PROFESSOR: So everybody ready to rock and roll today? Or at least roll? OK, if not rock. Welcome back to lecture 20.

On Thursday, we have a special guest appearance just for you from Professor Ron Weiss. He's going to be talking about synthetic biology. You know, on Richard Feynman's blackboard, when he died, was a little statement I've always really enjoyed that was actually encased in a little chalk line. And it said, "What I cannot create, I do not understand." And so synthetic biology is one way to approach questions in the biological sciences by seeing what we can make-- you know, whole organisms, new genomes, rewiring circuitry. So I think you'll find it to be a very interesting discussion on Thursday.

But first, before we talk about synthetic biology, we have the very exciting discussion today on human genetics, which of course concerns all of us. And so we're going to have an exploration today. And you know, as I prepared today's lecture, I wanted to give you the latest and greatest research findings. So we're going to talk today from fundamental techniques to things that are very controversial that have caused fistfights in bars, so I'm hopeful that you'll be as engaged as the people drinking beer are.

So we'll be turning to that at the end of the lecture, but first I wanted to tell you about the broad narrative arc once again that we're going to be talking about today. And we're going to be looking at how to discover human variation. We're all different, about 1 based in every 1,000. And there are two different broad approaches historically people have used. Microarrays for discovering variants, and we'll talk about how those were designed. And we'll talk about how to actually test for the significance of human variation with respect to a particular disease in a case and control study.

And then we'll talk about how to use whole genome read data to detect variation between humans and some of the challenges in that, because it does not make as many assumptions as the microarray studies, and therefore is much, much more complicated to process. And so we're going to take a view into the best practices of processing human read data so you can understand what the state of the art is.

And then we're going to turn to a study showing how we can bring together different threads we've talked about in this subject. In particular, we've talked about the idea that we can use other genomic signals such as histone marks to identify things like regulatory elements. So we're going to talk about how we can take that lens and focus it on the genome to discover particular genomic variants that have been revealed to be very important in a particular disease.

And finally, we'll talk about the idea that what-- the beginning we're going to talk about today is all about correlation. And as all of you go forward in your scientific career, I'm hopeful that you'll always be careful to not confuse association or correlation with causation. You're always respected when you clearly articulate the difference when you're giving a talk saying that this is correlated, but we don't necessarily know it's causative until we do the right set of experiments.

OK, on that note, we'll turn to the computational approaches we're going to talk about. We'll talk about contingency tables and various ways of thinking about them when we discuss how to identify whether or not a particular SNP is associated with a disease using various kinds of tests. And then we talk about read data, and we'll talk about likelihood based tests. How to do things like take a population of individuals and their read data and estimate the genotypic frequencies at a particular locus using that data in toto using EM based techniques. OK? So let us begin then.

Some of the things we're not going to talk about include non-random genotyping failure, methods to correct for population stratification, and structural variants and copy number variations. Point three we'll just briefly touch on, but fundamentally they're just embellishments on the fundamental techniques we're talking about, and so I didn't really want to confuse today's discussion.

Now a Mendelian disorder is a disorder defined by a single gene, and therefore, they're relatively easy to map. And they also tend to be in low frequency in the population because they're selected against, especially the more severe Mendelian disorders. And therefore, they correspond to very rare mutations in the population.

By point of contrast, if you thought about the genetics we discussed last time, if you think about a trait that actually perhaps is influenced by 200 genes and maybe that one of those genes is not necessary or sufficient for a particular disease. As a consequence, it could be a fairly common variant and it's only if you're unlucky enough to get all the other 199 variants will you actually come down with that syndrome. And therefore, you can see that the effect of variation in the human genome is inversely related to its frequency-- that fairly rare variants can have very serious effects, whereas fairly common variants tend to have fewer effects.

And in the first phase of mapping human variation, people thought that common variants were things that had an allelic frequency in the population of 5% or greater. And then to burrow down deeper, the 1000 Genomes Project surveyed a collection of different populations. And therefore, if you thought that a variant was prevalent in the population at frequency of 0.5%, how many people would have in the 1000 Genomes Project roughly? Just make sure among you we're phase-locked here. 0.5%, 1,000 people--

AUDIENCE: 5.

PROFESSOR: 5, great. OK. Good. Now of course, these are three different populations or more, and so it might be that, in fact, that variant's only present in one of the population. So it just might be one or two people that actually have a particular variant. So the idea is that the way that you design SNP chips to detect single nucleotide polymorphisms, otherwise known as "SNPs," is that you do these population-based sequencing studies and you design the array based upon all of the common variants that you find, OK? And therefore, the array gives you a direct readout in terms of the variation, in terms of all of these common variants. That's where we'll start today. And where we'll end today is sequencing based approaches, which make no assumptions whatsoever and just all hell breaks loose. So you'll see what happens, OK?

But I wanted just to reinforce the idea that there are different allelic frequencies of variants and that as we get down to rarer and rarer alleles, right, we have larger effects. But these higher frequency alleles could also have effects even though they're much smaller.

OK, so let's talk about how these variants arise and what they mean in terms of a small little cartoon. So long, long ago, in a world far, far away, a mutation occurred where a base G got mutated to a base A, OK? And this was in the context of a population of individuals-- all happy, smiling individuals because they all actually do not have a disease. And our story goes, what happens is that we had yet another generation of people who are all happy, smiling because they do not have the disease. Right? Yes, I do tell stories at night, too.

And then another mutation occurred. And that mutation caused some subset of those people to get the mutation and for them to get the disease. So the original mutation was not sufficient for people to get this genetic disease. It required a second mutation for them to get the genetic disease, OK? So we got at least two genes involved in it.

OK, the other thing that we know is that, in this particular case, some of the people have the disease that don't have this mutation. And therefore, this mutation is not necessary. It's neither necessary nor sufficient. Still, it is a marker of a gene that increases risk. And that's a fundamental idea, right, that you can increase risk without being necessary or sufficient to cause a particular genetic disorder. OK? And so you get this association then between genotype and phenotype, and that's what we're going to go looking for right now. We're on the hunt. We're going to go looking for this relationship.

So in certain older individuals-- hopefully, not myself in the future-- what happens is that your maculus, which is the center of your eye, degenerates as shown here. And you get this unfortunate property where you can't actually see in the center of your field of vision. It's called age-related macular degeneration. So to look for the causes of this, which is known to be genetically related, the authors of the study collected a collection-- a cohort, as it's called-- of these European-descent individuals, all who are at least 60 years old, to study the genetic foundations for this disorder. And so they found 934 controls that were unaffected by age-related macular degeneration and 1,238 cases, and they genotyped them all using arrays.

Now the question is, are any of the identified SNPs on the array related to this particular disorder? So I'll give you the answer first and then we'll talk about a couple of different ways of thinking about this data, OK? So here's the answer. Here's a particular SNP, rs1061170. There are the individuals with AMD and the controls. And what you're looking at up here, these numbers are the allelic counts, all right? So each person has how many alleles? Two, right? That's double the number of individuals. And the question is, are the C and T alleles associated with the cases or controls significantly?

And so you can compute a Chi-square metric on this so-called contingency table. And one of the things about contingency tables that I think is important to point out is that you hear about marginal probabilities, right? And people probably know that originally derived from the idea of these margins along the side of a contingency table, right? If you think about the marginal probability of somebody having a C allele, regardless of whether a case or control, it would be 2,192 over 4,344, right?

So the formula for computing the Chi-square statistic is shown here. It's this sort of scary-looking polynomial. And the number of degrees of freedom is 1. It's the number of rows minus 1 times the number of columns minus 1. And the P-value we get is indeed quite small-- 10 to the minus 62. Therefore, the chance this happened at random-- even with multiple hypothesis correction, given that we're testing a million SNPs-- is indeed very, very low. This looks like a winner. Looks like we've got a SNP that is associated with this particular disease.

Now just to remind you about Chi-square statistics-- I'm sure people have seen this before-- the usual formulation is that you compute this Chi-square polynomial on the right-hand side, which is the observed number of something minus the expected number of something squared over the expected number or something, right? And you sum it up over all the different cases. And you can see that the expected number of As is given by the little formula on the left. Suffice to say, if you expand that formula and manipulate it, you get the equation we had on the previous slide. So it's still that fuzzy, friendly Chi-square formula you always knew, just in a different form, OK?

Now is there another way to think about computing the likelihood of seeing data in a contingency table at random, right? Because we're always asking, could this just be random? I mean, could this have occurred by chance, that we see the data arranged in this particular form? Well, we've another convenient way of thinking about this, which is we could do Fisher's Exact Test, which is very related to the idea of the hypergeometric test that we've talked about before, right? What are the chances we would see exactly this arrangement?

Well, we would need to have, out of a plus b C alleles, we'd have 8 of them be cases, which is the first term there in that equation. And of the T alleles, we need to have c of them out of c plus d be there. And then we need to have a plus b plus c plus d choose a plus c-- that's the total number of chances of seeing things. So this is the probability of the arrangement in the table in this particular form. Now people-- I'll let you digest that for one second before I go on.

So this is the number of ways on the numerator of arranging things to get the table the way that we see it over the total number or ways of arranging the table, keeping the marginal totals the same. Is that clear? So this is the probability, the exact probability, of seeing the table in this configuration. And then what you do is you take that probability and all of the probabilities for all the more extreme values, say, of a. And you sum them all up and that gives you the probability of a null hypothesis. So this is another way to approach looking at the chance a particular contingency table set of values would occur at random.

So if people talk about Fisher's Exact Test-- you know, tonight at that cocktail party. "Oh yeah, I know about that. Yeah, it's like the hypergeometric. It's no big deal." You know? Right.

All right. So now let us suppose that we do an association test and you do the following design. You say, well, I've got all my cases. They're all at Mass General and I want to genotype them all. And Mass General is the best place for this particular disease, so I'm going to go up there. I need some controls but I'm running out of budgetary money, so I'm going to do all my controls in China, right? Because I know it's going to be less expensive there to genotype them.

And furthermore-- as a little aside, I was once meeting with this guy who is like one of the ministers of research in China. He came to my office. I said, so what do you do in China? And he said, well, I guess the best way to describe it is that I'm in charge of the equivalent of the NSF, DARPA, and the NIH. I said, oh. I said, would like to meet the president? Because I'd be happy to call MIT's president. I'm sure they'd be happy to meet with you. He said, no. He said, I like going direct. So, at any rate, I told him I was working in stem cell research. He said, you know, one thing I can say about China-- in China, stem cells, ethics, no problem.

[LAUGHTER]

At any rate, so you go to China to do your controls, OK? And why is that a bad experimental design? Can anybody tell me? You do all your cases here, you do your controls in China. Yes?

AUDIENCE: Because the SNPs in China are not necessarily the same.

PROFESSOR: Yes. The Chinese population's going to have a different set of SNPs, right, because it's been a contained population. So you're going to pick up all these SNPs that you think are going to be related to the disease that are simply a consequence of population stratification, right? So what you need to do is to control for that. And the way you do that is you pick a bunch of control SNPs that you think are unrelated to a disease and you do a Chi-square test on those to make sure that they're not significant, right? And methodologies for controlling for population stratification by picking apart your individuals and re-clustering them is something that is a topic of current research.

And finally, the good news about age-related macular degeneration is that there are three genes with five common variants that explain 50% of the risk. And so it has been viewed as sort of a very successful study of a polygenic-- that is, multiple gene-- genetic disorder that has been dissected using this methodology. And with these genes, now people can go after them and see if they can come up appropriate therapeutics.

Now using the same idea-- cases, controls-- you look at each SNP individually and you query it for significance based upon a null model. You can take a wide range of common diseases and ask whether or not you can detect any genetic elements that might influence risk. And so here are a set of different diseases, starting with bipolar disorder at the top, going down to type 2 diabetes at the bottom. This is a so-called Manhattan plot, because you see the buildings along the plot, right? And when there are skyscrapers, you go, whoa, that could be a problem, all right? And so this is the style-- you see, this came out in 2007-- of research that attempts to do genome wide scans for loci that are related to particular diseases.

OK, now, I'd like to go on and talk about other ways that these studies can be influenced, which is the idea of linkage disequilibrium. So for example, let us say that I have a particular individual who's going to produce a gamete and the gamete's going to be haploid, right? And it's going have one allele from one of these two chromosomes and one allele from one of these two chromosomes. We've talked about this before last time. And there are four possibilities-- AB, aB, Ab, and ab, OK? And if this was a coin flip, then each of these genotypes for this gamete would be identical, right?

But let us suppose that I tell you that there are only two that result-- this one, AB, and the small one, ab, OK? If you look at this, you say, aha, these two things are linked and they're very closely linked. And so, if they're always inherited together, we might think that the distance between them on the genome is small. So in the human genome, what's the average distance between crossover events during a meiotic event? Does anybody know, roughly speaking, how many bases? Hm?

AUDIENCE: A megabase?

PROFESSOR: A megabase. Little low. How many centimorgans long is the human genome? All right? Anybody know? 3,000? 4,000? Something like that? So maybe 50 to 100 megabases between crossover events, OK? So if these markers are very closely organized along the genome, the likelihood of a crossover is very small. And therefore, they're going to be in very high LD, right? And a way to measure that is with the following formula, which is that if you link the two locuses-- we have L1 and L2 here. And now we're talking about the population instead of a particular individual. If the likelihood of the capital allele A is piece of A and the probability of the big B allele is piece of B, then if they were completely unlinked, then the likelihood of inheriting both of them together with would be piece of A times piece of B, showing independence.

However if they aren't independent, we can come up with a single value D which allows us to quantify the amount of disequilibrium between those two alleles. And the formula for D is given on the slide. And furthermore, if it's more convenient for you to think about in terms of r-squared correlation, we can define the r-squared correlation as D squared over PA, QA, PB, QB, as shown in the lower left hand part of this slide, OK? This is simply a way of describing how skewed the probabilities are from being independent for inheriting these two different loci in a population. Are there any questions at all about that, the details of that? OK.

So just to give you an example, if you look at chromosome 22, the physical distance on the bottom is in kilobases, so that's from 0 to 1 megabase on the bottom. And you look at the r-squared values, you can see things that are quite physically close, as we suggested earlier, have a high r-squared value. But there are still some things that are pretty far away that have surprisingly high r-squared values. There are recombination hot spots in the genome. And it's, once again, a topic of current research trying to figure out how the genome recombines and recombination is targeted. But suffice it to say, as you can see, it's not uniform.

Now what happens as a consequence of this is that you get regions of the genome where things stick together, right? They're all drinking buddies, right? They all hang out together, OK? But here's what I'm going to ask you-- how much of your genome came from your dad? Half. How much came from your dad's dad?

AUDIENCE: A quarter.

PROFESSOR: And from your dad's dad's dad?

AUDIENCE: An eighth.

PROFESSOR: An eighth, OK? So the amount of your genome going back up the family tree is falling off exponentially up a particular path, right? So if you think about groups of things that came together from your great, great grandfather or his great, great grandfather, right, the further back you go, the less and less contribution they're going to have to your genome. And so the blocks are going to be smaller-- that is, the amount of information you're getting from way back up the tree.

So if you think about this, the question of what blocks of things are inherited together is not something that you can write an equation for. It's something that you study in a population. You go out and you ask, what things do we observe coming together? And generally, larger blocks of things are inherited together, occur in more recent generations, because there's less dilution, right? Whereas if you go way back in time-- not quite to where the dinosaurs roamed the land, but you get the idea-- the blocks are, fact, quite small.

And so, the HapMap project went about looking at blocks and how they were inherited in the genome. And what suffices to know is that they found blocks-- here you can see three different blocks. And these are called haplotype blocks and the things that are colored red are high r-squared values between different genetic markers. And we talked earlier about how to compute that r-squared value. So those blocks typically are inherited together. Yes?

AUDIENCE: Are these blocks like fuzzy boundaries?

PROFESSOR: No. Well, remember in this particular example, we're only querying at specified markers which are not necessarily at regular intervals along the genome. So in this case, the blocks don't have fuzzy boundaries. As we get into sequencing-based approaches, they could have fuzzier boundaries. But haplotype blocks are typically thought to be discrete blocks that are inherited, OK? Good question. Any other questions? OK.

So I want to impress upon the idea that this is empirical, right? There's no magic here in terms of fundamental theory about what things should be haplotype blocks. It's simply that you look at a population and you're look at what markers are drinking buddies and those make haplotype blocks and you empirically categorize and catalog them-- which can be very helpful, as you'll see.

And thus, when we think about genetic studies, when we think about the length of shared segments, if you're thinking about studying a family, like a trio-- a trio is a mom, a dad, and a child, right? They're going to share a lot of genetic information, and so the haplotype blocks that are shared amongst those three individuals are going to be very large indeed. Whereas if you go back generations, the blocks-- like the second cousins or things like that-- the blocks get smaller. So the x-axis on this plot is the median length of a shared segment. And as an association study, which is taking random people out of the population, has very small shared blocks indeed, OK? And so the techniques that we're talking about today are applicable almost in any range, but they're particularly useful where you can't depend upon the fact that you're sharing a lot of information along the genome proximal to where the marker is that's associated with a particular disease.

Now the other thing that is true is that we should note that the fact that markers have this LD associated with them means that it may be that a particular marker is bang on-- what's called a causative SNP. Or something that, for example, sits in the middle of a gene causing a missense mutation. Or it sits right in the middle of a protein binding site, causing the factor not to bind anymore. But also it could be something that is actually a little bit away, but is highly correlated to the causative SNP. So just keep in mind that when you have an association and you're looking at a SNP, it may not be the causative SNP. It might be just linked to the causative SNP. And sometimes these things are called proxy SNPs.

OK, so we've talked about the idea of SNPs and discovering them. Let me ask you one more question about where these SNPs reside and see if you could help me out. OK, this is a really important gene, OK? Call it RIG for short, OK? Now let us suppose that you know that there are some mutations here. And my question for you is, does it matter whether or not the two mutations look like this or the mutations look like this, in your opinion?

That is, both mutations occur in one copy or on one chromosome of the gene, whereas in the other case, we see two different SNPs that are different than reference, but they're referring in both mom and dad alleles. Is there a difference between those two cases? Yeah?

AUDIENCE: In terms of the phenotype displayed?

PROFESSOR: In terms of the phenotype, sorry.

AUDIENCE: It depends.

PROFESSOR: It depends?

AUDIENCE: Yes.

PROFESSOR: OK.

AUDIENCE: So if it causes a recessive mutation, then no, because other genes will be able to rescue it. But if it's dominant, then it'll still--

PROFESSOR: It's actually the other way around. If it's recessive, this does matter.

AUDIENCE: Oh, I see.

PROFESSOR: Because in this case, with the purple ones, you still have one good copy of the gene, right? However, with the green ones, it's possible that you have ablated both good copies of the gene, this really important gene, and therefore, you're going to get a higher risk of having a genetic disorder. So when you're scanning down the genome then-- you know, we've been asking where are there differences from reference or from between cases and controls down the genome, but we haven't asked whether or not they're on mom or dad, right? We're just asking, are they there?

But it turns out that for questions like this, we have to know whether or not the mutation occurred in only in mom's chromosome or in both chromosomes in this particular neighborhood, all right? This is called phasing of the variants. Phasing means placing the variants on a particular chromosome. And then, by phasing the variants, you can figure out some of the possible phenotypic consequences of them. Because if they're not they phased, in this case, it's going to be much less clear what's going on.

So then the question becomes, how do we phase variants, right? So phasing assigns alleles to the parental chromosomes. And so, the set of alleles along a chromosomes is a haplotype. We've talked about the idea of haplotypes. So imagine one way to phase is that if I tell you, by magic, in this population you're looking at, here are all the haplotypes and these are the only ones that exist. You look in your haplotypes and you go, aha, this haplotype exists but this one does not, right? That this is a haplotype-- this two purples together is a haplotype, and this green without one is another haplotype. So you see which haplotypes exist-- that is, what patterns of inheritance of alleles along a chromosome you can detect. And using the established empirical haplotypes, you can phase the variants, OK?

Now the other way to phase the variants is much, much simpler and much better, right? The other way to phase variants is you just have a single read that covers the entire thing, right? And then the read, it will be manifest, right? The read will cover the entire region and then you see the two mutations in that single region of the genome. The problem we're up against is that most of our reads are quite short and we're reassembling our genotypes from a shattered genome. If the genome wasn't shattered, then we wouldn't have this problem.

So everybody's working to fix this. Illumina has sort of a cute trick for fixing this. And PacBio, which is another sequence instrument manufacturer, can produce reads that are tens of thousands of bases long, which allows you to directly phase the variants from the reads. But if somebody, once again, comes up to you at the cocktail party tonight and says, you know, I've never understood why you have to phase variants. You know, this is a popular question I get all the time.

[LAUGHTER]

You can tell them, hey, you know, you have to know whether or not mom and dad both have got big problems or just all the problems are with mom, OK? Or dad. So you can put their mind at ease as you're finishing off the hors d'oeuvres.

OK, so I wanted just to tell you about phasing variants because we're about to go on to the next part of our discussion today. We are leaving the very clean and pristine world of very defined SNPs, defined by microarrays, into the wild and woolly world of sequencing data, right? Which is all bets are off. It's all raw sequencing data and you have to make sense of it-- hundreds of millions of sequencing reads.

So today's lecture is drawn from a couple of different sources. I've posted some of them on the internet. There's a very nice article by Heng Li on the underlying mathematics of SNP calling and variation which I've posted. In addition, some of the material today is taken from the Genome Analysis Toolkit, which is a set of tools over at the Broad, and we'll be talking about that during today's lecture.

The best possible case when you're looking at sequence data today is that you get something like this, all right? Which is that you have a collection of reads for one individual and you align them to the genome and then you see that some of the reads have a C at a particular base position and other of the reads have a T at that base position. And so it's a very clean call, right? You have a CT heterozygote at that position-- that is, whatever person that is is a heterozygote there. You can see the reference genome at the very bottom, right? So it's very difficult for you to read in the back, but C is the reference allele.

And the way that the IGV viewer works is that it shows non-reference alleles in color, so all those are the T alleles you see there in red, OK? And it's very, very beautiful, right? I mean you can tell exactly what's going on. Now we don't know whether or not the C or the T allele are a mom and dad respectively, right? We don't know which of the chromosomes they're on, but suffice to say, it's very clean.

And the way that all of this starts, of course, is with a BAM file. You guys have seen BAM files before. I'm not going to belabor this. There's a definition here and you can add extra annotations on BAM files. But the other thing I wanted to point out is that you know that BAM files include quality scores. So we'll be using those quality scores in our discussion.

The output of all this typically is something called a variant call file or a VCF file. And just so you are not completely scared by these files, I want to describe just a little bit about their structure. So there's a header at the top telling you what you actually did. And then chromosome 20 at this base location has this SNP. The reference allele is G, the alternative allele is A. This is some of the statistics as described by this header information, like DP is the read depth. And this tells you the status of a trio that you processed.

So this is the allele number for one of the chromosomes, which is 0, which is a G. The other one's a G, and so forth. And then this data right here is GT, GQ, GP, which are defined up here. So you have one person, second person, third person, along with which one of the alleles, 0 or 1, they have on each of their chromosomes, OK?

So this is the output. You put in raw reads and what you get out is a VCF file that for bases along the genome calls variance, OK? So that's all there is to it, right? You take in your read data. You take your genome, throw it through the sequencer. You take the BAM file, you call the variants, and then you make your medical diagnosis, right?

So what we're going to talk about is what goes on in the middle there, that little step- - how do you actually call the variants? And you might say, gee, that does not seem too hard. I mean, I looked at the slide you showed me with the CT heterozygote. That looked beautiful, right? I mean, that was just gorgeous. I mean, these sequencers are so great and do so many reads, what can be hard about this, after all? You know, it's a quarter to 2:00, time to go home. Not quite, OK? Not quite.

The reason is actual data looks like this. So these are all reads aligned to the genome and, as I told you before, all the colors are non-reference bases. And so you can see that the reads that come out of an individual are very messy indeed. And so we need to deal with those in a principled way. We need to make good, probabilistic assessments of whether or not there's a variant at a particular base. And I'm not going to belabor all the steps of the Genome Analysis Toolkit, suffice to say, here is a flow chart of all the steps that go through it. First, you map your reads. You recalibrate the scores. You compress the read set and then you have read sets for n different individuals. And then you jointly call the variants and then you improve upon the variants and then you evaluate, OK?

So I'll touch upon some of the aspects of this pipeline, the ones that I think are most relevant, so that you can appreciate some of the complexity in dealing with this. Let me begin with the following question. Let us suppose that you have a reference genome here, indicated by this line, and you align a read to it. And then there's some base errors that are non-reference, so it's variants calls down at this end of the read, OK? So this is the five prime, three prime. And then you align a read from the opposite strand and you have some variant calls on the opposite end of the read like this.

And you'll say to yourself, what could be going on here, you know? Why is it that they're not concordant, right, when they're mapped, but they're in the same region of the genome? And then you think to yourself, well, what happens if I map this successfully here correctly to the reference genome and this correctly to the reference genome here, but this individual actually had a chromosome that had a deletion right here, OK? Then what would happen would be that all these reads down here are going to be misaligned, all of these bases are going to be misaligned with the reference. And so you're going to get variant calls. And these bases will be also misaligned with the reference, you'll get variant calls.

So deletions in an individual can cause things to be mapped but you get variant calls at the end. And so that is shown here, where you have reads that are being mapped and you have variant calls at the end of the reads. And it's also a little suspicious because in the middle here is a seven-base pair homopolymer which is all T's. And as we know, sequencers are notoriously bad at correctly reading homopolymers. So if you then correct things, you discover that some fraction of the reads actually have one of the T's missing and all the variants that were present before go away. So this is a process of INDEL adjustment when you are mapping the region looking for variants.

Now this does not occur we're talking about SNP microarrays. So this is a problem that's unique to the fact that we're making many fewer assumptions when we map reads to the genome de novo.

The second and another very important step that they're very proud of is essentially-- I guess how to put this politely-- finding out that manufacturers of sequencing instruments, as you know, for every base that they give you, they give you an estimate of the probability that the base is correct-- or it's wrong, actually. A so-called Phred score-- we've talked about that before. And as you would imagine, manufacturers' instruments are sometimes optimistic, to say the least, about the quality of their scores. And so they did a survey of a whole bunch of instruments and they plotted the reported score against the actual score, OK?

And then they have a whole step in their pipeline to adjust the scores, whether you be a Solexa GA instrument, a 454 instrument, a SOLiD instrument, a HiSeq instrument, or what have you. And there is a way to adjust the score based upon the raw score and also how far down the read you are, as the second line shows. The second line is a function of score correction versus how far down the read or number of cycles you have gone. And the bottom is adjustments for dinucleotides, because some instruments are worse a certain dinucleotides than others.

As you can see, they're very proud of the upper left hand part. This is one of the major methodological advances of 1000 Genome Project, figuring out how to recalibrate quality scores for instruments. Why is this so important? The reason it's important is that the estimate of the veracity of bases figures centrally in determining whether or not a variant is real or not. So you need to have as best an estimate as you possibly can of whether or not a base coming out of the sequencer is correct, OK?

Now if you're doing lots of sequencing of either individuals or exomes-- I should talk about exome sequencing for a moment. Up until recently, it has not really been practical to do whole genome sequencing of individuals. That's why these SNP arrays were originally invented. Instead, people sequenced the expressed part of the genome, right? All of the genes. And they can do this by capture, right? They go fishing. They create fishing poles out of the genes that they care about and they pull out the sequences of those genes and they sequence them. So you're looking at a subset of the genome, but it's an important part.

Nonetheless, whether or not you do exome sequencing or you do sequencing of the entire genome, you have a lot of reads. And so the reads that you care about are the reads that are different from reference. And so you can reduce the representation of their BAM file simply by throwing all the reads on the floor that don't matter, right? And so here's an example of the original BAM file and all the reads, and what you do is you just trim it to only the variable regions, right? And so you are stripping information around the variant regions out of the BAM file and things greatly compress and the downstream processing becomes much more efficient. OK.

Now let's turn to the methodological approaches once we have gotten the data in as good a form as we possibly can get it, we have the best quality scores that we can possibly come up with, and for every base position, we have an indication of how many reads say that the base is this and how many reads say the base is that, OK?

So we have these raw read counts of the different allelic forms. And returning to this, we now can go back and we can attempt to take these reads and determine what the underlying genotypes are for an individual.

Now I want to be clear about the difference between a genotype for an individual and an allelic spectrum for a population. A genotype for an individual thinks about both of the alleles that that individual has, and they can be phased or unphased, right? If they're phased, it means you know which allele belongs to mom and which allele belongs to dad, so to speak, right? If they're unphased, you simply know the number of reference alleles that you have in that individual. Typically, people think about there being a reference allele and an alternative allele, which means that a genotype can be expressed as 0, 1, or 2, which is the number of reference alleles present at a particular base if it's unphased, right? 0, 1, or 2-- 0 mean there are no reference alleles there, 1 meaning that it's a heterozygote, 2 mean there are two reference alleles in that individual. OK?

So there are different ways of representing genotype, but once again, it represents the different allelic forms of the two chromosomes. And whatever form you choose, the probability over those genotypes has to sum to 1. You can think about the genotype ranging over all the possible bases from mom and from dad or over 0, 1, and 2. Doesn't really matter, depending upon which way you want to simplify the problem.

And what we would like to do is, for a given population-- let's say cases or controls-- we'd like to compute the probability over the genotypes with high veracity, OK? So in order to do that, we'll start by taking all the reads for each one of the individuals in a population, OK? And we're going to compute the genotype likelihoods for each individual. So let's talk about how to do that.

Now everything I've written on the board is on the next slide. The problem is, if I put it on the slide, it will flash in front of you and you'll go, yes, I understand that, I think. This way, I'll put it on the board first and you'll look at it and you'll say, hm, maybe I don't understand that, I think. And then you'll ask any questions and we can look at the slide in a moment, OK? But here's the fundamental idea, all right? At a given base in the genome, the probability of the reads that we see based upon the genotype that we think is there can be expressed in the following form, which is that we take the product over all the reads that we see, OK? And the genotype is going to be a composition of the base we get from mom and the base that we get from dad, or it could simply be 0, 1, and 2. We'll put that aside for a moment.

So what's the chance that we inherited something from a particular base from mom? It's this base over a particular read. What's the chance a particular read came from mom's chromosome? That's one half times the probability of the data given the base that we see. And once again, since it could be a coin flip whether the read came from mom or dad's chromosome, divide it by 2-- the probability of the data that we see with dad's version of that particular base.

So once again, for all the reads we're going to compute the probability of the read set that we see given a particular hypothesized genotype by looking at what's the likelihood or the probability of all those reads. And for each read, we don't know if it came from mom or from dad. But in any event, we're going to compute the probability on the next blackboard, this bit right here. OK? Yes?

AUDIENCE: So if you assume that mom and dad have a different phase at a particular base, couldn't that possibly skew the probability of getting a read from mom's chromosome or dad's chromosome?

PROFESSOR: A different phase?

AUDIENCE: So the composition of bases affect what you get. I think certain base compositions are more likely to be sequenced, for example.

PROFESSOR: Yes.

AUDIENCE: Could that bias--

PROFESSOR: Yes. In fact, that's why on, I think, the third slide, I said non-random genotyping error was being excluded.

AUDIENCE: Oh.

PROFESSOR: Right? From our discussion today? But you're right that it might be that certain sequences are more difficult to see, but we're going to exclude that for the time being. OK?

So this is the probability of the reads that we see given a hypothesized genotype. And I'll just show you, that's very simple. That we have a read, let's call the read D sub j, and we have the base that we think we should see. And if the base is correct, then the probability that that's correct is 1 minus the error, right, that the machine reported. And if it isn't correct, the probability is just the error at that base that the machine reported. So we're using the error statistics from the machine and if it matches what we expect, it's 1 minus the error. And if it doesn't match, it's just simply going to be the error that's reported, OK?

So this is the probability of seeing a particular read given a hypothesized base that should be there. Here's how we use that, looking at all the possible bases that could be there given a hypothesized genotype. Remember, this genotype is only going to be one pair. It's only going to be AA or TT or what have you, right? So it's going to be one pair. So we're going to be testing for either one or two bases being present.

And finally, we want to compute the posterior of the genotype given the data we have observed, OK? So we want to compute what's the probability of the genotype given the reads that we have in our hand. This is really important. That is what that genotype likelihood is up there. It's the probability of the read set given the genotype times the probability of the genotype-- this is a prior-- over the probability of the data. This is simply Bayes' Rule. So with this, for an individual now, we can compute the posterior of the genotype given the read set. Very simple concept.

So in another form, you can see the same thing here, which is the Bayesian model. And we've talked about this haploid likelihood function, which was on the blackboard I showed you. And we're assuming all the reads are independent and that they're going to come equally from mom and dad, more or less, et cetera, OK? And the haploid likelihood function, once again, just is using the error statistics for the machine. So I'm asking if the machine says this is an A and I think it's an A, then the probability that that's correct is 1 minus the error of the machine. If the two are not in agreement, it's simply the error in the machine that I'm using.

So this allows me now to give a posterior probability of a genotype given a whole bunch of reads. And the one part that we haven't discussed is this prior, which is how do we establish what we think is going on in the population and how do we set that? So if you look back at the slide again, you can see that we have these individuals and there's this magic step on the right-hand side, which is that somehow we're going to compute a joint estimate across all the samples to come up with an estimate of what the genotypes are in a particular SNP position.

And the way that we can do that is with an iterative EM procedure-- looks like this-- so we can estimate the probability of the population genotype iteratively using this equation until convergence. And there are various tricks. As you'll see if you want to delve further into this in the paper I posted, there are ways to deal with some of the numerical issues and do allele count frequencies and so forth. But fundamentally, in a population we're going to estimate a probability over the genotypes for a particular position.

And just to keep it simple, you can think about the genotypes being 0, 1, or 2-- 0, no reference alleles present; 1, one reference allele present to that site; 2, two reference alleles present to that site, OK? So we get a probability of each one of those states for that population. Any questions at all about that? The details or anything at all? People get the general idea that what we're just doing is we're taking a bunch of reads at a particular position for an individual and computing the posterior probability of a genotype seeing all of those reads and then, when we think about the entire population-- say either the cases or the controls-- we're computing the probability over the genotypes within that population using this kind of iterative procedure.

OK, so going on then, if we go back to our 0, 1, 2 kind of genotype representation, we can marginalize psi, which is the probability of the reference allele being in the population, and 1 minus psi being the probability of the non-reference allele, where the capital alleles are reference and the little ones are non-reference. And then we could also for epsilon 0, epsilon 1, And Epsilon 2, those are the probabilities of the various allelic forms, the various genotypes. And actually, I think epsilon 0 should be little A, little A. Must have got the two of them flipped, but it's not really that important.

OK, so what do we know about a population? Who's heard about Hardy-Weinberg before? Hardy-Weinberg equilibrium? OK. So Hardy-Weinberg equilibrium says that, for example, in a population, if the allelic frequency of the reference allele is psi, right, what's the chance that an individual should be AA, big A, big A, reference, reference? In a population? Pardon? I think I heard it. Psi squared, right? We're going to assume diploid organisms, we're going to assume random mating, we're going to assume no selection, we're going to assume no bottlenecks, and so forth, right? That over time, the population will come to its equilibrium in the exchange of alleles.

However, if there's strong selection or if part of the population gets up and moves to a different continent or something of that sort, you can get out of equilibrium. And so one question, whenever you're doing a genetic study like this-- is your population in equilibrium or not? And we have a direct way for testing for that because we're actually estimating the genotypes, right?

So what we can do is this test. We can do a log likelihood test directly, right? And we can compare the probability of the observed genotypes-- these are E1 and E2, where the number indicates the number of reference copies-- over the probability of the genotypes being composed directly from the frequency of the reference allele. And this will tell us whether or not these are concordant or not. And if the Chi-square value is large enough, we're going to say that this divergence couldn't have occurred at random and therefore the population is not in equilibrium.

And you might say, well, gee, why do I really care if it's in equilibrium or not? I mean, you know, what relevance does that have to me when I'm doing my test?

Well, here's the issue. The issue is this-- you're going to be testing whether or not genotypes are different between a case and a control population, let's say, OK? And you have a couple different tests you can do. The first test is the test on the top and the second test is the test on the bottom. Let's look at the test on the bottom for a moment, OK?

The test in the bottom is saying you consider the likelihood of the data in group one and group two multiplied together over the probability of the data with the groups combined and you ask whether or not the increased likelihood-- you're willing to pay for that given the two degrees of freedom that model implies, right? Because you have to have two additional degrees of freedom to pay for that in the bottom.

On the other hand, in the top, you only have one degree of additional freedom to pay for the difference in simply the reference allele frequency. And so these are two different metrics you can use to test for associations for a particular SNP in two different case and control populations. The problem comes is that if the population is in equilibrium, then the bottom has too many degrees of freedom, right? Because the bottom, in some sense, can be computed directly from the top. You can compute the epsilons directly from the size if it's in equilibrium. So you need to know whether or not you're in equilibrium or not to figure out what kind of test to use to see whether or not a particular SNP is significant.

OK, so just a brief review where we've come to at this point. I've handed you a basket of reads. We're focusing on a particular location in the genome. For an individual, we can look at that basket of reads and compute a posterior probability of the genotype at that location. We then asked if we take all of the individuals in a given interesting population, like the cases, we could compute the posterior or the probability of the genotype over all those cases. We then can take the cases and the controls and test for associations using this likelihood ratio, OK? So that is the way to go at the question of SNPs. And I'll pause here and see if there any other questions about this.

OK, now there are lots of other ways of approaching structural variation. I said I would touch upon it. There is another method which is-- we've not been here assigning what haplotypes to things or paying attention to which chromosome or mom or dad a particularly allele came from. But suffice to say-- I'll let you read this slide at your leisure-- the key thing is that, imagine you have a bunch of reads. What you can do in a particular area there's going to be a variant is you can do local assembly of the reads. We want to do local assembly because local assembly handles general cases of structural variation.

And if you then take the most likely cases of the local assembly supported by reads, you have the different possible haplotypes or collections of bases along the genome from the assembly. And we've already talked about, earlier in class, how to take a given sequence of bases and estimate the probability of the divergence from another string. So you can estimate the divergence of each one of those assemblies from the reference and compute likelihoods, like we did before for single bases, although it's somewhat more complex.

And so another way to approach this is, instead of asking about individual bases and looking at the likelihood of individual bases, you can look at doing localized assembly of the reads that handle structural variation. And if you do that, what happens is that you can recover INDEL issues that appear to call SNP variants that actually are actually induced by insertions and deletions in one of the chromosomes. So as you can see, as things get more sophisticated in the analysis of human genome data, one needs a variety of techniques, including local assembly, to be able to recover what's going on. Because of course, the reference genome is only an approximate idea of what's there and is being used as a scaffold.

So we've already talked about the idea of phasing, and we talked about why phasing is important, especially when we're trying to recover whether or not you have a loss of function event, right? And so the phasing part of genomics right now is highly heuristic, relies partly upon empirical data from things like the HapMap Project and is outside the scope of what we're going to talk about, because we could talk about phasing for an entire lecture. But suffice it to say, it's important.

And if you look at what actually goes on, here's a trio, mom, dad, and the daughter. You can see down at the bottom, you see mom's reads. And you see dad actually has two haplotypes. He got one haplotype from one of his parents and the other haplotype from both of his parents. And then the daughter actually has the haplotype number one from dad and no haplotype number one from mom. So you can see how these blocks of mutations are inherited through the generation in this form.

And finally, if you look at a VCF file, if you see a vertical bar, that's telling you that the chromosomal origin is fixed. So it tells you, for example, in this case, the one where the arrow's pointed to is that variant number one, which is the alternative variant T came from mom and a T came from dad. So VCF, when it has slashes between the two different alleles of a genotype is unphased, but you can actually have a phased VCF version. And so there's a whole phase in the GATK tool kit that does phasing.

And finally, we get through all this, the question is how important are the variants you discover? And so there's a variant analysis phase and it will go through and annotate all the variants. For example, this is a splice site acceptor variant which is in a protein coating gene and it's thought to be important. And so at the end of this pipeline, what's going to happen is you're going to be spitting out not only the variants but, for some of them, what their annotated function is.

OK, so that ends the part of our lecture where we're talking about how to process raw read data. And I encourage you to look at the GATK and other websites to actually look at the state of the art. But as you can see, it's an evolving discipline that has a kernel of principal probabilistic analysis in part of its core surrounded by a whole lot of baling wire, right, to hold it all together. And you get the reads marshalled, organized, compressed, aligned properly, phased, and so forth, OK?

Let's talk now about how to prioritize variants. And I just wanted to show you a very beautiful result that is represented by this paper, which came out very recently. And here is a portion of the genome. And we're looking here at-- you see here-- a pancreatic disorder. And there is an enhancer on the right-hand side of the screen where that little red square is.

And recall that we said that certain histone marks were present typically over active enhancers. And so you can see the H3K4 mono-methyl mark being present over that little red box as well as the binding of two pancreatic regulators, FoxA2 and Pdx1. And within that enhancer you can see there are a whole lot of variants being called. Well, not a whole lot. There are actually five different variants being called. And in addition to those five specific SNP variants, there's also another variation that was observed in the population of a 7.6-kb deletion right around that enhancer. And if you look at the inheritance of those mutations with disease prevalence, you can see that there's a very marked correlation. That when you inherit these variants, you get this Mendelian disorder. So it's a single gene disorder really.

And in order to confirm this, what the authors of this study did was they took that enhancer and they looked at all the different SNPs and they mutated it. And in the upper left-hand corner, you can see that the five little asterisks note the decrease in activity of that enhancer when they mutated the bases indicated at SNP positions. The three C's on the panel on the right with the y-axis being relative interaction shows that that enhancer is interacting with the promoter of that gene. And they further elucidated the motifs that were being interacted with and the little arrows point to where the mutations in those motifs occur. So they've gone from an association with that particular disease to the actual functional characterization of what's going on.

I'll leave you with a final thought to think about. This is the thing that caused the brawls in the bars I told you about at the beginning of today's lecture. Let's suppose you look at identical twins, monozygotic twins, and you ask the following question. You say, OK, in monozygotic twins, the prevalence of a particular disease is a 30% correlative. That is, if one individual in a twin has the disease, there's a 30% chance that the other twin's going to get it. Now there are two possibilities here, which is that for all pairs of twins, there's a very low risk that you're going to get it, or that there are some subset of genotypes where if one twin has it, the other one's always going to get it.

So the author of the study actually looked at a block of identical twin data and asked two questions. The first was if you look at the percentage of cases that would test positive, making no assumptions about genotype, using twin data, you can see the percentage of people that actually have a disease that will test positive is actually fairly low for a wide variety of diseases. All these diseases were studied in the context of identical twins. And this suggested to them that, in fact, personal genomic sequencing might not be as predictive as one might like. That is, if you have a disease, for over half of these, the test will not be positive. And furthermore, if you test negative for the disease, your relative risk-- that is, the chance you're going to get this disease to the chance you get it at random in the population-- is not really reduced that much.

And so, since this study relied solely on twin data and didn't make any other assumptions, it raised the question in the field of to what extent is personal genome sequencing going to be really helpful? So on that note, we've talked a lot today about the analysis of human genetic variation. I hope you've enjoyed it. Once again, on Thursday we have Ron Weiss coming. Thank you very much. This concludes lecture 20 of the formal material in the course and we'll see you in lecture. Thank you very much.