PROFESSOR: All right, well, good afternoon and welcome back. We have an exciting fun-filled program for you this afternoon. I'm David Gifford. I'm delighted to be back with you again, here in computational systems biology.

Today we're going to talk about chromatin structure and how we can analyze it. And to give you the narrative arc for our discussion today, we're first going to begin with looking at computational methods that we can break the, quote unquote code, that describes the epigenome.

Now, epigenetic state is extraordinarily important and one way you can visualize this is that the genome is like a hotel filled with lots of different rooms. And a lot of the doors are locked and some of the doors are unlocked.

And only in the doors that we can go into, where the genome is open and accessible can there actually be work done, regulation performed and transcripts and proteins made.

So we're going to talk about how to actually analyze epigenetic state. And then we're going to talk about how to use epigenetic information to understand the entire regulatory occupancy of the genome.

We've already talked about ChIP-seq and the idea that we can understand where individual regulators sit on the genome, and how they regulate proximal genes.

We're now going to see if we can learn more about the genome. How it's state-- whether it's open or closed. Is it self-regulated? And answer a puzzle.

The puzzle is, if there are hundreds of thousands of possible binary locations that are equally good for a regulator, why are only tens of thousands occupied? And

how are those sites picked? Because that level of regulation is extraordinarily important to establish a basal level of what genes are accessible and operating.

And finally, we're going to talk about how we can map, which regulatory regions in the genome are affecting which genes. It turns out that about 1/3 of the regulatory sites in the genome skip over a gene that's closest to them to regulate a gene that's farther away.

This is a million genomes. And so given that rough approximation, how is it that we can make connections between regulatory sites and the genes that they control?

Now, in computational systems biology, we always talk a lot about biology, but we also need to reflect upon the computational methods that we're bringing to bear on these questions.

And so, today, we're going to be talking about three different methods. We'll talk about dynamic Bayesian networks as a way to approach, understanding the histone code.

We'll talk about how to classify factor binding, using log likelihood ratios. And finally, we'll turn to our friend, the hypergeometric distribution to analyze which locations in the genome are interacting with one another.

So let's begin with establishing a vocabulary. I'm sure some of you have seen this before. This is the way that chromatin can be thought of being organized at different levels. There's the primary DNA sequence, which can include methylated CPGs.

That's cysteine, phosphate, guanine. And the nice thing about that is that it's symmetrical so that when you have a CPG, a methyltransferase during DNA replication can copy that methy mark over. So it's a mark that's heritable.

The next level down are histone tails. On the amino terminus of histones H3 and H4, different chemical modifications can be made, and they serve as sign posts, as we'll see, to give us clues about what's going on in the genome in that proximal location.

The next level down is, whether or not the chromatin is compacted or not. Whether it's open or closed. And that relates to whether or not DNA binding proteins are actually on the genome.

And finally, certain domains of the genome can be associated with the nuclear lamina. And so they're different levels of organization of chromatin. And we'll be exploring all of these today.

So the cartoon version of the way that the genome is organized is that at the top we have a transcribed gene. And you can see that there's an enhancer that is interacting with the RNA polymerase II start site.

And you can see varied histone marks that are associated with this activated gene. There are also marks that are associated with that active enhancer.

Down below, you see an inactive gene. And you can see that there's a boundary element that's bound by CTCF, which, one of its function is to serve as a genomic insulator, which insulates the effect of the enhancer above from the gene below.

So through careful biochemical analysis over the years, these different marks have been analyzed and characterized. And a general paradigm for understanding how the marks transition as genes are activated is shown here.

So genes that are fairly active and cycle between active and inactive states typically have a high CPG content in their promoters. And transition is shown on the left.

Where in the repressed state on the bottom, they're marked by H3K27 trimethyl marks. When they're poised, they have both H3K4 trimethyl and H3K27 trimethyl. And when they're active, they only have H3K4 trimethyl.

And on the right hand side are genes that are less active. So in their completely shut down state, they may have no marks, but the DNA is methylated, silencing that region of the genome. And other marks then, culminating in H3K4 trimethyl once again when they become active at the top.

So I'm summarizing for you here, decades of research in histone marks. And it has been summarized in figures like this, where you can look at different classes of genetic elements-- whether they be promoters in front of genes, gene bodies themselves, enhancers, or the large scale repression of the genome-- and you can look at the associated marks with those characteristic elements.

OK, so, how can we learn this de novo? That is, you could memorize, and of course it's important to understand, for example, if you want to look for active enhancers in the genome, that looking for things like H3K4 monomethyl and H3K7 27 acetyl marks together, would give you a good clue where the enhancers are in the genome that are active.

But if we want to learn all this de novo, without having to memorize it or rely upon the literature, the great thing is that there's a lot of data out there now that characterizes, or profiles all these marks, genome-wide, in variety of cellular states. And there's the epigenome roadmap initiative to look at this in hundreds of different cell types.

So, what is the histone code? That is, how can we unravel the different marks present in the genome and understand what they mean? Because the genome doesn't come ready-made with those little cute labels that we had on it-- enhancer, gene body, and so forth.

So somehow, if we want to understand the grammar of the genome and its function, we're going to need to be able to annotate it, hopefully with computational help.

So here's a picture of what typical data looks like along the genome. So, obviously you can't read any of the legends on the left-hand side. If you want to look at the slides that are posted on Stellar, you can see the actual marks.

But the reason I posted this is because you can see the little pink thing at the top-- that's where the RNA transcript has been mapped to the genome. The actual annotated genes are above. And then down below you can see a whole collection of histone marks and other kinds of chromatin information that have been mapped to

the genome and spatially create patterns that are suggestive of the function of the genomic elements, if they're properly interpreted.

And below, you see in blue, the binding of different TFs, as determined by ChIP-seq.

So, what we would like to do then, is to take this kind of information and automatically learn, or automatically annotate the genome as to its functional elements.

Let me stop here and ask, how many people have seen histone mark information before? OK. And how many people have used it in their research? Not too many-- a couple people? OK.

So it's getting quite easy to collect and there are a couple of ways of analyzing this kind of data, genome-wide. One way is that we could run a hidden Markov model over these data and predict states at regular intervals. For example, every 200 bases down the genome, and see how the HMM transition from state to state and let the state suggest what the underlying genome elements that we're doing.

Another way is to use a dynamic Bayesian network. So a dynamic Bayesian network is simply a Bayesian network. We've talked about those before. And it models data sampled along the genome. And so it's a directed acyclic graph.

There are tools out there that allow us to learn these models directly. And it allows us, as we'll see, to analyze the genome at high resolution, and to handle missing data.

So we'll be talking about Segway, which is a particular dynamic Bayesian network that takes the kind of data we saw on the slide before and essentially parses it into labels that allow us to assign function to different genomic elements. And it does this in an unsupervised way. What I mean by that is that it is automatically learning the states, and then afterwards we can look at the states and assign meaning to them.

So here is the dynamic Bayesian network that Segway uses. And let me explain this

somewhat scary looking diagram of lots of little boxes and pointers to you.

The genome is described through the variables on the bottom-- the observation variables, going from left to right, where each base is a separate observation variable which consists of the level of a particular histone mark at a particular based position as described by mapped reads to that location.

The little square box-- the little boxes that says "x" on it with the other small print you can't read-- is simply an indicator, whether or not the data is present. If the data is absent, we don't try and model it. If that box contains a zero, we don't model the data. If the box is one, then we attempt to model the data.

And the most important part of the dynamic Bayesian network is the q box above, where those are the states. And each state describes an ensemble of different histone marks that are output.

And so the key thing is that for each state we learn what marks it's outputting. And the model learns this automatically through a learning phase. The boxes above simply are a counter.

And the counter allows us to define maximum lengths for particular states, so states don't run on forever. So unlike a hidden Markov model that doesn't have that kind of control, we can adjust how long we want the states to last.

So this model, if you turned it 90 degrees and rotated it clockwise, would be more familiar to you because all the arrows would be flowing from the top of the screen down. There are no cycles in this directed acyclic graph.

And therefore, it can be probabilistically viewed and learned in the same framework that we learn a Bayesian network. In fact, it is a Bayesian network. The reason it's called dynamic is because we are learning temporal information, or in this case, spatial information with these different observations along the bottom of the model.

Now before I go on, perhaps somebody could ask me a question about the details of these dynamic Bayesian networks, because the ability to automatically assign

labels to genome function, given the histone marks is really a key thing that's gone on the last couple of years. Yes?

AUDIENCE: Could you re-explain that-- what the labeled-- the second [INAUDIBLE] was all about?

PROFESSOR: Sure. So the Q label is right here, these labels. And each of these Q labels defines one of a number of states. For example, 24 different states. In a given state, describes the expected output in terms of what histone marks are present in that state.

So it's going to describe the means of all those different histone marks. 24 different means, let's say, of the marks it's going to output. And the job of fitting the model is picking the right states, or a set of 24 states, each of which is most descriptive of its particular subset of chromatin marks. And then defining how we transition between states.

So we not only need to define what a state means in terms of the marks that it outputs, but also when we transition from one state to another. Does that make sense to you?

AUDIENCE: So I know it states the information that tells at each of the Q boxes. Is that a series of probabilities? Or is it something else?

PROFESSOR: It's actually a discrete number, right. So it actually is a single-- there's only a single state in each Q box. So it might be a number between 1 and 24 that we're going to learn. And based upon that number, we're going to have a description of the marks that we would expect to see at the observation at that particular genomic location.

And so our job here is to learn those 24 different states and what they output in the training phase, and then once we've trained the model, we can go back and look at other held out data, and then we can decode the genome.

Because we know what the states are, and we know what they are supposed to be producing, we can use a Verterbi decoder and go back and-- as we did with the

HMM and we learned the HMM-- go back and read off on the histone mark sequence and figure out what their relative states are for each base position of the genome. Is that helpful? Yes?

Any other questions about dynamic Bayesian networks? Yes?

AUDIENCE: How do you choose the number of states?

PROFESSOR: That's a very good question. How do you choose the number of states? Well, if you choose too many states, they obviously don't really become descriptive and you can become over fit and then can start fitting noise to your model.

And if you choose too few states, what will happen is, that states can get collapsed together and they won't be adequately descriptive. The answer is, it's more or less trial and error. There really isn't a principled way to choose the right number of states in this particular context. Now, you could do--

AUDIENCE: What's the trial, then? You run it and you get a set of things, and what do you do with those labels?

PROFESSOR: What do you do with labels?

AUDIENCE: Yeah, how do you evaluate it?

PROFESSOR: You typically, in both of these cases-- both in the case of chrome HMM and this-- you rely upon the previous literature. And we saw on that slide earlier, what marks are associated with what kinds of features.

So you use the prior literature and you use what the states are telling you they're describing to try and associate those states with what's known about genome function. All right, yes?

AUDIENCE: Where does that information concerning the distance between states go again? Like, the counter? Like, how does that control how long the states go on and whether or not--

PROFESSOR: What happens is that the counter at the top, the C variables, influence the J variables you can see there. When the J variable terms to a 1, it forces the state transition.

So the counters count down and can then force a state transition which will cause the Q variable to change. It's sort of a-- that particular formulation of this model is a bit of a, sort of Rube Goldberg kind of hackish kind of thing. I think to make it get out of particular states. But it works, as we'll see in just a moment. OK.

So here's an example of it operating. And you can see the different states on the y-axis here. You can see the different state transitions as we go down the genome. And you can see the annotations that it's outputting, corresponding to the histone marks.

And so what this is doing is it's decoding for us what it thinks is going on in the genome, solely with reference to the histone marks, without reference to primary sequence or anything else. And this kind of decoding is most useful when we want to discover things like regulatory elements. When we want to look for H3K4 mono or dimethyl, and H3K27 acetyl for example, and identify those regions of the genome that we think are active enhancers. OK. OK.

So, any questions at all about histone marks and decoding? Do you get the general idea that you can assay these histone marks through ChIP-seq using antibodies that are specific to a particular mark. Pull down the histones that are associated with DNA with that mark and map them to the genome.

So we get one track for each ChIP-seq experiment. We can profile all the marks that we think are relevant, and then we can look at what those parks imply about both the static structure of our genome, and also how it's being used as cells differentiate or in different environmental conditions. OK. OK.

So, let's go on, then, to the next step, which is that if we understand the sort of epigenetics state, how is that established and how is the opening of chromatin regulated and how is it that factors find particular places in the genome to bind?

So, the puzzle I talked to you about earlier was that there are hundreds of thousands of particular motifs in the genome, but a very small number are actually bound by regulatory factors.

And you might think that the difference is that the ones that are bound have different DNA sequences. But in fact, on the right-hand side, what we see is that identical DNA sequences are bound differentially in two different conditions.

Shown there are sites that are only bound, for example, in endodermal tissues or in ES cells. So it isn't the sequence that's controlling the specificity of the binding, it's something else. And we'd like to figure out what that something else is. We'd like to understand the rules that govern where those factors are binding in the genome.

So a set of factors are known that bind to the genome and open it. They're called pioneer factors. There are some well known pioneer factors like FoxA and some of the iPS reprogramming factors. And the idea is that they're able to bind to closed chromatin and to open it up to provide accessibility to other factors.

So what we would like to do, is to see if there's a way that we could, both understand how to discover those factors automatically, using a computational method, and secondarily, understand where factors are binding in a single experiment across the genome.

So the results I'm going to show you can be summarized here. I'm going to show you a method called PIQ that can predict where TFs bind from DNase-seq data that I'll describe in a moment.

We'll identify pioneer factors. We'll show that certain of these pioneer factors are directional and only operate in one way on the genome. And finally, that the opening of the genome allow subtler factors to come in and to bind to the genome.

So let's begin with what DNase-seq data is, and how we can use it to predict where TFs are binding to the genome. So DNase-seq is a methodology for exploring what parts of the genome are open. So here's the idea. You take your cell and you expose it, once you've isolated the chromatin to DNase-1 which will cut or nick DNA

at locations where the DNA is open.

You then can collect the DNA, size separate it and sequence it. And thus, you're going to have more reads where the DNA has been open, and less reads were it's protected by proteins.

So the cartoon below gives you an idea that, where there are histones-- each histone has about 147 bases of DNA wrapped around it. Or where there are other proteins hiding the DNA, you're going to cast shadows on this.

So we're going to be looking at the shadows and also the accessible parts, by looking directly at the DNase-seq reads.

So if we sequence deeply enough we can understand that each binding protein has its own particular profile of protection. So if you look at these different proteins, they cast particular shadows on the genome.

I'm showing here a window that's 400 base pairs wide. This is the average of thousands of different binding instances. So this is not one binding instance on the top row. You can see how CTCF and other factors have particular shadows they cast or profiles. Yes?

AUDIENCE: How do you know which factor was at which site? [INAUDIBLE].

PROFESSOR: How do we know which factor is at which site? By the motifs that are under the site. And what's interesting about CTCF is that you can actually see how it phase the nucleosomes. You can see the, sort of, periodic pattern in CTCF. And those dips are where the nucleosomes are. There's a lot you can tell from these patterns about the underlying molecular mechanism of what's going on.

Now, you can see at the very bottom, the aggregate CTCF profile. And if all the CTCF bindings looked like that, it'd be really easy. But above it, as I've shown you what an individual CTCF site looks like, you can see how sparse it is. We just don't get enough re-density to be able to recover a beautiful protection profile like that.

So we're always working against a lot of noise in this kind of biological environment. And so our computational technique will need to come up with an adequate model to overcome that noise.

But if we can, right, the great promise is that with a single experiment we'll be able to identify where all these different factors are binding to the genome from one set of data.

So, just reiterating now, if you think about the input to this algorithm-- we're going to have three things that we input to the algorithm. We input the original genome sequence. We input the motifs of the factors that we care about, that we think are interesting. And we input the DNase-seq data that has been aligned to the genome.

So those are the three inputs. And the output of the algorithm is going to be the predictions of which motifs are occupied by the factors, probabilistically. And in order to do that, for each protein we need to learn its protection profile.

And we need to score that profile against each instance of the motif to see whether or not we think the protein is actually sitting at that location in the genome. Any questions at all about that? No? OK. Don't hesitate to stop me.

So the design goals for this particular computational algorithm, as I said earlier, is resistance to low coverage and lots of noise. To be able to handle multiple experiment once, it has to work on the entire mammalian genome. It has to have high spatial accuracy and it has to have good behavior in bad cases.

So in order to model the underlying re-distribution of the genome, what we're going to do is something that is in principle quite straightforward. Which is that we're going to model all accounts that we see in the genome by a Poisson distribution.

So in each base of the genome, the counts that we see there in the DNase-seq data are modeled by a Poisson. And this is assuming that there's no protein bound there.

So what we're trying to do is to model the background distribution of counts without any kind of binding. And the log rate of that Poisson is going to be taken from a

multivariate normal. And the particular structure of that multivariate normal provides a lot of smoothing.

So we can learn from that multivariate normal how to fill in missing information. It's very important to build strength from neighboring bases.

So, even though we may not have lots of information for this base, if we have information for all the bases around us, we can use that information to build strength to estimate what we should see at this base if it's not occupied.

So the details of how we learn the mean and the sigma matrix you see up there for estimating the multivariate normal are outside the scope of what I'm going to talk about today. But suffice to say, they can be effectively learned.

And the second thing we need to learn are these profiles. And so each protein is going to have a profile. Here shown 400 bases wide. And it describes how that protein, so to speak, casts a shadow on the genome. And we judge the significance of these profiles-- and remember that one of my points was I wanted this to be robust.

So I will not make calls for proteins where I cannot get a robust profile that is significant above background. And I also exclude the middle region of the profile because it's been shown that the actual cutting enzymes are sequence specific to some extent. The DNase-1 cutting enzyme. And so we don't simply want to be but picking up sequence bias in our profile.

So we learn these profiles that describe for each particular motif-- and typically we can take in hundreds of motifs, over 500 motifs at once-- for each motif, what its protection looks like.

So what we then have-- we're going to learn this, actually, in an iterative process, but what we're going to have is-- now we have a model of what the unoccupied genome looks like. And we have a model of the reads that a particular protein at a motif location is going to produce.

And we can put those two things together and the way that we do that is that we have a binding variable. Showing there is delta. And we can either add or not add the binding profile of a particular protein in a location in the genome. And that will change the expected number of counts that we see.

So the key part of this is that we use a likelihood ratio shown as the second probability. It's not really a probability, it's a ratio, which is the probability of a count, given that a protein j is binding at that location, versus the probability of the counts, were it not binding. And that quantity is key because it's going to be-- once we log transform it, will be a key component of our test statistic to figure out whether or not a protein's binding at a particular location.

And so the way that we go about that is it we log that ratio and we add it to some other prior information that gives us an overall measure for whether or not the protein is binding at a particular location. And then we can rank these for all the motifs for that particular protein in the genome.

And then we can make calls using a null set. So we could look in the genome for locations that we know are not occupied, compute a distribution of that statistic, and then we can say, for what values of this statistic that we observe, at the actual motif sites, is it so unlikely that this would occur at random. At some desired p value by looking at the area in the tail of the null set.

So, just summarizing, we learn a background model of the genome, which is a Poisson that takes log rates from a multivariate normal. We learn patterns, or profiles of protection, or the production of reads for each motif. And at each motif location, we ask the question whether or not, it's likely that the protein was there and actually caused the reads that we're seeing, using a log likelihood ratio.

So what we're integrating together, when we take all these things, is that we're taking our original DNA seq-reads, we're taking our TF-specific specific binding profiles. We can build strength across experiments for the background model and we can also learn, to what extent, the strength of binding is influenced by the match of the position-- a specific weight matrix-- to a particular location in the genome. And

then we can produce binding calls. And when we do so, it works quite well.

So here you see three different mouse ESO factors. And the area under this receiver operating curve-- we've talked about this before. Remember a receiver operating characteristic curve-- has false positives increasing on the x-axis and true positives increasing on the y-axis. And if we had a perfect method, the area under that curve would be 1.0.

And so for this method, the area under the ROC curve for these three factors, using ChIP-seq data, is the absolute gold standard, is over 0.9.

And you might say, well that's great, but how well does it work in general? I mean, for example, the On-code project has used hundreds and hundreds of ChIP-seq experiments to profile where factors are binding in different cellular states.

If you take the DNase-seq data from those matched cell types and you ask, can you reproduce the ChIP-seq seq data? The answer is, a lot of the time we can, using this kind of methodology. And that is, the AUC mean is 0.93 compared to 313 different ChIP-seq experiments.

So this methodology of looking at open chromatin allows us to identify where lots of different factors bind to the genome. And about 75 different factors are strongly detectable using this methodology. So it's detectable if it has a strong motif, if it binds in DNase-accessible regions and has strong DNA-binding affinity.

So I tell you this just so you know that there are new methods coming that allow us to take a single experiment and analyze it and determine where a large number of factors bind from that single experimental data set.

Now, a second question we wanted to answer was, how is it that chrome, opening and closing is controlled? And since we had a direct read out of what chromatin is open, because reads are being produced there, we could look in a experimental system where we measured chromatin accessibility through developmental time.

And the idea was that as we measured this accessibility, we could look at the places

that changed and determine what underlying motifs were present that perhaps were causing the genome to undergo this opening process.

So we developed an underlying theory that pioneer factors would bind to closed chromatin as shown in the middle panel and open it up, and that we could observe those by looking at the differential accessibility of the genome at two different time points that were related.

And we couldn't observe pioneers they didn't open up chromatin. And for non-pioneers-- obviously the left-hand panel-- they would not, in our design here, lead to increased accessibility.

So we then looked at designing computational indices that measured the-- oh, question, yes?

AUDIENCE: When you say pioneer factors, are you looking at what proteins are pioneer factors, or are you looking at what sequences they bind to that are [INAUDIBLE].

PROFESSOR: So the question is, are we looking at what proteins are factors, or are we looking at what sequence, right? What we're doing is, we're making an assumption that the underlying sequence denotes one or more proteins and thus, we are hypothesizing, there's the proteins that are actually binding to the sequence, that's causing that. And then later on, we'll go back and test that experimentally, as you'll see in a second. OK?

So here there are three different metrics, which is the dynamic opening of chromatin from one time point to the next, the static openness of chromatin around a particular factor, and a social index showing how many other factors are around where a particular factor binds.

And you can see that these things are distributed in a way that certain of the factors have a very high index in multiple of these scores. And thus, we were able to classify a certain set of factors as what we classified as computational pioneers, that would open up the genome.

Now, in any kind of computational work, we're actually looking at correlative analysis, which is never causal. Right. So we have to go back and we have to test whether or not our computational predictions are correct.

So in order to do that, we built a test construct where we could put the pioneers in on the left-hand side and ask, whether or not the pioneer would open up chromatin and enable the expression of a GFP marker. And the red bars show the factors that we thought were pioneers.

And as you can see, in this case, all but one of the predictive pioneers produces GFP activity. And this construct was designed in an interesting way. We had to design it so that the pioneers themselves were not simply activators.

And so it was upstream of another activator, which is a retinoic acid receptor site. And so in the absence of retinoic acid receptor, we had to ensure that when we turned on the pioneer, GFP was not turned on. It was only with the addition of the pioneer to open the chromatin and the activator that we actually got GFP expression.

OK. So, through this methodology we discovered about 120 different motifs corresponding to proteins that we found computationally open-- chromatin out. Yes?

AUDIENCE: [INAUDIBLE] concentrations of different pioneer factors are different, wouldn't that show up differentially [INAUDIBLE]?

PROFESSOR: The question is, if the concentration of different pioneer factors was different, wouldn't that show up differentially? And that's precisely, we think how chromatin structures are regulated.

That we think that the concentration, or presence of different pioneer factors, is regulating the openness or closeness of different parts of the genome, based upon where their motifs are occurring. Is that, in part, answering your question?

AUDIENCE: Yes, but, if a concentration of a particular pioneer factor is low, do they necessarily have lesser binding sites on the genome?

PROFESSOR: So you're asking, how is the concentration of a pioneer factor related to its ability to open chromatin and whether or not a higher dosage would open more chromatin?

AUDIENCE: Yes.

PROFESSOR: I don't have a good answer to that question. Those experiments haven't been done.

However, one thing you may have noticed about these profiles-- remember these are the same profiles that we talked about earlier of DNase-1 read reproduction around a particular factor. And what you might notice is that some of these profiles are asymmetric. And that they appear to be producing more region one direction than the other direction.

And so this is all computational analysis, right. But when you see something like that you say, well gee, why is that going on? Why is it that for NRF-1 the left-hand side has a lot more reads than the right hand side.

Now, of course, the only reason that we can produce an oriented profile like that is that the NRF-1 motif is not palindromic, right. We can actually orient it in the genome and so we know that the more reads, in this case, are coming from the five prime end then from the three prime end.

So what do you think would cause that? Does anybody have a-- when we first saw this, we didn't know what it was. But anybody have an idea of what that could be? Oh, yes.

AUDIENCE: It's the remodelers that these transcription factors are calling in tend to open the chromatin more on one side of the motif than the other.

PROFESSOR: Right, so if the remodelers are working in some sort of directional way, right. So that's what we thought. We didn't know whether they were or not. And so we went back to our assay and we tested the motifs, both in the forward and the reverse direction. Right.

To see whether or not it mattered which way the motif went into the construct,

based upon selecting factors, based upon a symmetry score that we computed for their read profile, right?

And what we found was that, in fact, it was the case that when the motif was properly oriented it would turn on GFP and was in the other direction it would not.

So it appeared, for the factors that we tested, that they did have directional chromatin opening properties. And so that's an interesting concept that you actually can have chromatin being opened in one direction but not the other direction, because it admits the idea of some sort of genomic parentheses, where you could imagine part of the genome being accessible where the other part is not.

And overall this led us to classifying protein factors that are operating in genome accessibility into three classes. Here shown as two, where we have pioneers which are the things that open up the genome, and settlers that follow behind and actually bind in the regions where the chromatin is open.

That is, it's much more likely that those factors are going to bind where the doors of the rooms are open, and the pioneers are the proteins that come along and open the doors, in particular, chromatin domains.

And there were a couple of other tests that we wanted to do. We wanted to test whether or not we could knock out this pioneering activity by taking a pioneer and just only including its DNA-binding domain and knocking out the rest of its domain which might be operative in doing this chromatin remodeling.

And then asked, whether or not, when we expressed this sort of poisoned pioneer, whether or not it would affect the binding of nearby factors. And, in fact, when you do express the sort of poison pioneer, it does reduce the binding of nearby factors.

Here, we have a dominant negative for NFYA and dominant negative for NRF1. It reduces the binding of nearby factors. And finally, we wanted to know, if we included a dominant negative for the directional pioneer, if it actually would preferentially affect the binding of [INAUDIBLE] on one side of its binding occurrences or the other side.

And so we looked at mix sites that were oriented with respect to NFYA. And when we add the NFYA, you can see that it actually-- the dominant negative NFYA-- when the mix site is down of where we think NFYA is opening up the chromatin, the binding is substantially reduced. Whereas, when the Myc site is not on the side where we think that NFYA is opening, it doesn't really have an effect.

So this is further confirmation of the idea that in vivo, these factors are actually operating in a directional way.

Now I tell you all this because, you know, we do a lot of computational analysis and it's important to follow up and understand what the correlations tell us. So when you do computational analysis and you see a very interesting pattern, the thing to keep in mind is, what kind of experiment can I design to test whether or not my hypothesis is correct or not?

We also did an analysis across human and mouse data sets and found that for a given motif, and thus, protein family, it appeared that the chromatin opening index was largely preserved, evolutionarily. So that there are similar pioneers between human and mouse.

Are there any questions at all about the idea? So I told you, I mean, when you go to cocktail party tonight, you say hey, you know, did you know that DNase-seq is this really cool technique that not only tells you whether or not chromatin is open or not, but, you know, where factors bind?

And some of those factors open up the chromatin itself and, plus, get this, some of the factors only do it in one direction, right. That'd be a good conversation starter, right? That'd be the end of the conversation, no. You get the idea, right. So are there any questions about DNase-1 seq analysis? Yes?

AUDIENCE: A little unrelated, but I was just wondering-- in the literature where people have identified factors that neither directly reprogram between different cell types, or go through some sort of [INAUDIBLE] intermediate--

PROFESSOR: Yes.

AUDIENCE: There are a number of transcription factors that have been identified. [INAUDIBLE] but there are others. Do you often see, or always see some of the pioneers that you've identified in those cases. And then--

PROFESSOR: Yes.

AUDIENCE: And then, a follow-up question would be, do you think that if you took some of the pioneers that you generated that were not known before and expressed them in cell types, that they would open up the chromatin sufficiently to potentially reprogram the mistakes?

PROFESSOR: Right. So the question was, is it the case that known reprogramming factors, at times are powerful pioneers? The answer is yes.

The second question was, now that you have a broader repertoire of pioneer factors, and you can identify what they're doing, is a possible to, in a principled way, engineer the opening of chromatin by perhaps expressing those factors to see whether or not you could match a particular desired epigenetic state, let's say?

Our preliminary results are yes on the second count as well. That there appear to be pioneer factors that operate, sort of at a basal level that keep, sort of, the sort of usual rooms open in the genome.

And then there are factors that operate in a lineage-specific specific way. And when we express lineage-specific pioneer factors, they don't completely mimic but largely mimic the chromatin state that's present in the corresponding lineage committed cells. And so we think that for principal reprogramming of cells, the basal level of establishing matched open states is going to be an interesting and important avenue to explore. Does that answer your question? Yeah. OK.

So, now we're going to turn to another-- well let me just first summarise what I just told you about, which is that we can predict where TFs bind from DNase-seq data. We can identify these pioneer factors. Some of them are directional. And other

factors follow these pioneers and bind sort of in their wake. In where they are actually open up the chromatin.

And returning to our narrative arc for today, we've talked about the idea of histone marks. We've talked about the idea of chromatin openness and closeness. And now I'd like to talk about the important question of how we can understand which regulatory regions are regulating which genes.

Now the traditional way to approach this, is that if you have a regulatory region, the thing that you do is you look for the closest gene. And you go, aha, that's the one that that regulatory region is controlling. This applies not only for regulatory regions but for snips, right. If you find a snip or a polymorphism you are likely to assume that it's regulating the closest gene. It could have an effect on the closest gene.

But there are other ways of approaching that question with molecular protocols. And drawing you once again a cartoon of genome looping, you can see how an enhancer is coming in contact with the Pol II holoenzyme apparatus. And this enhancer will include regulators that will cause Pol II to begin transcription.

And if somehow we could capture these complexes so that we could examine them and figure out what bits of DNA are associated with one another, we could map, directly, what enhancers are controlling what genes, when they're active in this form.

So the essential idea of a variety of different protocols, whether it be protocols like high c or ChIA-PET that we're going to talk about are the same. The difference is that in the case of ChIA-PET, we're only going to look at interactions that are defined by a particular protein.

So what we're going to do in the slides I'm going to show you today, is we're going to only look at interactions that are mediated through RNA polymerase II. And those are particularly interesting interactions as you can see, because they involve actively transcribed genes. So if we could capture all the RNA polymerase II mediated interactions, we'd be in great shape.

So, we have a lot of very talented biologists here. So would anybody like to make a suggestion for a protocol for actually revealing these interactions? Does anybody have any ideas how you'd go about that? Or what enzyme might be involved? Any ideas? Don't be bashful now. Yes.

AUDIENCE: How about fixing everything in place where it is and then getting [INAUDIBLE] through DNA.

PROFESSOR: OK. Fixing everything where it is in place. That's good. So we might cross link this whole thing, for example. OK. And then any other ideas what we would do? That's done, this protical-- yes.

AUDIENCE: Well, [INAUDIBLE] that you've going to be [INAUDIBLE]. And then digesting the DNA that's coming out, and then that lingers to the DNA that are closest together in the sequence.

PROFESSOR: OK. So I think what you're suggesting goes something like this. All right. Which is, that imagine that we cross link those complexes and we precipitate them. And then what we do is we, in a very dilute solution, we ligate the DNA together.

And so we get two kinds of ligation products. On the left-hand side we get self-ligation products where a DNA molecule ligates to itself. And on the right-hand side we get inner ligation products, where the piece of DNA that the enhancer was on, ligates to the pieces of DNA that the RNA polymerase was transcribing the gene on.

And those inter-ligation bits of DNA, the ones that are red and blue, are really interesting, right. Because they contain both the enhancer sequence and the promoter sequence.

And all we need to do now is to sequence those molecules from the ends and figure out where they are in the genome. Yes?

AUDIENCE: How much variation would there be in the sequence? I guess I'm just wondering-- the RNA polymerase is not static, is it? In terms of its interaction with the intenser and the gene. I just don't know what would be capturing in this--

PROFESSOR: Right.

AUDIENCE: [INAUDIBLE] doesn't just touch at the beginning and then [INAUDIBLE].

PROFESSOR: Right. And I think that's a very good question. And in fact, a PhD thesis was just written on this topic. Which is, when you have proteins that are moving down the genome, in some sense, you're looking at a blurred picture.

So how do you de-blur the picture so that it's brought sharply into focus? And so a compute is something called a point spread function which describes how things are spread out down the genome. And then you invert that to get a more focused picture of where the protein is actually, primarily located.

But you're right. Things like RNA polymerase II are not thought of as point-binding proteins. They're actually proteins in motion most time when they're doing their work.

AUDIENCE: [INAUDIBLE] that it's polymerizing, does that it mean that it's still continually bound to the [INAUDIBLE]?

PROFESSOR: No. Although, I don't think we really understand all of the details of that mechanism. But, suffice to say that what I can do is I can start showing you data and from the data we can try and understand mechanism. These are all great questions, right. Yes.

AUDIENCE: When we did the citations and ligation, you're going to get a lot of random ligation, right?

PROFESSOR: A lot of random ligation?

AUDIENCE: Yeah, between DNA sequences that aren't aren't, I guess, as close? Or you shouldn't really be ligating certain things?

PROFESSOR: Well, this picture is a little bit deceiving, right? Because there's actually another complex just like the one at the top, right to its left, right? And you could imagine those things ligating together. And so now you're going to get ligation products that

are noise. They don't mean anything.

AUDIENCE: Do you just throw those out, I guess?

PROFESSOR: Well, the problem is, you don't know which ones are noise and which ones aren't. Right? Now, there are some clever tricks you can play.

One clever trick is to change the protocol to do these kinds of reactions, not in solution, but in some sort of gel or other thing that keeps the products apart. The other thing you can do is estimate how bad the situation is. And how might you do that?

What you do is, you take one set of-- you take your original preparation and you split it into two. OK. And you color this one red and this one blue using linkers, right. And then you put them together and you do this reaction.

And then you ask, how many molecules have the red and the blue linkers on them. And then you know those are bad ones because they actually came from different complexes, right.

And so by estimating the amount of critical chimeric products you get, from that split and then recombined approach, you can optimize the protocol to reduce the chimeric production rate.

Current chimeric production rates are about 20%. Something of that order. OK. It used to be 50%, that's really bad. OK. So you can try and optimize that.

Now, if the protocol has these issues-- you have a moving protein that was brought up here, right, that you're trying to capture. You've got a lot of noise coming from the background of these reactions, right.

Why are we doing this? Well, it's the only game in town right now. If you want to have a mechanistic way of understanding what enhancers are communicating with what genes, this and its family-- I broadly call this a family of protocols-- is really the only way to go. OK.

The interesting thing is that when you do, you get data like this. And so, what you're looking at here is exactly the same location in the genome. It's about 600,000 bases across from left to right. OK.

And at the very bottom, you see the SOX2 gene. And you have three different cellular states. The top state-- our motor neurons have been programmed through the ectopic expression of three transcription factors.

The second set of interactions are motor neurons that have been produced by exposure to small molecules over a 7-day period.

And the bottom set of interactions are from mouse ES cells that are plueripotent. And what's interesting is that you can see how-- I'm going to point here.

You can see here-- this is the SOX2 gene down at the bottom. And you can see here-- this regulatory region is interacting heavily with the SOX2 gene at the ESL state. And above here, I have put SOX2 ChIP-seq data. So you can actually see that SOX2 is regulating itself.

And up here, we have the same SOX2 gene locus. And OLIG2 is a key regulator of this motor neuron fate. And you can see that it appears that OLIG2 is now regulating SOX2.

And we don't have as complete dependence upon the SOX2 locus as we had before. And up here in the induced motor neuron state, LHX4 is one of the reprogramming factors and you can see how it is interacting with SOX2 here and over here.

So what this methodology allows us to do, is to tie these regulatory regions to the genes that they are regulating, albeit it with some issues.

So, we'll talk about the issues in just a second. Are there any questions at all about the idea of capturing, in essence, the folding of the genome with this methodology to link regulatory regions to genes? Yes?

AUDIENCE: I have a question. So in each of those charts you've got parts describing regions that are interacting.

PROFESSOR: Yes.

AUDIENCE: Is that correct?

PROFESSOR: Yes. The little loops underneath are the actual read pairs that came out of the sequencer. And the green dotted lines are the interactions I'm suggesting are significant.

So I'm showing you the raw data and I'm showing you the hypothesized or purported interactions with the green dotted lines. Right? Right?

AUDIENCE: So how is you raw sequencing then transformed into this set of interactions?

PROFESSOR: How is the raw sequencing data-- remember that what came out of the protocol were molecules on the right-hand side that had little bits of DNA from two different places in the genome.

AUDIENCE: I'm sorry, I meant, how did you determine-- because I'm assuming each of these arcs has to have a single base start side and a single base end site.

PROFESSOR: Correct.

AUDIENCE: However, your reads are going to span-- your joined paired reads are going to span a number of bases. So you have a number of bases coming from the red part and a number of bases coming from the blue part.

PROFESSOR: We've got 20, 20 something, yeah.

AUDIENCE: How do you determine which of these red bases and which of these blue bases are your start and end points for the [INAUDIBLE].

PROFESSOR: Well, you are looking at a 600,000 base pair window of the genome and we're not quite at the resolution of 28 bases yet.

AUDIENCE: OK.

PROFESSOR: So, you know--

AUDIENCE: So this is not necessarily single base pair resolution, but this is a region resolution? Is that correct?

PROFESSOR: Once again, the question of how to improve the spatial resolution of these results is a subject of active research. And once again, you can deconvolve things like the shearing to actually get things down to within, say, 10 to 100 base pairs resolution.

AUDIENCE: OK.

PROFESSOR: OK?

AUDIENCE: Got it.

PROFESSOR: But you can't identify the exact motif that the things land on, right. They can get in the ballpark, so to speak, right. You can figure out where you need to look for motifs.

And so one thing that we and others do is look at these regions and we ask what motifs are present into these regions. Or if you have match DNase-seq data, you can go back and you can say, aha, I have DNase-seq data.

I have this data and I know that there's something going on at that region of the genome. What proteins do I think are sitting there, based upon the protection profiles I see. Right.

So you can take an integrative approach where you use different data types to begin to pick apart the regulatory network. Where you see the connections directly molecularly, and you see the regulatory proteins that are binding at those locations. OK? Was that helpful? Good. Good questions. Any other questions? Yes?

AUDIENCE: Would you consider Hi-C and 5C and all of those to be the same family of technique?

PROFESSOR: I would. They're all, sort of the same family and they're improving. I'm about to tell you why this doesn't work very well. But, that said, it's the best thing we have going. Right. 5C is not any to any. It's to one to any.

This protocol, when you do one experiment with this, it tells you all the interacting regions in the genome. Right. I believe 5C-- help me if I'm wrong. You pick one anchor location and then you can tell all the regions and genomes that are interacting with that anchor location.

AUDIENCE: Isn't that 3C?

PROFESSOR: What?

AUDIENCE: 3C's one to one. 4C's one to any.

AUDIENCE: And 5C is--

AUDIENCE: 5C's any to any.

PROFESSOR: And 5C's any to any? OK. I stand correct. Thank you. Yeah. OK. You didn't critique my bond type. See I was trying to get you and you didn't. OK. And other questions about this? OK.

What could go wrong? What could go wrong? Well, I can tell you what will go wrong. What will go wrong is that it has a low true positive rate. OK.

And how can you tell that? You do the experiment twice and you get thousands of interactions from each experiment in exactly matched conditions and there's a very small overlap between the conditions. Oops.

So, that's a pretty big oops, right? Because you would like it to be the case that when you do an experiment multiple times, you get the same answer.

So let us just suppose that you get 10,000 interactions in experiment one. 10,000 interactions in experiment two, but only 2,000 of them are the same.

What could possibly be going wrong? Any ideas? If you're looking at the data, what

would you think? Well? Yeah?

AUDIENCE: [INAUDIBLE] could be really high, so you're just seeing a couple of things that are above the background. And they don't necessarily--

PROFESSOR: Right. So is it maybe that, you know, it's just tough to get these interactions out. And so you got a lot of background trash. And the things that are significant are tough to pick out. Yeah?

AUDIENCE: Maybe it's a real biological noise issue? So rather than the technique, actually any given time that the interactions as so diverse that when you take the snap shot you can't--

PROFESSOR: I like that explanation because it's very pleasing and makes me feel good. And I would be hopeful that that would be true that there's enough biological noise that that's actually what I'm observing. It doesn't make me feel too warm and fuzzy, but you know, I'd go with that, right.

The other thing you might think is, gee, if we just sequenced that library more, we'd get more interactions out them, right? So you go off and you compute the library complexity of your library and you go, oops, that's not going to work.

There just isn't enough diversity in the library. Meaning that the underlying biological protocol did not produce enough of those interesting inner ligation events to allow you to reveal more information about what's going on. OK.

Now if I ask you to judge the significance of an interaction pair here. Let's think about this using what we know already from the subject. OK.

So I'm going to draw a picture. So I have my genome. And let's just say that I have a location, CA and a location CB and I have a pile of ends that wind up in those two locations. OK.

And what I would like to know is-- and I have, let me just see what variable I used for this. And I have a certain number of interactions between a and b. That is I have a certain number of reads that cross between these two locations in the genome.

And I'd like to know whether or not this number of reads is significant. OK. How could I estimate that? Any ideas? Oh, I'm also going to tell you that n is the total number of read ends observed. OK.

Well, here is the idea. I've got n total read ends, right? I've got ca read ends here. I've got cv read ends here, and I have iab that are overlapping.

So now, this is just our old friend, the hypergeometric, right. We can ask what is the probability of that happening at random? This many interactions or fewer would happen at random.

And if it's very unlikely, we would reject the null hypothesis and accept that there's really an interaction going on here. OK? So, just to be more precise about that. This is what it looks like. You've seen this before.

That the probability of those interactions happening on a null model, given a total number of interactions end in ca and cb is given by the hypergeometric. OK.

So that's one way of going about assessing whether or not the interactions we see are significant.

Now, let me ask you a slightly different question. Right. Imagine that I have-- and I'm being very generous here. Imagine that I have two experiment-- that's the wrong size bubbles. I don't want to mislead you.

One of your friends comes to you and say, "I've done this experiment twice." Twice, OK. "And each time I get 1,000 interactions. So each one gives you 1,000, let's say.

And I have 900 that are common between the two replicates. And your friend says, "how many interactions do you think there are in total?"

How could we estimate that? Well, what's interesting about this problem is that what we're asking is what's n? Right.

What's the total number of interactions of which we're observing this set and this set

of which 900 is overlapping. There's the hyperlink geometric again.

So all we need to do is to find the maximum value, the best value for n that predicts the observed overlap given that we have two experiments of size, with m and n different observations, and we have an overlap of k. OK.

Does that makes sense to everybody? Of how to estimate the total number of interactions out there making a set of assumption that they're all equally likely. Any questions about that at all? OK.

And, just so you know, you can approximate this, this way. Which is that the maximum likelihood estimate of the total number of interactions is approximately n times n over k, as seen by the approximation on the bottom. OK? Just so that you can approximate how many things are out there that you haven't seen when you've done a couple of replicates.

OK, you guys have been totally great. We've talked about a lot of different things today in chromatin architecture and structure. Sort of the DC to light version of chromatin structure and architecture lecture.

Next time we're going to talk about building genetic models of EQTLs. And the time after that we're going to talk about human genetics.

Thank you so much. Have a great, long weekend. We'll see you next Thursday.