Next Generation Sequencing Technologies: Data Analysis and Applications

Genome-wide Transcription Factor (TF) Binding Sites

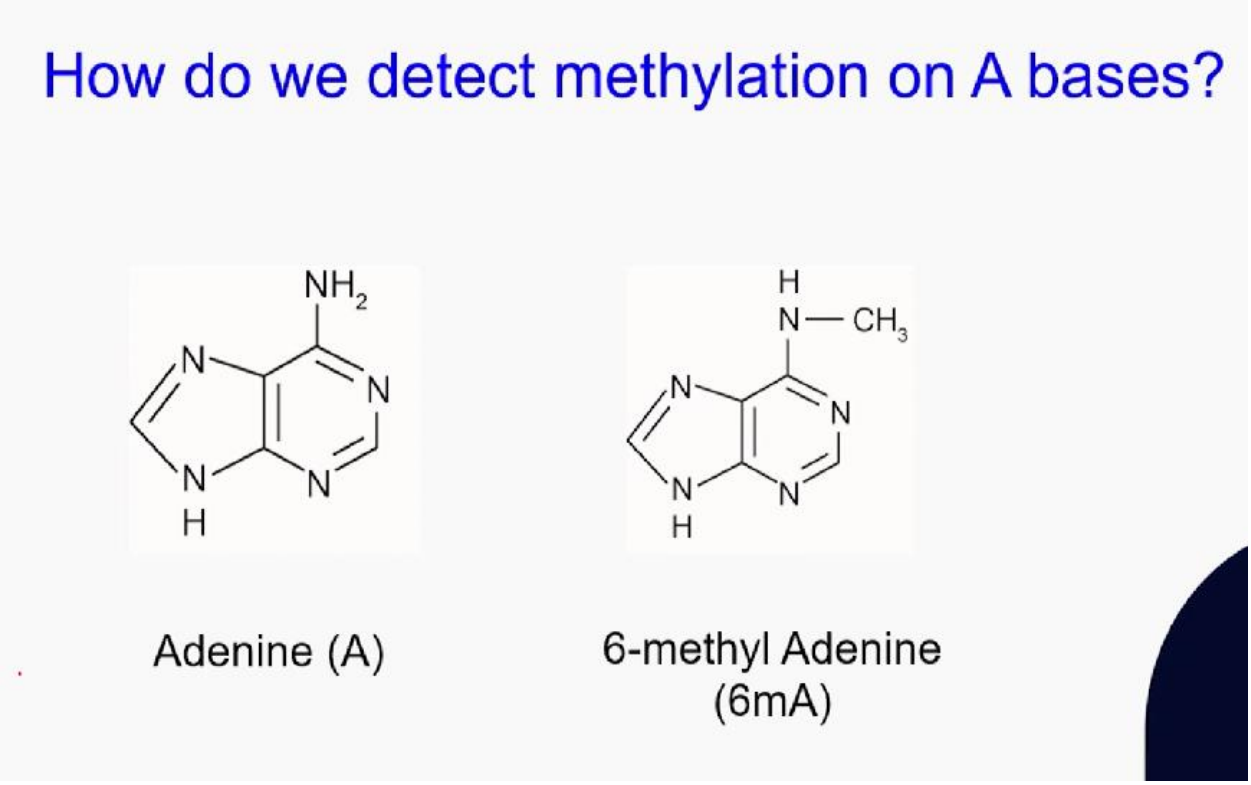
Dr. Riddhiman Dhar, **Department of Biotechnology**

Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the course on Generation Sequencing Technologies, Data Analysis, and Applications. In the last class, we discussed cytosine methylation and different methods that could detect the methylation of cytosine. We also talked about different methods that could distinguish between two different types of modifications on cytosine, for example, the methylation right and the hydroxy methylation. So, we talked about different methods that could do that. So, one of the questions that still remains is: how do we actually identify the methylation pattern on adenine? So, on A base, ok. So, this is something that we have not talked about so far, and it turns out that this was not so focused on because these were actually not very common in the human genome. So, in this class, we will talk about methods that could allow us to identify these 6 MA modifications or 6 MA bases, and then we will talk about different methods that could actually help us identify the transcription factor binding sites across the genome. So, this is something that is very important because all these methylations actually impact the binding of transcription factors, as we will see in this presentation. So, these are the two topics that will be discussed in this class. So, direct detection of DNA methylation patterns would actually allow us to identify the six MA bases, and then we will talk about transcription factor binding site identification using a method called chromatin immunoprecipitation, or QI. These are the keywords for this class: inter-pulse duration, pulse width, and nanopore, and then we will also see immunoprecipitation. So, just to summarize what we know about DNA methylation so far, these are methylations on A or C bases, mostly, of course; we have talked about some other modifications as well, and we have talked about the detection of 5 mC and 5 hmC bases.

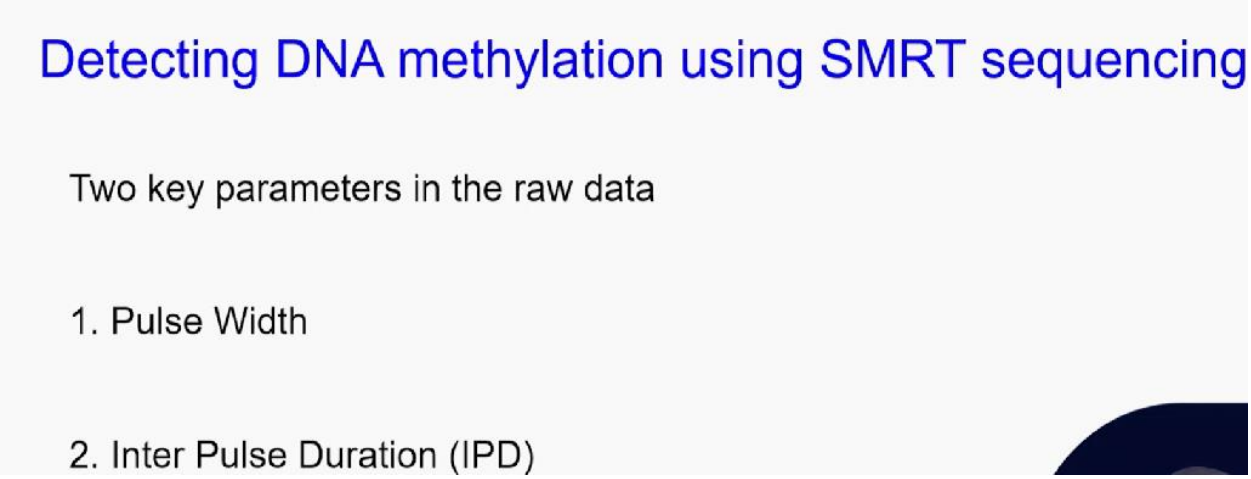
These are modifications on cytosine, and we specifically talked about bisulfite sequences that could actually easily allow us to identify these modifications at the genome level. So, across the whole genome, we also talked about tap sec and OXPS sec, which actually allow us to distinguish between 5 mC and 5 hmC modifications because these are important for certain applications. So, the question now is: how do we detect methylation on A bases? This is something that we want to

identify, and just to remind you, the modification that we see on A is called 6 mA. So, the modification of A is very common in prokaryotes, and it is thought to be uncommon in eukaryotes, especially in mammalian cells.



However, recently, papers have shown that 6-mA methylation is also present in the human genome, and they actually mark actively transcribing regions, and the levels of these modifications change in human diseases. So, here is the reference; if you are interested, you can go and read further. So, how do we detect these 6MA bases in the genome? So, it turns out we will have to resort to something called direct detection of DNA methylation, and there are two sequencing technologies that actually allow us to detect this DNA methylation directly. So, remember, in bisulfate sequencing, we need to do a treatment with sodium bisulfite, and this is quite a harsh treatment for DNA. You can also have other sources of damage to the DNA and other bases. So, if we can avoid that, and if we can go about detecting these methylation patterns just by sequencing, that would be really ideal, ok? And it turns out two methods actually offer that opportunity: the first is single-molecule real-time, or SMRT, sequencing, and the second is nanopore sequencing. So, we will discuss a little bit about these methods and how they allow us to detect this DNA methylation directly from the sequencing data. And between these two methods, what is common between them is that they are single-molecule sequencing. So, they are

sequencing a single molecule, and we also get real-time data, unlike the earlier methods. So, let us now see how we can apply these methods and how they have been applied to actually identify DNA methylation in practice. So, first, we will talk about SMRT sequencing and how we can apply this method to detect DNA methylation. So, if you remember SMRT sequencing, the sequencing happens at the bottom of the zero-mode wave guides, or ZMWs, right? So, you have the DNA polymerase attached to the bottom of those ZMWs, and the DNA molecule comes in OK. So, the polymerase is there, and the DNA molecule comes along with the sequencing of the complementary strand. The synthesis of the complementary strand happens, and the nucleotides come and bind, and these nucleotides are attached to fluorophores. And this means when the nucleotides come and are attached to the DNA, they are held in place by the polymerase, and this is where the signal detection happens, ok? Now, since this is single-molecule real-time sequencing, the time of addition of this individual base is recorded in the raw data, ok? So, you can find out from this data right at which time these additions are happening. And then, if you know the time, you can look at the interval between two base addition events. And you can also look at another parameter, which is the time or duration of the signal. So, the time for which you can actually detect the signal is okay. So, these are the two parameters.



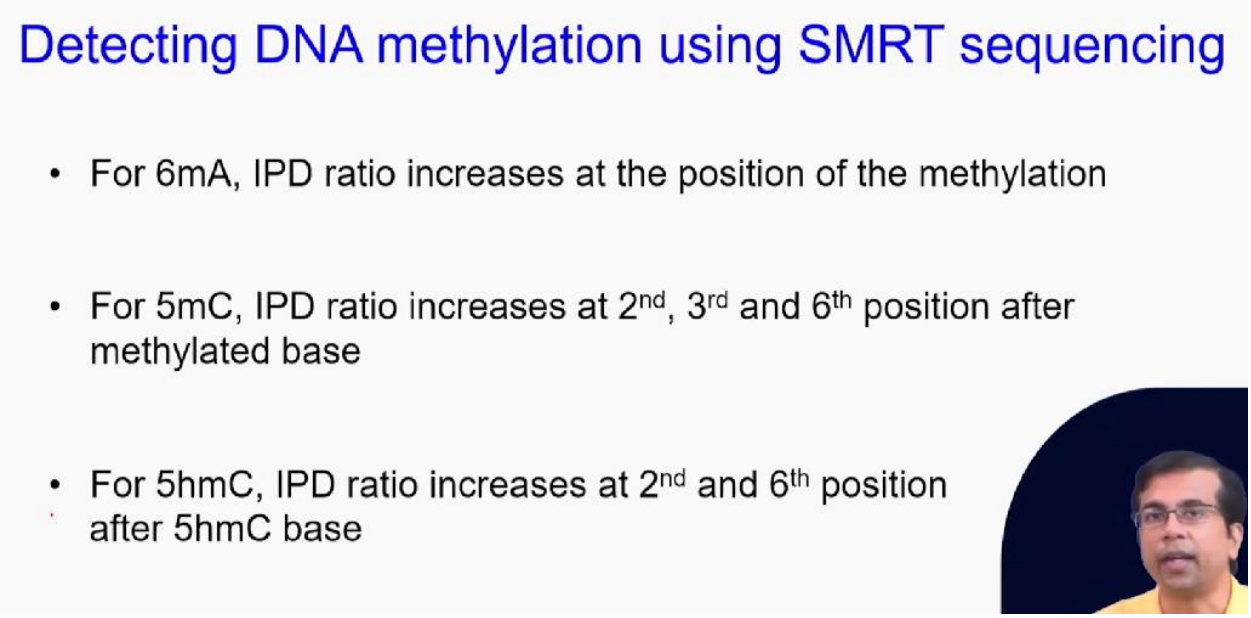
The technical terms for those are: the first one is pulse width. So, a pulse is the addition of a base, which is the event right and the width of that pulse, or the time interval for which we can detect the fluorescence signal. So, let us say you are adding an A base, and you can say, Let us say the detector can detect this A base fluorescence for, let us say, a few milliseconds. So, that would be the pulse width, and the second is inter-pulse duration, or IPD. So, this is the time interval between

two consecutive pulses, ok? So, you have let us say addition of base A and then addition of T, ok? So, the time interval between these two base addition events is called the inter-pulse duration. So, you can then observe this interpulse duration for the full process, and this data is stored in the raw data. So, using these two data sets now, we can actually see that if you have methylation, there is an increase in interpulse duration.

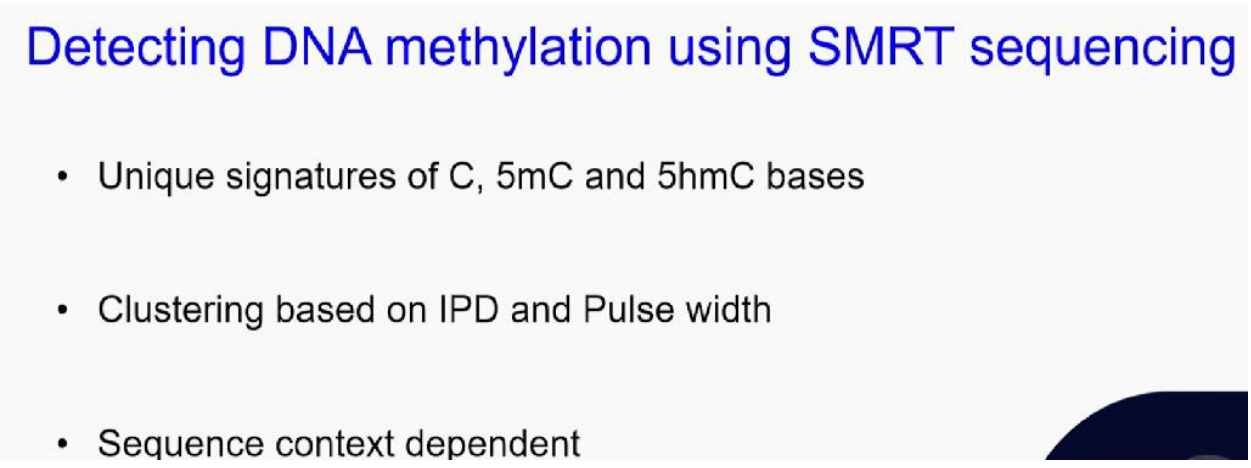
So, here is the paper that actually looks at the direct detection of DNA methylation using this SMRT sequencing, and what they observe is that if you have methylation on A, for example, you can see this difference in interpulse duration, and there will be an increase in interpulse duration. So, if you have a base and there is methylation, the time taken would be longer for the next base to come. So, of course, you do not understand why that is the case; perhaps this is something to do with the steric hindrance of the structure of the whole thing, etcetera. So, it turns out that this method can detect 6 mA as well as 5 mC and 5 hmC. So, you can imagine that these chemical structures are different.

So, they affect the signal in different ways, or these parameters in different ways, which we can then identify for identifying these methylation patterns. And what we look at is the IPD ratio between the sample with methylation and the control. This is what this paper actually read, and they actually looked at this modification, which is a certain specific genomic context or specific sequence context, and then looked at these IPD ratios between the sample with methylation and the control without methylation. And what they observed is that for 6 mA, there is an increase in this IPD ratio, and this increases at the position of the methylation. So, where this methylation is present, there is an increase in the IPD ratio in the sample compared to the control. Now, for 5 mC, this IPD ratio increases to the second, third, and sixth positions after the methylated basis.

Again, we still do not understand why that is the case and how they influence this base, like IPD values, at these positions after the methylation, but this is what they observed. Again, this is with respect to a specific sequence context. So, it might be the case that if you change the sequence context, these values might change. So, again, before you actually apply these methods, you probably have to optimize and look at these controls. For 5 hmC, there was a small difference.

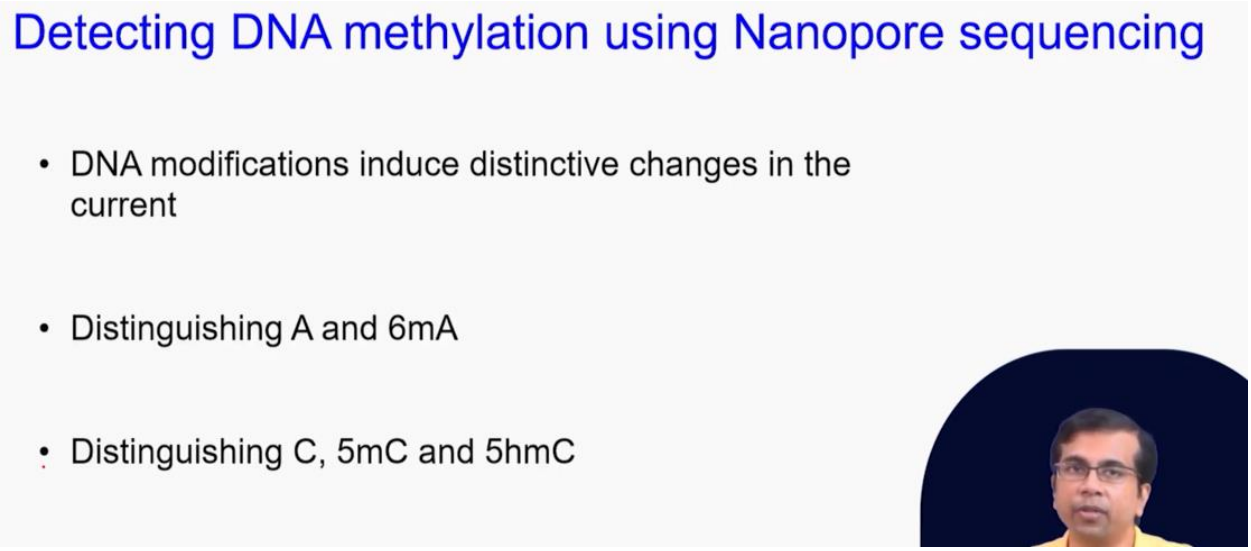


So, the IPD ratio actually increased at the second and sixth positions, but not at the third position after the 5 hmC base. So, what they could do is actually differentiate between unmethylated C, 5 mC, and 5 hmC bases based on these two parameters that we just described. So, with these IPD and pulse width parameters, if they were used, they could actually cluster out these unmethylated cytosines at 5 mC and 5 hmC bases, and they could see these different clusters based on which they could identify the modification. But what I have just mentioned is that this is dependent on the sequence context, right? So, if you change the sequence context, it is likely that the signals, IPD ratios, etcetera, will change.



Now, let us move on to nanopore sequencing. sequencing, can you apply nanopore sequencing to

detecting DNA methylation? So, if you remember, in nanopores, you have these channels right through which the DNA molecule passes, and it actually changes the current. So, there is a current at the picoampere level going through the nanopore, and the moment DNA enters the nanopore, it changes the resistance, which actually alters the current. This change in the current signal is detected by the detector, and then the base call happens okay. We talked about this in detail in the beginning, ok? So, what happens is that if you have DNA modifications right these are actually unique again unique signals chemical signals right.

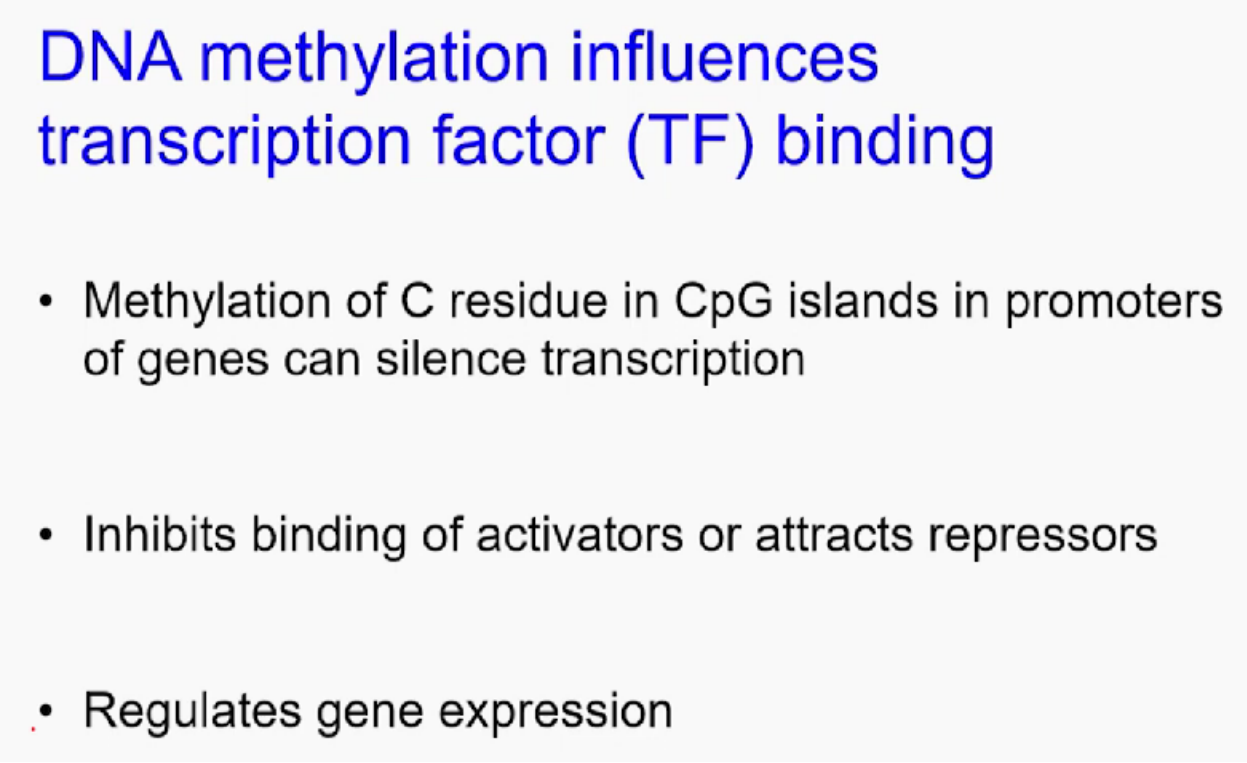


So, they actually induce distinctive changes in the current system. So, the moment you change the chemical structure, the resistance will be different, which will induce distinctive changes in the current, and this has actually been used for distinguishing A and 6 mA as well as C, 5 mC, and 5 hmC bases. So, here are two papers that are looking at this detection and mapping with nanopores for these different types of modifications. So, you can go and look into their data and see how these current signals are different from each other. Now, what we have learned about DNA methylation detection is that traditional methods for detecting 5 mC and 5 hC require chemical treatment of DNA, and that might be harsh in some cases.

It also increases the experimental complexity as well as the complexity of the data analysis. So, if you have to do this three-way comparison between samples, that actually increases the complexity of the data analysis. What we have seen is that SMRT and nanopore sequencing actually enable

us to directly detect these DNA methylation patterns, and they can differentiate between different types of methylation patterns. But, of course, the accuracy is not actually perfect. So, there is a lot of scope for improvement in these direct detection methods, but nevertheless, they provide direct identification methods.

Now, why are we interested in detecting these DNA methylation patterns? So, one thing we have learned is that these methylations are important for influencing gene expression. For example, we have seen that methylation of C residues in CpG islands in promoters of genes can silence transcription. This has been observed for a long time, and what we know is that this methylation actually inhibits the binding of activators that will induce gene expression, or it attracts repressors that actually block this expression. So, this means that DNA methylation is actually intricately linked with gene expression. So, they are controlling gene expression.

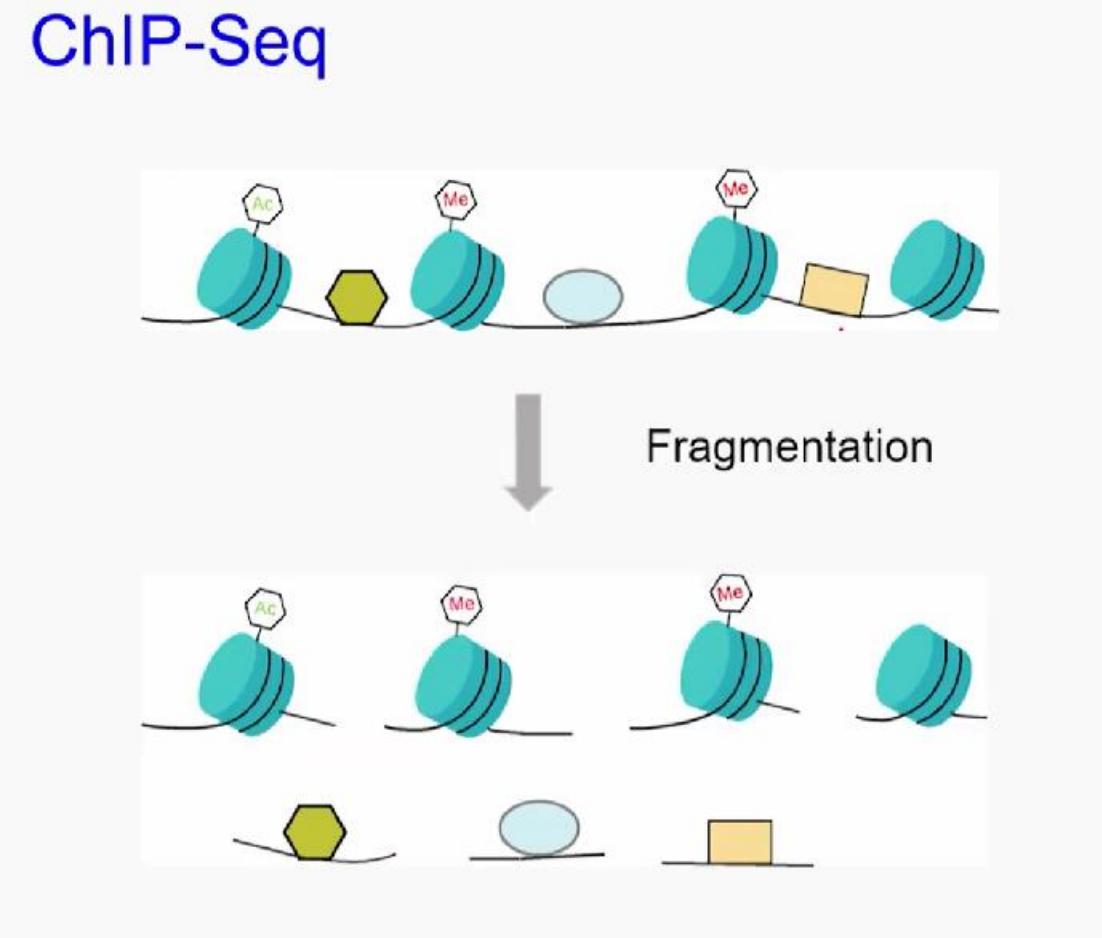


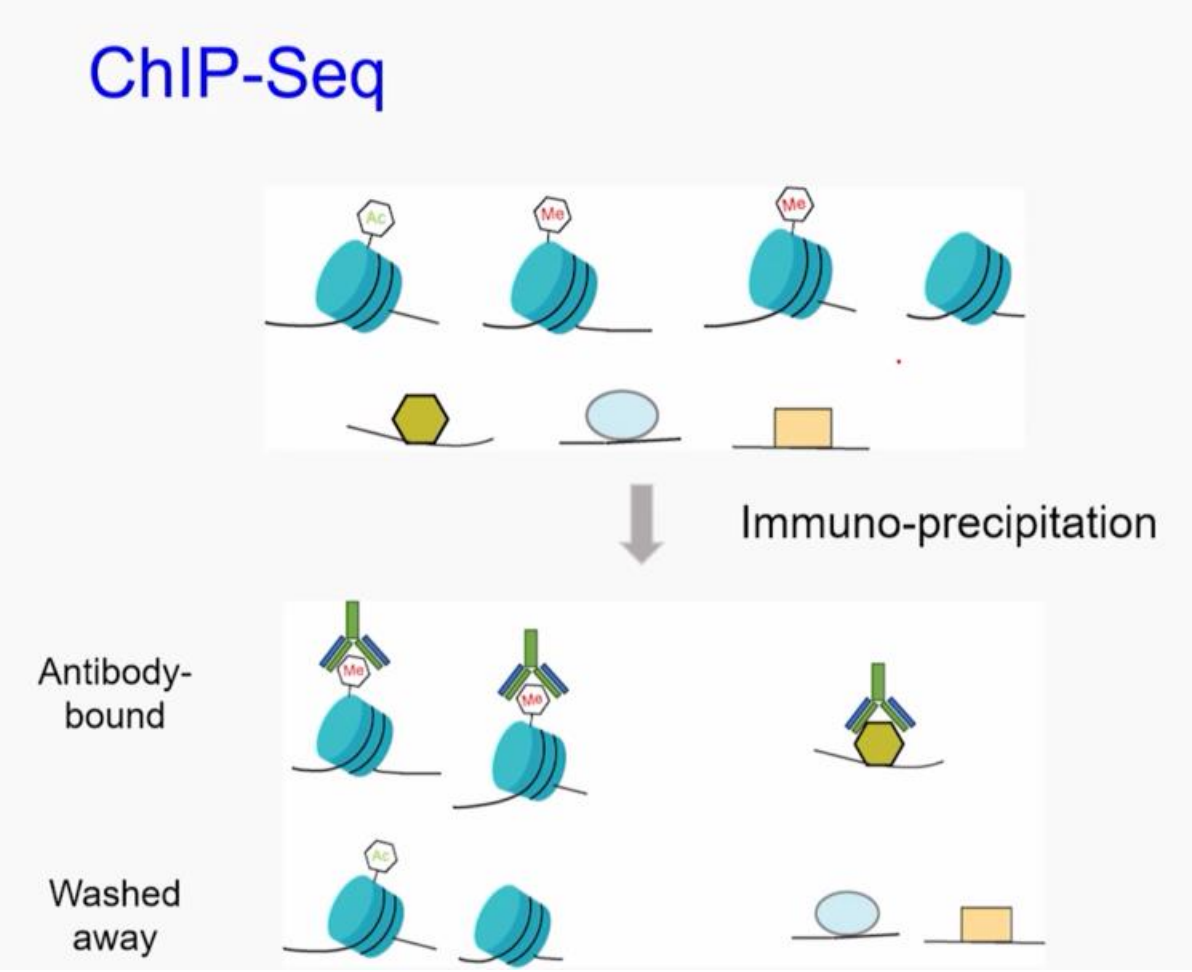
So, can we identify these sites of transcription factor binding in the genome directly? So, instead of relying on this methylation information, can we directly identify the binding sites of transcription factors in the genome? So, the ultimate goal is to understand how the epigenome

regulates gene expression. So, if we can identify these transcription factor binding sites and see whether transcription factors are bound to two specific sites, then we can correlate that with the gene expression levels. So, as I said, this actually will influence gene expression, and then, of course, we can also connect this information with the DNA methylation studies. So, we can then say that this kind of methylation pattern blocks these transcription factors or that this kind of methylation pattern attracts these transcription factors. So, this is something that would be more informative in association with the DNA modifications, ok? So, we will talk about this method called chromatin immunoprecipitation that actually allows us to do so, and it can be combined with next-generation sequencing. This method is called ChIP-Seq, and earlier it used to be applied with microarray technologies; it is called ChIP-on-chip. So, these are the full forms of the right chromatin immunoprecipitation chip with sequencing, and chromatin ChIP-on-chip would be chromatin immunoprecipitation with DNA microarray. So, these methods actually allow us to detect the binding sites of transcription factors in the genome, and they can also detect the binding of other DNA-binding proteins, such as chromatin remodelers. So, for any protein that can bind to the DNA, we can detect their binding sites and their locations, etcetera, and we can also detect the sites of histone modifications. So, we have not talked about this at all, and ChIP-Seq offers us a way to actually identify these histone modifications. So, what are the steps? So, the first step is actually the cross-linking of DNA and protein using formaldehyde. So, this actually preserves this connection between DNA and proteins. So, at certain sites, we have proteins bound to the DNA, and we want to preserve those connections right before we actually proceed with DNA fragmentation, etcetera, because otherwise these connections might get lost when we are preparing the sample. So, the first step is cross-linking, where we preserve the connections that we want to study.

The second step is DNA fragmentation, right? So, we create smaller fragments of DNA. The third step is immunoprecipitation using antibodies. So, this antibody will be targeted against the protein that we are interested in or against the histone modification that we want to identify. And then the final two steps are very familiar to us: the first is library preparation, and the second is sequencing. So, let us move on to this animation. So, you will understand this process better. So, what you have is this DNA, right? So, you can see these histones and their modifications on the top right.

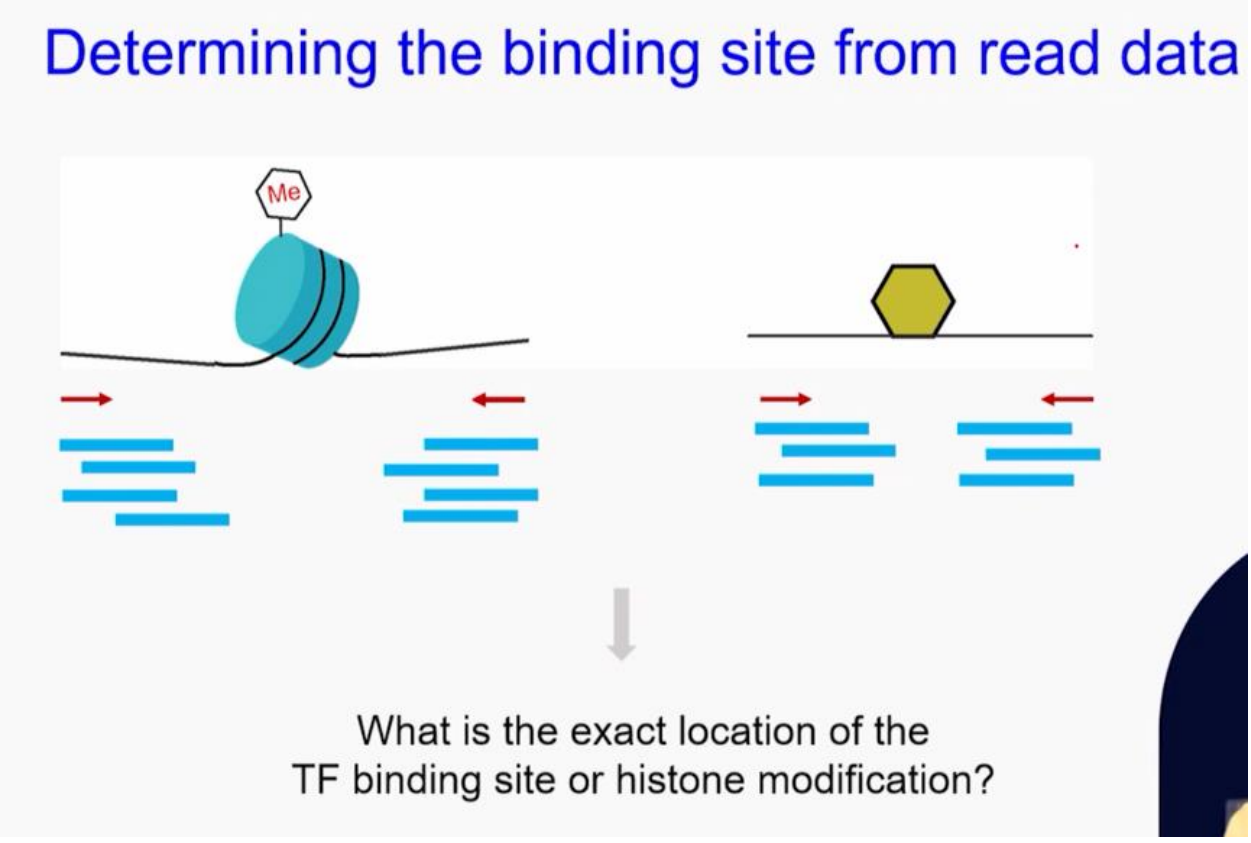
So, for some histones, you have methylation and some acetylation, just for illustration purposes. And then you also see certain proteins that are also bound to DNA.





This could be transcription factors or other proteins, but the process remains the same. So, what we do is do the cross-linking right, and then we have the fragmentation, okay? So, once we have this fragmentation, we get the fragments, and we do the immunoprecipitation using antibodies. Now, we can design antibodies that will detect these specific methylation patterns on histones, or we can design antibodies against these proteins, transcription factors, chromatin remodelers, or any other protein that binds DNA. So, once we have this immunoprecipitation, what happens is that we are pulling down these regions of the genome, and the rest of them are washed away. So, these regions are washed away in this process. So, once we have pulled these regions down, we purify the DNA fragments and then subject them to next-generation sequencing. So, we should identify these specific regions where we have this DNA binding event, whether it is histone modification or transcription factor binding. So, the next step is that we want to identify the binding site from the data. So, here is the challenge, ok? When you isolate this DNA fragment and sequence it, the reads will appear at the end, ok? So, if we sequence from either end, we can use parent

sequencing and sequence from both ends. So, the reads that I have that I am showing in blue color here right will mostly appear at the end. So, in these regions, we cannot control the size of these regions. So, they could be quite big; they could be 1 kB; they could be even longer, right? So, we cannot control the size, but then remember that the sequencing method is, if it is illuminated shortly, sequencing traditionally right. So, the reads will be concentrated at the end of this DNA fragment, ok? So, similarly, in this case, if you are interested in determining the binding site of this protein, we will end up being left with these reads that are coming from the end of this DNA fragment. So, the problem is that we have to now determine the actual binding site from the distribution of reads here. So, the actual binding site is somewhere in between these paired end data points, right? So, between these two sets of reads, somewhere in between, but from the distribution of reads, we have to infer the binding site. So, this is an inference problem; we cannot directly determine the binding site.



Now, the question is: what is the exact location of the tape binding site or the histone modification that we are interested in? So, there is a computation analysis, and this is called peak calling. I will

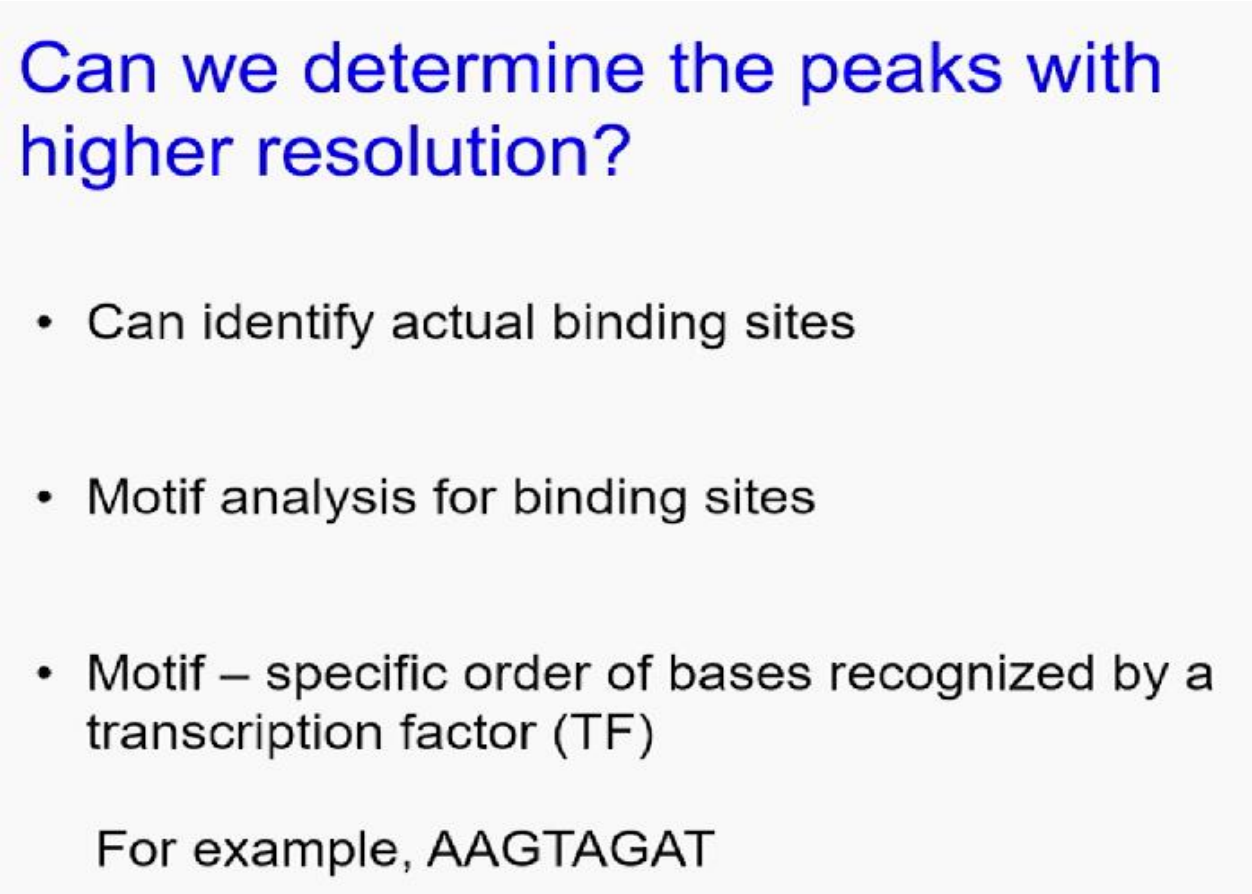
not go into the details. So, the idea is that by looking at the read distribution pattern, we will have to infer the actual binding site, ok, but again, this is an inference or prediction based on the distribution and not actual observation, ok.

So, this is what this method will do: it will determine the binding site or site of modification, and the resolution depends on the fragmentation sites. So, again, these fragmentation sites are not controlled in most cases, right? So, they happen randomly, and this resolution is dependent on the fragmentation sites. If the fragments are small, then the reads will come close to the actual binding site, and in that case, inference would be better, but if the fragments are large, the reads will be concentrated towards the end of the fragment, and then inference will be much more difficult. So, this resolution is usually 50 to 100 base pairs, but the actual binding sites of transcription factors are less, even less than 10 base pairs, up to 12 to 15 base pairs. So, you can see that there is a huge scope for improvement using this chip set method.

Now, we also do something called de-sales, or differential binding analysis. So, this is a comparative analysis between two or more samples, and this is focused on changes in binding sites or sites of modification. The process remains the same: we are just doing this for two different samples, and then we are comparing them. So, this could be very useful if you want to compare these transcription factor binding sites between two different conditions or between disease and healthy samples. It is very similar to what you talked about in the case of transcriptome analysis. So, we talked about differential gene expression analysis, and similarly, we can do differential binding analysis for transcription factors, and then, of course, we can connect this to the transcriptome data. So, whether these differences in the binding of transcription factors can influence the gene expression pattern is now one of the questions that we have: can we determine the peaks with higher resolution? So, as we have seen, the chip set actually does not give us the desired resolution. So, can we determine the peaks with better resolution? And then we can identify or determine the actual binding sites, not just infer or predict; we can actually see the actual binding sites, and we can do something called motif analysis for binding sites. So, what are these motif analyses? So, this motif is a specific order of basis recognized by a transcription factor.

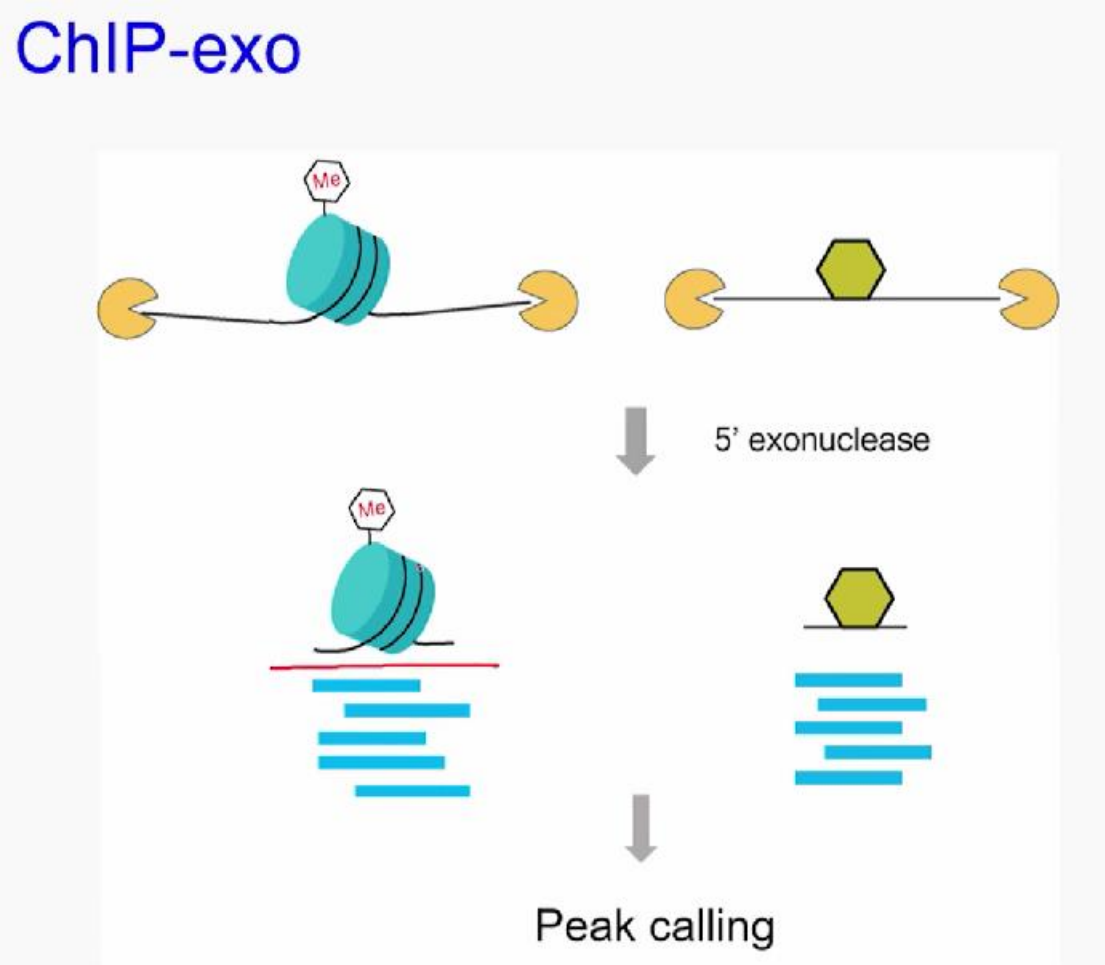
So, for example, a transcription factor recognizes these bases, or the sequence A G T A G A T,

and each transcription factor has a signature sequence, or maybe sometimes more signature sequences that it can bind to because of the chemical interactions, etcetera. And if you know this motif sequence, you can then search across the genome for the putative binding sites of that transcription factor.



So, this is actually very useful if we can determine the actual binding sites. And this method was proposed to do that; it is called chip exo. So, this is chromatin immunoprecipitation with exonuclease treatment, okay? So, the steps are mostly similar. So, as you can see, the first three steps are exactly identical as before. So, we have this cross-linking, then fragmentation, then immunoprecipitation using antibodies, but there is an extra step here, which is the exonuclease treatment of the purified DNA. So, or the purified fragments that we get? And then we have library preparation and sequencing, okay? So, what happens is that after immunoprecipitation, once you take out or pull down these fragments, we subject them to exonuclease treatment.

So, these are the exonuclease enzymes animated, and what they will do is chew away these bases from both ends. So, they will chew away the five prime ends, ok? And what will happen is that you will get much smaller DNA fragments. Of course, they cannot chew the sites that are bound by these proteins. So, what will happen is that they will stop once they encounter these proteins. So, you are left with these very small fragments. And now, if you sequence from both ends, you will get reads mostly focused on the binding sites, ok? So, you do not have to infer these peaks anymore, right? You can simply directly identify the peaks from this data, and this is a much simpler peak-calling method. So, from ChIP-exo, we get high-resolution peaks and sometimes even single-base resolution. So, this is great; we can identify the actual binding sites, and we can then do the motif analysis for transcription factors. As I mentioned, we can identify the specific set of bases that transcription factors identify. And then we can also look for putative binding sites for those transcription factors across the genome.

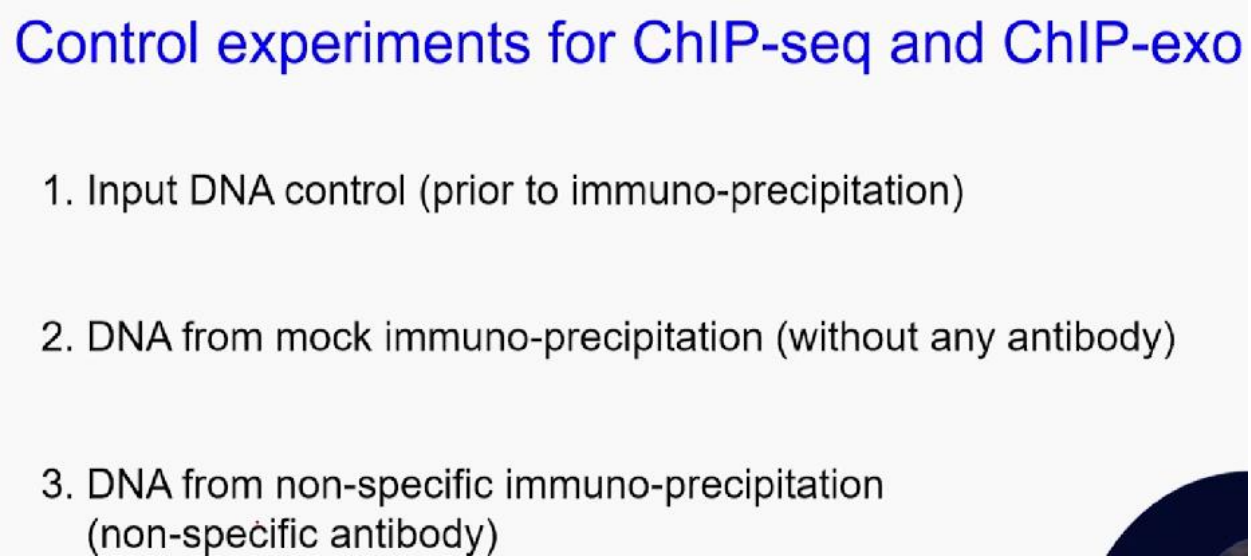


Now, there are several factors that actually affect the quality of the ChIP-seq and ChIP-exo experiments. As you can see, these are quite complex experiments. You first need an antibody, then you have to pull down, and if you are going for ChIP-exo, you need to do the exonuclease treatment. So, this is quite complex, and we have to be aware of the factors that can affect the quality of the data. So, the first thing is the antibody quality and specificity. As I mentioned, this is very critical for the process as well as the sample quality. So, the antibody quality is very important; it has to be very specific for the protein that we are interested in.

If it also binds to other non-specific targets, those will also appear in the results. So, this is something that we need to take care of, and we do so by having control experiments. So, these are experiments that we also do along with the original ChIP-exo or the actual ChIP-exo or ChIP-Seq experiments, just to ensure that the results that we get are robust and accurate. So, there are three types of controls that we need. The first is input DNA control, which is prior to immunoprecipitation.

So, remember what we will get after ChIP-Seq or ChIP-exO. We will get reads mostly focused on or mostly from specific regions of the genome, right? So, if you are looking at, let us say, transcription factor binding sites, we will get reads mostly around that region. So, there should be an enrichment of reads around that region, and that has to be compared with the input DNA control. So, because this enrichment could also happen at random, as you can see, the coverage could fluctuate across the genome. So, this fluctuation is also something we need to consider when we are analyzing or looking for enrichment in the case of ChIP-Seq or ChIP-exo. So, that is why we also need to have input DNA control; this is priority mini precipitation. So, whatever we have prepared before we add the antibody, we also take those fragments and sequence them. So, before and after mini precipitation, we see enrichment, and that gives us the transcription factor binding sites. In addition to controlling for other non-specific bindings, So, we also do ChIP experiments from mock mini precipitation. So, this is DNA from a mock mini-precipitation without any antibodies, ok? This is again looking for biases in the sample preparation protocol, and the third one is DNA from non-specific immunoprecipitation. So, this is a non-specific antibody that should not actually lead to enrichment of any kind in those specific regions. So, we also include that,

ideally, because this will tell us how much of the enrichment that we see is occurring by chance. So, these three types of controls would be ideal for any ChIP-Seq or ChIP-exo experiments.



Here are the references for this class, and to summarize, we have talked about direct detection of DNA methylation patterns and using two different next-generation sequencing methods. single-molecule real-time sequencing, or SMRT sequencing, and nanopore sequencing, and we have seen that they apply two different principles to actually identify these methylation patterns. The SMRT utilizes two parameters, which are the inter-pulse duration and the pulse width. So, based on the distribution of values for these two different parameters, the methods could allow us to distinguish between these methylated bases. And nanopore sequencing looks at the current signals, and apparently each molecule or each modification has a distinctive influence on the current signal, which could then be utilized to identify the modification. We have talked about chromatin immunoprecipitation, which can identify binding sites for specific transcription factors or specific histone modifications. So, if you are working with antibodies against transcription factors, we will be looking at binding sites for specific transcription factors, but we can also design antibodies against specific histone modifications, and then we will also know the histone modification sites in the genome. We can apply ChIP-exo, ChIP-Seq, and ChIP-Chip right. So, we have talked about ChIP-exo and ChIP-Seq, and if we talk about resolution, ChIP-exo has a higher resolution than ChIP-Seq, and ChIP-Seq has a higher resolution than ChIP-Chip. This is something that is seen in the data across different samples. And one of the major challenges to ChIP-exo is obtaining a

good-quality specific antibody, and as I said, you need to have an antibody that is specific against the target.

If you have non-specific interactions, that can bias your results. So, this is a challenge for these experiments. So, we will talk about other different types of methods that can actually allow us to look at these different binding sites or look at the nucleosome-occupied sites in the next class. Thank you very much.