. We've really been through a very wide variety of things actually in the past three lectures, but in this particular lecture, we talked about the cost of sequencing, referred to as the idea that as it gets cheaper, there are more and more applications and it's more and more accessible to us.

We talked about SNPs, single nucleotide polymorphisms, and other sources of variation. We talked about how the Hardy-Weinberg equilibrium could be used to predict allele frequencies. We also said that we could use variation to infer evolutionary trajectories, and we looked at examples of the great apes as well as archaic humans and modern day humans. We looked at early human geographical distributions and we said that we were able to do this because of our ability to extract DNA from ancient artifacts, archaeological remains actually.

We also talked about genetic screening. We talked about the possibility of mitochondrial editing after we covered three-parent babies. So you can see that the array of applications from genetic technology is really wide. So a few things that I'd like you to take away from this set of lectures is one, the idea of sequencing, how much more accessible it is becoming, and so that there are really very large scale sort of applications and societal implications for us. I would also like you to go back and think about how genetic technology has been used for us to understand really disease, right? So we very briefly touched on that in the last few lectures, but we will discuss that in a lot more detail in the next lecture where we will talk about specifically how mutations can lead to epilepsy, an example of a disorder that affects many people within the human population. Thank you.