PROFESSOR: OK. Well hello, everyone. And welcome back to Computational Systems Biology. I am David Gifford and I am delighted to be back with you today.

We're going to talk, today, about understanding transcription. Specifically, how we're going to understand transcription is a technique called RNA-seq. And RNA-seq is a methodology for characterizing RNA molecules through next generation sequencing.

And we'll talk, first, about RNA-seq principles. We'll then talk about how to take the data we learn from RNA-seq and analyze it using tools for characterizing differential gene expression and principal component analysis. And finally, we'll talk about single cell RNA-seq, which is a very important and growing area of scientific inquiry.

But first, let's talk about RNA-seq. How many people have heard of RNA-seq before? Fantastic. How many people have done it before? Some? Great.

So RNA-seq is fairly simple in concept. What we're going to do is we're going to isolate RNA species from a cell or collection of cells in the desired condition. And note that we can choose which kind of RNA molecules to isolate.

We can isolate molecules before any selection, which would include molecules that are precursor RNAs that have not been spliced yet, including non-coding RNAs. As you probably know, the study of non-coding RNAs is extraordinarily important. There are over 3,300 so-called long non-coding RNAs that have been characterized so far. Those are non-coding RNAs over 200 bases long. We'll be talking about those later on, when we talk about chromatin function in the genome.

And of course, there are the precursor messenger RNAs that are spliced, turned

into messenger RNAs that are then translated into protein. And the specifics of the RNA-seq protocol will give you various of these species depending upon what kinds of purification methodologies you use. But as you're aware, there are many isoforms that are possible in most mammalian genes.

This is a short summary, produced by the Burge Laboratory, of different kinds of splicing events that can occur. And the splicing events are often regulated by cis-regulatory sequences that live in the introns. And these introns contain recognition sequences for splicing factors where the splicing factors can be conditionally expressed. And so you get different combinations of these exons being glued together to produce variant forms of proteins.

So we have to be mindful of the idea that the RNA molecules that we're going to be observing, typically, are going to be reverse transcribed. So we'll see entire transcripts that came, perhaps, from distinct exonic locations in the genome.

And the essential idea of RNA-seq is that we take the RNA molecules we care about-- in this case, we're going to purify the ones that have poly-A tails. We will take those molecules. We'll reverse transcribe them. We'll sequence fragments of them and then map them to the genome.

Now if you don't purify for poly-A tails, you get a lot of things mapping to the genome that are intronic. And so you get a lot of data that is very difficult to analyze. So typically, people, when they're looking for gene expression data, will do poly-A purification. And when you do this-- when you sequence the result of doing the reverse transcription of these RNA molecules-- and you map the results to the genome, what you find is data that looks like this.

This is the SOX2 gene. This is typical expression data. You can see all of the individual reads mapping to the genome, the blue in the plus strand, the pink on the minus strand. And our job is to take data like this and to analyze it.

Now we can take these data and we can summarize them in various formats. One way is to simply count the number of times that we see a read at a particular base,

like here, for the SMUG1 gene. And here you see something else going on, which is that we have reads from-- the sequencing experiments have been polyadenylated and purified. So we're only seeing the reads that occur over the exonic sequences, more or less. There are a few intronic reads there, scattered about.

The other thing that we see, which is very important, is that we see reads that are split across exons. Because the splicing event has occurred, the RNA molecule is a contiguous sequence of a collection of exons. And sometimes you'll get a read that spans an exon-exon boundary. And when we get this, you can see, in the bottom part of the slide that I'm showing you, these reads can map across the exons.

Typically, in order to get good performance out of something like this, we want to use reads that are about 100 bases long. And we'll use a mapper that is capable of mapping both ends of a read to account for split reads for these exon crossing events. So that gives you an idea of the kinds of data that we have. And part of the challenge now is looking at how we can determine which particular isoforms of a gene are being expressed by evaluating these data.

So there are two principal ways of going about this. One way is we simply take our read data, and we use the ideas that we talked about in genome assembly, and we assemble the reads into transcripts. That's typically done de novo. There are reference guided assemblers.

It has the benefit that you don't need a reference genome. So sometimes, when you're working with an organism that does not have a well characterized reference genome, you will de novo assemble the transcripts from an RNA-seq experiment. But it also has problems with correctness, as we saw when we talked about assembly. The other approach, which is more typically used, is to map reads, or aligned the reads, to a reference genome and identify the different isoforms that are being expressed using constraints. And ultimately, the goals is to identify the isoforms and quantitate them so we can do further downstream analysis.

And you'll hear about two different metrics, sometimes, in the literature, for expression. One is the number of reads per kilobase of transcript per million reads.

So you might have, for example, an RPKM metric of 1,000, which means that one out of every thousandth read is mapping to a particular gene. So it gives you a metric that's adjusted for the fact that longer genes will produce more reads.

An alternative metric is fragments per kilobase per million. And that's sometimes used when we're talking about paired end data. And you're considering that you're sequencing both ends of a fragment. So here we're talking about how many fragments we see for a particular gene per 1,000 bases of the gene per million fragments.

OK. So the essential idea, then, is to take the reads that we have-- the basket of reads-- align it to the genome, both to exons and to exon crossings, and to determine, for a given gene, what isoforms we have and how they're being expressed. So from here on in, I'm going to assume that we're talking about a gene and its isoforms. And that's OK. Because typically, we can map reads uniquely to a gene. And there are details, of course, when you have genes there are paralogs that have identical sequences across the genome where this becomes more difficult.

So if we consider this, once again, we're going to take our reads, we're going to map them to the genome, and we're going to look for all possible pairings of exons. What we would like to do is to enumerate all of the possible isoforms that are possible given the data that we have. And we can use junction crossing reads and other methodologies to enumerate all the possible isoforms.

But what we're going to assume is that we've enumerated all the isoforms. And we're going to number them 1 through n. So we have isoform 1, isoform 2, isoform 3 for a given gene. And what we want to compute for each isoform is its relative contribution to the read population we're seeing that maps to that gene.

So in order to do that, what we could do is use some constraints. So if I show you this picture, which suggests that we have possible splice events here for the event C in the middle, if I told you that A, C, and E were very highly covered by reeds, you might think that A, C, and E represented one isoform that was highly expressed.

And so if we think about how to use our read coverage as a guide to determining isoform prevalence, we'll be in good shape.

And really, that's the best evidence we have. We have two sources of evidence, right? We have our junction crossing reads, which tell us which exons are being spliced together, that helps us both compute the set of possible isoforms and estimate their prevalence. The other thing we have is the reads that actually simply cover exons. And their relative prevalence can also help us compute the relative amounts of different isoforms.

So in order to do this, we can think about what reads tell us. And some reads will exclude certain isoforms. So if we consider the reads that we have, we can think about reads that cross particular junctions that are inclusion reads saying that-- for example, in this case, the top reads are indicating that the middle white exon is being included in a transcript whereas the bottom reads are exclusion reads indicating that that white exon in the middle is being spliced out.

So what we would like to do, then, is to build a probabilistic model that takes into account what we know about what a read tells us. Because each read is a piece of evidence. And we're going to use that read like detectives. We're going to go in and we're going to try and analyze all of the different reads we see for a gene and use it to weight what we think is happening with the different isoform expressions in the pool of reads that we're observing.

And in order to do so, we will have to build a function that describes the probability of seeing a read given the expression of a particular isoform. So the essential idea is this-- for the next three slides, I want to build a model of the probability of seeing a read conditioned upon a particular isoform being expressed. All right? So there are three ways to approach this.

One is that I can see a read that I know is incompatible with a given isoform. And therefore, the probability of seeing that read given the isoform is 0. And that's perfectly fine. And this can either happen with a single ended read or with a paired end read. And it's a conditional probability. So the probability of seeing read i given

and isoform j can be 0.

Another possibility is that, if I have a transcript-- now recall that the transcript has been spliced. So what we're looking at is the entire sequence of the spliced isoform. And we see a read. And the read can land anywhere within that transcript. Let's assume, for the time being, we're looking at single ended read. Then, the probability of seeing that read land in that transcript is 1 over the length of the transcript-- all right-- at a particular base.

So this read is compatible with that transcript. And we can describe the probability in this fashion. It's also possible for us to consider paired end reads. And if we have paired end reads, we can describe a probability function that has two essential components.

The denominator is still the same. That is, the likelihood of the read aligning at a particular position is going to be 1 over the length of the entire transcript. The numerator is different though. We're going to compute the length of the read that we have, or the implied length of the paired end read, and ask, what's the likelihood of seeing that.

So we don't know the exact length, recall, of the insert. When we're looking at paired end reads, we can only estimate how long the fragment is that we're sequencing. And so we are going to have a probabilistic interpretation of how long the piece of RNA is that we actually wound up sequencing the ends of. And that is placed in the numerator, which scales the 1 over l sub j. So this gives us a probability for seeing a read in a particular transcript that accounts for the fact that we have to back both ends to that transcript. OK?

So we have three possibilities that we've described, one, where a particular read is incompatible with an isoform, two, where we had a single end read mapping to an isoform, which is simply 1 over the length of the isoform, and three, where we're mapping paired end reads to an isoform, which includes uncertainty about where it will map, which is the 1 over l sub j, and also uncertainty about the length of the fragment itself, which is encoded in the F function, which is a distribution over

fragment lengths.

OK. So once we have this structure, we can then estimate isoform expression. Now we talked before, when we talked about ChIP-seq last time, the idea of estimating proportions. And the essential idea here is that if we want to compute the probability of a read given, in this case, a mixture of isoforms, that's simply going to be-- let's see, what variable did I use? Yeah-- the estimated concentration of that isoform times the probability of the read as seen in that isoform.

So for an individual read, we can estimate its likelihood given this mixture. And then the product around the outside describes how to estimate the probability of the entire basket of reads that we see for that gene. And what we would like to do is to pick psi, in this case, to maximize the likelihood of the observed reads. So what psi is going to do is it's going to give us the fraction of each isoform that we see.

Are there any questions about the idea of isoform quantitation? Yes.

AUDIENCE: I'm a little lost in-- in the last full slide review, you were describing these three cases for excluded, single, and paired reads. So we're computing the different probabilities for both ends to happen in a transcript, of for just one, or--

PROFESSOR: It depends. The second and third cases depend upon whether we're analyzing single ended reads or paired end reads. And so we wouldn't use both of them at the same time. In other words, if you only have single ended data, you would use the second case that we showed. And if you had paired end data, you would use the third case that we showed.

AUDIENCE: OK.

PROFESSOR: Question, yes.

AUDIENCE: Sorry. I noticed that the single end reads case-- could you explain the intuition behind that probability [INAUDIBLE]?

PROFESSOR: Sure. The intuition behind that probability is that we're asking-- so here is a

transcript. And we're assuming, what is the probability of a read given that it came from this transcript. OK? What's the probability of observing a particular read?

And the probability of observing it lining up at a particular position is 1 over the length of this transcript. And so the probability actually includes the idea of alignment at a particular position in the transcript. OK? So obviously, the probability's 1 if we assume it comes from here if we don't consider this fact. But if we want to ask where it lines up in the transcript, it's going to be 1 over l sub j. OK?

AUDIENCE: So we assume that it's uniformly possible?

PROFESSOR: That it's uniformly possible, which gets us to a good point. I'm glad you asked that question. Sometimes we would have more reads at the three prime end of a transcript than the five prime end. Can anybody imagine why? Yes.

AUDIENCE: Because you're pulling on the poly-A tails.

PROFESSOR: Yeah. So this actually was purified by the poly-A tail. And we're pulling on this. We're purifying by it. And if there's breakage of these transcripts as it goes through the biochemical processing, we can get shorter and shorter molecules. They all contain this bit. And the probability to contain that whole thing actually decreases.

So oftentimes there is three prime bias in RNA-seq experiments one needs to be mindful of. But we're assuming, today, that there isn't such bias and that the probability's equal that it maps someplace in this transcript. OK? Does that answer your question? Yes.

AUDIENCE: Sorry. I do have one more. Can you show us how that extends, then, to the paired end read and where the probability distribution--

PROFESSOR: Right. So if we go to paired end reads, right, like this, this component is going to be where it aligns, right? And then, the probability of the length of this entire molecule is what I had up there before, which is-- exactly how did I do that?

So this is going to be-- this is this bit, which is the implied length of this. OK? So if I map the left and the right-- this component is where the left end maps. OK? I take

the left and the right ends of the read that I have from that transcript. And it has a particular length on this transcript. OK? I'll call that length l sub j of R sub i.

Now remember, that is not the same as the length in the genome. That's the length in this transcript as it is spliced. OK? F is going to be the probability distribution of the length of the fragments. So let's just say that they're centered around 200 bases. OK?

So if this is exactly 200 bases, it's going to be quite likely. OK? But imagine that when I map this fragment, this wound up being 400 bases apart. Then, this distribution would tell is it's very unlikely that I would see a fragment that mapped here and mapped 400 bases up here, because my fragment with distribution defined by F is 200 bases.

So it's going to discount that, the probability of that. So this term that the probability of the read given the transcript is the component of where it's aligning times the likelihood that the implied fragment length agrees with what we think we have empirically. OK? Does that make sense? OK. Those are good questions.

OK. So given this framework, we can either use EM or other machine learning like frameworks to maximize psi and to learn the fraction of expression of each different isoform from the observed data given the functions that we have. And just to give you an idea, when this was done for myogenesis, a program called Cufflinks, which does this kind of process of identifying isoform prevalences, was able to identify a large number of transcripts.

70% of the reads were in previously annotated transcripts. But it also found 643 new isoforms of genes in this single time series. And I posted one of the papers that describes some this technology on the Stellar site. But note that certain of the genes have light coverage.

And what we're seeing here is that for genes are expressed in low copy numbers, it's obviously more difficult to get reads out of them. And I'm presuming, in this particular experiment-- although I can't recollect-- that the reason they don't see

that many intronic reads is they did poly-A purification.

OK. So we've talked about how to take our reads, build a probabilistic model, estimate isoform prevalences. And we know how many reads are mapping to a given gene. The question now is whether or not we see differential expression of a gene in different conditions. So I'd like to turn to the analysis of differential expression unless there are any final questions about the details of RNA-seq and isoform estimation.

This is your chance to ask those hard questions. Yes.

AUDIENCE: OK. I have a really silly question. But can you explain, really quickly, what are isoforms?

PROFESSOR: What an isoform is?

AUDIENCE: Yeah.

PROFESSOR: Sure. An isoform is a particular splice variant of a gene. So a gene that has a particular splicing pattern is called an isoform. So imagine we have three exons, one, two, and three. And a transcript that has all three would be one isoform. And another variant that omits two would be a second isoform. So just one in three would be an isoform.

And each gene has a set of isoforms it exhibits. And that depends upon how it's regulated and whether or not any splicing is constitutive-- it always happens-- or whether or not it's regulated. And so in theory, a gene with n exons has how many potential isoforms?

It's 2 to the n. Because you can consider each exon being included or omitted. All right. But that isn't typically the case-- that there are many fewer isoforms than that. But in general, an isoform refers to a particular splice variant of a gene. Yes.

AUDIENCE: I just want to make sure I have everything correctly. When you're using single legged or pair end reads, you can get excluded ends, right? So you can get that in both cases, whether or not you're--

PROFESSOR: Well, it depends. Once again, it's somewhat probabilistic where the reads actually hit. Because if all the reads only hit exons, and you didn't get any junctions, and none of your paired end reads crossed junctions, then you wouldn't actually have exclusion. All right?

AUDIENCE: But it's possible for using both types of sequencing?

PROFESSOR: Yes, it's possible with both types of sequencing. In fact, oftentimes, what people do is that they will count junction crossing reads. If you have a large enough number of reads in your sequencing, say, 100 base pairs at least, then a large number of your reads are going to be-- not a large, but a significant fraction-- will be exon crossing. And you'll be able to count the number of exon-exon junctions you have of different types.

And that will give you an estimate of how much splicing is going on and will help validate the kinds of conclusions that come out of programs like Cufflinks or MISO, which is another program from the Burge Laboratory that is used to estimate isoform prevalence. Yes.

AUDIENCE: So even given this information, you can't say which exons go with which exons necessarily, except paralogs, right? Because the reads, in general, aren't long enough to span an exon. And therefore we wouldn't know, for example, that a given transcript is exons one, five, and six. You could only know that exons five and six went together.

PROFESSOR: That is not strictly true if you have paired end reads and your fragments are long enough to span exons. But in general, you're correct. And that's why modern sequencing technologies that are coming down the pike that can do 25 kilobase reads are so important for doing things just like that. Yes.

AUDIENCE: In your diagram of the read up there, are the boxes the actual genome or RNA sequence and the line in between the artificial linker that you added when you seq ref'd?

PROFESSOR: Ah, that's a good question. The question is, is this the linker and these are the actual sequences. No. What I'm drawing here is that these are the bits that we actually get to see the sequence of. We sequence from both ends of a molecule. This is the part of the fragment that we haven't sequenced because our reads aren't long enough.

And so the entire fragment might be 300 bases long. If this is 100 and this is 100, then the unobserved part is 100 in the middle. OK? And that's called the insert length, the entire length of the molecule.

And we get to choose how long these fragments are up to a maximum size. Contemporary sequencers don't really like fragments over 1,000 bases. And the performance starts falling off when you get close to that number. So people, typically, are operating in a more optimal range of fragments that are a few hundred bases long. Any other questions? OK.

So I wanted to briefly talk about hypothesis testing. Because we're going to be needing it for determining when things are really differentially expressed. So I'm just going to show you some data and ask you a few questions about it.

So here are two different scatters of data. Well, actually, it's exactly the same data. But we have two different bits to it. We have two independent Gaussians that are fit to the data, from gene one and gene two. And another fit uses two Gaussians that have a correlation structure between them.

And the question is whether or not the null hypothesis or the alternative hypothesis is more reasonable. And typically, when we say reasonable, we want to judge whether or not it's significant. Significance typically talks about, what's the chance that the data we saw occurred at random given the null hypothesis. So what's the chance it was generated by the null hypothesis versus the chance that it was generated by the alternative hypothesis?

Now the problem is that alternative hypotheses, typically, are more complex. And a more complex model will always explain data better. So we need to have a

principled way of asking the question, given that the alternative hypothesis is always going to do a better job, does it do such a better job that it can exclude the null hypothesis at a particular probability level. OK?

So here are two different models for these data. The null model, H0, is that they came from two independent Gaussians. The alternative model, H1, is that they came from two correlated Gaussians. And then we can ask whether or not H1 is sufficiently more likely to warrant our rejecting H0 and accepting H1.

Now as I said, H1 is always going to fit the data better. So the probability of that collection of points evaluated with the H1 model, fit to the data, is always going to be superior. So we need to have a way to compare the probability of the data given H1 versus the data given H0 in a way that allows us to judge the probability that the data via H0 occurred at random.

In this particular case, the data supports H1. And let's see why. So this is a key idea here. How many people have heard of likelihood ratio statistics before? OK. About half the class. OK. So here's the idea.

The idea is that what we're going to do is we're going to compute a test statistic. And the test statistic is going to be a function of the observed data. And it's 2 times the log of the probability of the observed data given H1 over the probability of the observed data given H0. OK?

Now we know that this is always going to have a higher value than the probability in the denominator. So this is always going to be greater than 1. So the test statistic will always be greater than 0 since we're operating in the log domain. OK?

The question is-- we know that this is always going to be better, even when the data was generated from H0. But when is this sufficiently better for us to believe that H0 is not true and we should accept H1? What we need is a distribution for this test statistic that occurred if H0 was true. And that distribution allows us to compute the probability that an observed value for the test statistic occurred, even in the presence-- assuming that H0 is true.

OK. So this depends upon the number of degrees of freedom difference between H1 and H0. How many degrees of freedom are there in H1 in this model up here? How many parameters do we get to pick?

AUDIENCE: Six.

PROFESSOR: Hmm?

AUDIENCE: Six.

PROFESSOR: Six?

AUDIENCE: Two means and four--

PROFESSOR: Two means and four coherences. And for H0?

AUDIENCE: Just four.

PROFESSOR: Four. So what's the difference in the number of degrees of freedom between H1 and H0? It's two. So the test statistic is parametrized by the difference in number of degrees of freedom. And so what we see, then, is something that looks like this.

We see a test statistic where this is the probability of it, on the y-axis, and the test statistic on the x-axis. But as the test statistic gets larger and larger, the probability that it occurred with H0 being true gets smaller and smaller. So let us just suppose that we took our data from our model that we observed. And we computed the test statistic at a particular value call T observed.

So this is the actual value that we computed out of our likelihood ratio test. What we would like to ask is, what's the probability that our test statistic is greater than or equal to T observed given that H0 is true, which means that we're going to consider all the tail of this distribution. Because we want to also consider the case where T observed was even greater than what we saw. And this gives us a way of computing the probability that H0 is true given the test statistic. And this gives us our p-value. OK?

So this is a way of, in general, comparing two probabilistic models and analyzing the significance of adding extra degrees of freedom to the model. Typically, what we'll be doing in today's lecture is asking whether or not-- if we let the means change, for example, between two conditions-- we get a sufficient improvement in our ability to predict the data that our test statistic will allow us to reject the null hypothesis that the means are the same.

OK. I'm going to stop here and see if there are any questions at all about this.

AUDIENCE: Yes. Where did the degrees of freedom enter into the equation?

PROFESSOR: Where did the degrees of freedom enter into this? Great question. The chi-square tables that you look up are indexed by the number of degrees of freedom of difference. OK? And so whenever you compute a chi-squared, you will compute it with the number of degrees of freedom difference. Any other questions?

OK. So let's now turn to evaluating RNA-seq data once again. And I'm going to describe a method called DESeq for determining differential expression. And in our analysis, what we're going to do is we're going to let i range over a gene or an isoform. j is an experiment. And there may be multiple experiments in the same condition that are replicates. And Kij is the number of counts observed for i and j. So that's the expression of gene or isoform i in experiment j.

Now what we need to do, however, is to normalize experiments against one another. And the normalization factor s sub j is computed for a particular experiment. And it's used to normalize all of the values in that experiment. And if all the experiments were completely identical-- the read depth was identical and everything was the same-- then all of the s sub js would be 1.

If you had an experiment that had exactly twice as many reads as the other experiments, it's s sub j would be 2. So this scale factor is used to normalize things in our next slide, as we'll see. And the essential idea is that we're going to take the median value of this ratio.

And the reason that the denominator is a geometric mean is so that no one

experiment dominates the average. They all have equal weight for the average. But the geometric mean is simply the product of all of the expressions for a particular median gene taken to the root m power to get them back to the value for a single experiment. And that is the normalizing factor for the numerator, which is the number of counts for a particular gene. OK?

So we're just doing median style normalization where s sub j is a scale factor. Once again, if all the experiments were the same, s sub j would be 1. If one particular experiment had twice as many counts as another experiment uniformly, s sub j would be 2, just for that experiment. Any questions about the scale factor? Yes.

AUDIENCE: Sorry, what is the term on the bottom-- in the denominator?

PROFESSOR: That's a normalizing term across-- that's the geometric mean of all the experiments put together. All right? So because it's the product of all the experiments-- m experiments-- then, rooted m, it's equal to the geometric mean of a single experiment. Any other questions? Yes.

AUDIENCE: Are we normalizing different experiments to each other or different replicates of a single experiment?

PROFESSOR: In this particular case, each one of these is a different replicate. OK? So j is ranging over different replicates, not over conditions, right now. So each replicate, each experiment, gets its own normalizing factor. We'll see, in a moment, how to put those replicates together to build statistical strength. But we need-- since each replicate has its own read depth, we have to normalize each one independently. OK?

So what we then do is we compute an expression for a condition. Now a condition, we're going to call p. And q sub ip is the normalized expression for gene slash isoform i in condition p.

So a condition may have multiple replicates in it. So we're going to average, over all of the replicates, the average expression, as you can see here. So we're summing over all the replicates for a given condition. We're going to take each replicate,

normalize it by its scale factor we just computed, and then compute the normalized expression for a gene or an isoform in that particular condition. Is that clear to everybody, what's going on here?

Now I'm describing this to you because the key fact is the next line, which is that we compute the mean for a particular replicate by taking the normalized expression for a gene and then reverse correcting it back to scaling it back up again for that particular replicate by multiplying by s sub j. But the most important thing is what's on the right hand side, which is that the variance is equal to the mean plus this function of the expression.

And the reason this is important is that most other models for modeling expression data use Poisson models. And we've already seen, when we talked about library complexity, that Poisson models don't work that well all the time. So this is using a negative binomial function, once again. We saw negative binomials before. We're modelling both the mean and the variance. And the variance is a function, a linear function and a non-linear function, of the mean.

So what's going to happen, then, is that we're going to use the negative binomial to compute the probability of observing the data in a given condition. And we can either combine conditions and ask, what's the probability of seeing the conditions combined with a single mean and variance, or we can separate them into a more complex H1 hypothesis and ask, what's the probability of seeing them with separate means and variances, and then do a test to see how significant the difference is to determine whether or not we can exclude the fact that the genes are expressed at the same level.

And to give you an intuitive idea of what's going on, this is a plot from the paper showing the relationship between mean expression and variance. Recall, for Poisson, that variance is equal to mean. We only have one parameter to tune for Poisson, which is lambda. The purple line is Poisson. And you can check and see that the mean's equal to the variance in that case. What DESeq does is it fits the orange line to the observed data. Yes.

AUDIENCE: I don't know where the v sub p comes from. Where is that, again?

PROFESSOR: That's the function. V sub p, in that equation up there, is the function that we're fitting-- that's the solid orange line-- to the observed relationship between mean and variance in the data. OK? So DESeq fits that function. EdgeR is another technique that does not fit and instead uses the estimate that's the dotted line, which isn't as good.

So a lot of people use DESeq these days for doing differential expression analysis because it allows the variance to change as the mean increases. And this is the case for these type of count data. Before I go on though, I'll pause and see if there are any questions at all about what's going on here. Yes.

AUDIENCE: You said that mu sub p, then, is the [INAUDIBLE] to the data. I'm confused. How do we fit that function, or where does that come from?

PROFESSOR: You mean the mu sub p?

AUDIENCE: Yes.

PROFESSOR: That function is fit. And the paper describes exactly how it's fit, which I posted on the Stellar site. But it's a nonlinear function of q, in this case. Good question. Any other questions? OK.

So once again, we have two hypotheses, the null hypothesis that A and B are expressing identically, H1, A and B differentially express. We can compute the number of degrees of freedom. And we can do a likelihood ratio test, if we'd like, to compute the probability of H0.

And our model in this case is the negative binomial model of the data, which fits the data better. And that's why DESeq does a better job than other methodologies. Because it provides a better approximation to the underlying noise.

And the next slide shows what you get out of this kind of analysis, where the little red dots are the genes that have been called significant using Benjamini-Hochberg

correction, which we talked about previously. And you can see how, as the mean increases, the required log 2 fold change comes down to be significant. So oftentimes, you'll see plots like this in papers that describe how they actually computed what genes were differentially expressed. Any questions at all? Yes.

AUDIENCE: Why is it that the significance is lower as the mean increases?

PROFESSOR: Why the significance is lower?

AUDIENCE: Or the threshold.

PROFESSOR: Oh, because as you increase the number of observations, the mean value theorem is going to cause things to actually get closer and closer to 0. And so you need less of a fold change difference to be significant as you get more and more observations. And other questions? OK.

So now we're going to delve into one other area. How many people have done hypergeometric tests before? OK. So we're going to talk about hypergeometric tests.

So imagine that we have a universe, for simplicity, of 1,000 genes. OK? So we have this universe. And we have, B is a set of genes that there are 30 of them. And there's another set, A, of which there are 20. And the overlap between these two is a total of three genes.

So it might be that A is the set of genes that are differentially expressed between two conditions. B is the set of genes that, you happen to know, have a particular annotation. For example, they're involved in stress response in the cell. And you'd like to know whether or not the genes that are differentially expressed have a significant component of stress response related genes or whether or not this occurred at random. OK?

So we need to compute that. So how many ways could we choose B? Well, if we are going to use-- this is n1, n2, this is big N, and this is k. All right? So the number of ways I can choose B is big N choose n2. That's the number of ways I can choose B.

Is everybody with me on that? Yeah? OK.

How many ways can I choose three elements out of A, these three that are going to overlap? Well, that's going to be n1 choose k. So that's how many ways I can choose these three elements.

And then, how many ways could I choose the other elements of B? So once again, I'm figuring out how I can choose B. Well, how could I choose the rest of B? Well, how many elements do I have to choose, of B, here?

Well, B is n too big. But I've already chosen k of them. Right? All right. Sorry, it's the other way around. The universe I can pick from is 1,000, which is all the elements, minus the elements of A that I've already chosen from to get those three.

And then I need to pick the 27 things they don't overlap with A. So 27 things that don't overlap with A would be n2 minus k. So this is the number ways to choose B given this set of constraints. This is the number of ways to choose B given no constraints. So the probability that I have overlap of exactly k is equal to this, which is, how many ways are there with no constraints and how many ways are there given that I have an overlap of k. All right?

And typically, what I want to ask is, what is the probability that my observed overlap is greater than or equal to k. So this case, the overlap would be three. But I also would need to consider the fact that I might have four, or five, or six, which would be even more unlikely, but still significant. So if you look at the exact computation, the probability of three here is 0.017. And the probability that I have three or more is 0.02.

So that's still pretty significant. Unlikely that would occur by chance. Right? That I have three or more genes overlapping in this situation could only happen two out of 100 times. Does everybody understand what's going on here? Any questions at all? So you're all now hypergeometric whizzes, right? All right? Fantastic.

OK. Now we're going to turn to a final kind of analysis. How many people have heard of principal component analysis before? How many people know how to do

principal component analysis? A few. OK. Great. Yes.

AUDIENCE: Sorry, could you just briefly mention, again, where exactly do we use the hypergeometric test? What kinds of questions are we asking when we do that?

PROFESSOR: Typically, they're overlap questions. So you're asking-- you have a universe of objects, right-- like, in this case, genes. And you have a subset of 20 and a subset of 30. Let's say these are the differentially expressed genes. These are genes in the stress response pathway. They overlap by three genes. Does that actually occur at random or not? All right?

If I told you that there are a much smaller number of genes and the stress response genes were very much larger, it could be much easier for that overlap to occur at random. Good question. Any other questions?

OK. So the next part of lecture is entitled "multivariate Gaussians are your friends." OK? They are friendly. They're like a puppy dog. They are just wonderful to play with and very friendly. And the reason most people get a little turned off by them is because they get this-- the first thing they're shown is this very hairy looking exponential which describes what they are.

And so I'm going to shy away from complicated looking exponentials and give you the puppy dog, my favorite way of looking at multivariate Gaussians. OK? Which, I think, is a great way to look at them. And the reason we need to look at multivariate Gaussians is that they help us understand what is going on with principal component analysis in a very straightforward way.

And the reason that we want to use principal component analysis is that we're going to be able to reveal hidden factors and structures in our data. And they're also going to allow us to reduce the dimensionality of the data. And we'll see why in a moment.

But here is the friendly version of multivariate Gaussians. And let me describe to you why I think this is so friendly. So we're all familiar with unidimensional Gaussians like this. Centered at zero. They have variance 1. Just very friendly univariate Gaussians, right? Everybody's familiar with those? Normal distributions?

So let's suppose that we just take a whole collection of those. And we say that we have a vector z that is sampled from a collection of univariate Gaussians. And it can be as long as we like. OK? But they're all sampled from the same distribution. And what we're going to say is that our multivariate Gaussian, x, is going to be a matrix times z plus a mean.

And so what this matrix is going to do is it's going to take all of our univariate Gaussians and combine them to produce our multivariate Gaussian. All right? So the structure of this matrix will describe how these single variate Gaussians are being mixed together to produce this multivariate distribution. And you can imagine various structures for this matrix A. Right?

And the covariance matrix, sigma, which describes the structure of this multivariate Gaussian, is shown on this slide to be equal to A A transpose. And thus, if we knew this matrix A, which we may not know, we'd be able to compute the covariance matrix directly. OK?

Let me take that one more time. We take a bunch of univariate Gaussians, make a vector z out of them. And just for clarity, right, we're going to talk about matrices and vectors as rows across columns. So this is n by 1. This is n by n. This is n by 1. And this is n by 1. OK? That's the dimensionality of these various objects we're dealing with here.

So we get this vector of univariate Gaussians. We apply this matrix n to combine them together. We get out our multivariate Gaussian offset by some mean. Is everybody happy with that so far? Yes? No? You're suspending disbelief for the next slide. Is that-- OK.

Well, here's the next thing I'd like to say, is that the variance of a vector-- we'll call it v-- times x, which is a random variable, is going to be equal to-- x is derived from this distribution-- v transpose sigma v. And the demonstration of that is on the top of this page. So the variance of this vector-- sorry, the projection of this random variable onto this vector-- is going to give you a variance in this direction as that

product.

So what we would like to do is this. We would like to find v sub i, which are vectors, to maximize variance of v sub i transpose x such that v sub i transpose v sub i is equal to 1. In other words, they're unit length vectors. So these are going to be called the eigenvectors.

And if we think about the structure that we desire, what we'll find is that they satisfy the constraint that the covariance matrix times an eigenvector is equal to the eigenvalue associated with that vector times the vector itself. And with a little manipulation-- if we multiply both sides by v sub i transpose-- v sub i equals v sub i transpose lambda sub i sigma squared v sub i-- and we move these guys around, v sub i transpose sigma v sub i is equal to-- these two guys, multiplied together, equal 1-- lambda sub i squared.

This, we see up above, is equal to the variance when it's projected in the direction of v. And so lambda sub i squared is simply the variance associated with that direction. So the question then becomes, how do we find these things. And how do we discover these magic eigenvectors that are directions in which this multivariate Gaussian has its variance maximized?

And we can do this by singular value decomposition. So we can compute this covariance matrix. So we compute sigma from the data. And then we do a singular value decomposition such that sigma is equal to U S U transpose.

And that's what the singular value decomposition does for us is it decomposes the sigma matrix into these components where S is a diagonal matrix that contains the eigenvalues and U is a column. Each column is an eigenvector. So in doing a singular value decomposition, we get the eigenvalues and the eigenvectors.

The other thing, you recall, was that sigma was equal to A A transpose when we started off. A was the matrix that we used to make our multivariate Gaussian out of our univariate Gaussians. And thus, what we can observe is that our multivariate Gaussian x is equal to U S to the 1/2 times z plus a mean. So here is what's going

when we make a multivariate Gaussian is we're taking a bunch of univariate Gaussians, we're scaling them, and we're rotating them. OK?

And that makes a multivariate Gaussian. And then we offset this whole thing by a mean. Because we also have to do rotations around the origin.

So the way I think about multivariate Gaussians is that it is a scaling and a rotation of univariate Gaussians. And implicit in that scaling and rotation is the discovery of the major directions of variance in the underlying data as represented by the eigenvectors. And the eigenvalues tell you how much of the variance is accounted for in each one of those dimensions. Are there any questions about that? I hope there are. Yes.

AUDIENCE: How do you compute sigma from the data? Is it some magical process?

PROFESSOR: The sigma of the value of decomposition?

AUDIENCE: No. So--

PROFESSOR: How do you compute sigma from the data?

AUDIENCE: Yeah, the first step.

PROFESSOR: That is shown in equation seven. So you can compute the means. And you know you have x, which are observed values. So you compute that expectation. And that is sigma. OK? Good question. Any other questions?

OK. So we have these eigenvectors and eigenvalues, which represent the vectors of maximum variance in the underlying data. And we can use these to organize data by projecting observations onto these eigenvectors-- or they're sometimes called principal components-- defined dimensions of variability that help us organize our underlying data. We'll come back to that in a moment. OK. Any other questions about principal component analysis? Yes.

AUDIENCE: [INAUDIBLE] e was expectation when you were calculating the second?

PROFESSOR: Yes, e is the expectation is correct.

AUDIENCE: And also what that means.

PROFESSOR: It's the average expected value. So in the case of computing sigma, you would compute the expected value of that inner equation across all the data points that you see. So you'd sum up all the values and divide by the number of things that you had. Any other questions? OK.

So just for calibration for next year, how many people think they've got a general idea of what principal component analysis is-- a general idea? Uh-oh. How many people who thought it was really interesting were sort of completely baffled about halfway through? OK. All right.

Well, I think that recitation can help with some of those questions. But if anybody has a question they'd like to ask now-- No? It's that far gone? I mean, the thing with this sort of analysis is that if your matrix algebra is a little rusty, then, when you start looking at equations like that, you can get a little lost sometimes.

All right. Well, let's turn, then-- if there aren't any brave souls who wish to ask a question, we'll turn to single cell RNA-seq analysis. So I'm a firm believer that single cell analysis of biological samples is the next big frontier. And it's being made possible through devices like this.

This is a Fluidigm C1 chip, which has 96 different reaction wells, which allows you, in each well, to process a single cell independently. And the little winds are ways to get reagents into those cells to do things like produce RNA-seq ready materials. And when you do single cell analysis, you can take apart what's happening in a population.

So an early paper asked some fairly fundamental but simple questions. For example, if you take two 10,000 cell aliquots of the same culture, and you profile them independently, and you ask how well to the expression values for each gene agree between sample A and sample B, you expect there to be a very good agreement between sample A and sample B in these 10,000 cell cultures.

A second question is, now, if you take, say, 14 cells from those cultures and you profile them independently, and you ask, how well do they correlate with what you saw in the 10,000 cell experiment, that will tell you something about the population heterogeneity that you're observing. Because if they correlate perfectly with the 10,000 cell experiment, then you really know that there's no point in looking at individual cells in some sense, because they're all the same. Seen one, seem them all. Right? But if you find that each cell has its own particular expression fingerprint and what you're seeing in the 10,000 cell average experiment wipes out those fingerprints, then you know it's very important to analyze each cell individually.

So the analysis that was done asked exactly that question. So here's what I'll show you in these plots. So here is, on the upper left, the 10,000 cell experiment versus the 10,000 cell experiment. And as you can see, the correlation coefficient is quite high-- 0.98-- and looks very, very good of experiment one versus experiment two, or rep one, rep two. Here is a separate experiment which is looking at two individual cells and asking-- and plotting, for each gene, the expression in one cell versus the gene in the other cell.

And you can see that the correlation coefficient's 0.54. And there's actually a fairly wide spread. In fact, there are genes that are expressed in one cell that are not expressed in the other cell, and vice versa. So the expression of these individual cells it's quite divergent.

And the final panel shows how-- down here-- how a single cell average on the y-axis relates to the 10,000 cell experiment. But given the middle panel-- the panel B there-- that is showing the fact that two single cells don't really relate that well to another, it bags other questions. For example, are the isoforms of the genes that are being expressed the same in those distinct cells? And so panel D shows isoforms that are the same across each one of the single cells being profiled, which is the solid bar at the top. And the bottom couple of rows in figure D are the 10,000 cell experiment average.

But panel E is the most interesting, perhaps, which is that the isoforms for those

four genes are being differentially expressed in different individual cells. And that's further supported by taking two of those genes and doing fluorescent in situ histochemistry and microscopy, and looking at the number of RNA molecules for each one of those, and noting that it corresponds to what's seen in the upper right hand panel. So we see that in individual cells, different isoforms are being expressed.

Now these cells were derived from bone marrow. And they're exposed to lipopolysaccharide to activate an immune response. So they are clearly not all behaving exactly the same.

And to further elucidate this, the authors of this paper took the gene expressions that they saw for a given cell as a large vector and computed the principal components, and then projected the cells into the first and second principal component or eigenvector space. And as you can see, there is a distinct separation of three of the cells from the rest of the cells, where three of the cells, which correlate well with principal component one, are thought to be mature cells that express certain cell surface proteins whereas the ones on the left-- the maturing cells-- the triangle depicted cells-- express certain cytokines under the maturing legend there, on the clustergram on the right-hand side.

And thus, the first principal component was able to separate those two different broad classes of cells. So it looks like there are at least two different kinds of cells in this population. And then, the authors asked another question, which is, can they take individual cell data and look at the relationship between pairs of genes to see which genes are co-expressed.

And the hypothesis is that genes that are co-expressed in individual cells make up individual regulatory circuits. And so they hypothesize that the genes LRF7 and IFIT1 and STAT2 and LRF7 are all in an anti-viral regulatory circuit. They then ask the question, if they knocked out LRF7, which is the second panel on the right-hand side, would they oblate downstream gene expression.

And they partially did. And they thought that since STAT2 and LRF7 are both

thought to be regulators of the circuit and they're both downstream of the interfering receptor, they thought if they knocked out the interfering receptor, they would oblate most of the anti-viral cluster, which, in fact, they did.

So what this is suggesting is that, first, single cell analysis is extraordinarily important to understand what's going on in individual cells. Because in a cell culture, the cells can be quite different. And secondarily, it's possible, within the context of individual single cell analysis, to be able to pick out regulatory circuits that wouldn't be as evident when you're looking at cells en masse.

And finally-- I'll thank Mike for the next two slides-- I wanted to point out that quality metrics for RNA-seq data for single cells is very important. And we talked about library complexity earlier in the term. And here, you can see that as library complexity increases, expression of coefficient of variation, which is the standard deviation over the mean, comes down as you get sufficient library complexity. And furthermore, as library complexity increases, mean expression increases.

And the cells that are in red were classified as bad by microscopy, from the Fluidigm instrument processing step. So I think you can see that single cell analysis is going to be extraordinarily important and can reveal a lot of information that is not present in these large batch experiments. And it's coming to a lab near you.

So on that note, I'll thank you very much for today. And we'll see you later in the term. And Professor Burge will return at the next lecture. Thanks very much.