Next Generation Sequencing Technologies: Data Analysis and Applications

Chromatin Accessibility

Dr. Riddhiman Dhar, Department of Biotechnology

Indian Institute of Technology Kharagpur

Good day, everyone. Welcome to the course on Next Generation Sequencing Technologies, Data Analysis, and Applications. In the last class, we talked about looking at transcription factor binding sites or identifying histone modification sites using chromatin immunoprecipitation. So, we will extend that to a much more general accessibility study, and we will talk about chromatin accessibility in this class. So, this is the agenda for this class. So, we will talk about methods that can allow us to assess chromatin accessibility, and again, the idea is very similar.

We want to look at the role or impact of the epigenome on gene expression. So, chromatin accessibility is a much broader term. So, we are mostly looking at open chromatin and closed chromatin regions, right? So, as we have discussed in the introductory class,

So, open chromatin regions are mostly transcriptionally active, whereas closed chromatin regions are transcriptionally inactive, and this can actually help us understand or associate chromatin accessibility with the transcription process. Here are the keywords that we will come across: So, nuclease digestion, cross-linking, and fragmentation are okay. So, we talked about chromatin immunoprecipitation experiments in the last class, and we also talked about the challenges that we face when doing this experiment. So, the first challenge is obtaining a good-quality and specific antibody.

This is something that is very critical for the experiment. If you have non-specific interactions, it will bias your result. This is something we talked about in the last class. One thing you should remember, or you should understand, is that if you want to identify, let's say, 10 different transcription factors, you need 10 different antibodies against those factors. So, you have to generate antibodies against each protein or each histone modification that you want to study, and this means this is a very tedious process.

So, the generation of antibodies that are also specific and also quite expensive is okay. So, what are our options? So, can there be alternate methods? So, we will talk about some

methods. These are actually not actual alternates, but they can actually supplement these cheap experiments. So, these methods actually look at a much broader aspect, and they look at chromatin accessibility.

So, these are four methods that we will discuss very briefly. The first one is called MNase-seq, the second one is called DNase-seq, the third one is FAIRE-seq, and finally, we will talk about ATAC-seq. This is the most recent method for looking at chromatin accessibility. So, let us first look at what we mean by chromatin accessibility, as I mentioned. So, we have open chromatin, which is accessible to transcription factors and polymerases.

So, this region is transcriptionally active. On the other hand, you have closed chromatin, and this region is closed, which means they are bound by nucleosomes, mostly occupied by nucleosomes and histones. So, they are densely packed, and they are not accessible to transcription factors. So, open chromatin is transcriptionally active, whereas closed chromatin is mostly inactive. So, let us now look into the methods, the steps, and the different considerations for each of these methods.

So, the first method that we will talk about is called MNase-seq, or MAINE-seq. So, this is the full-form micrococcal nucleus combined with sequencing. So, MNase stands for micrococcal nucleus, or what we also call MAINE-seq, is micrococcal nucleus-assisted isolation of nucleosomes. So, this is again combined with sequencing. So, what exactly does this method do? It actually looks at the location of nucleosomes and DNA-binding proteins in the genome.

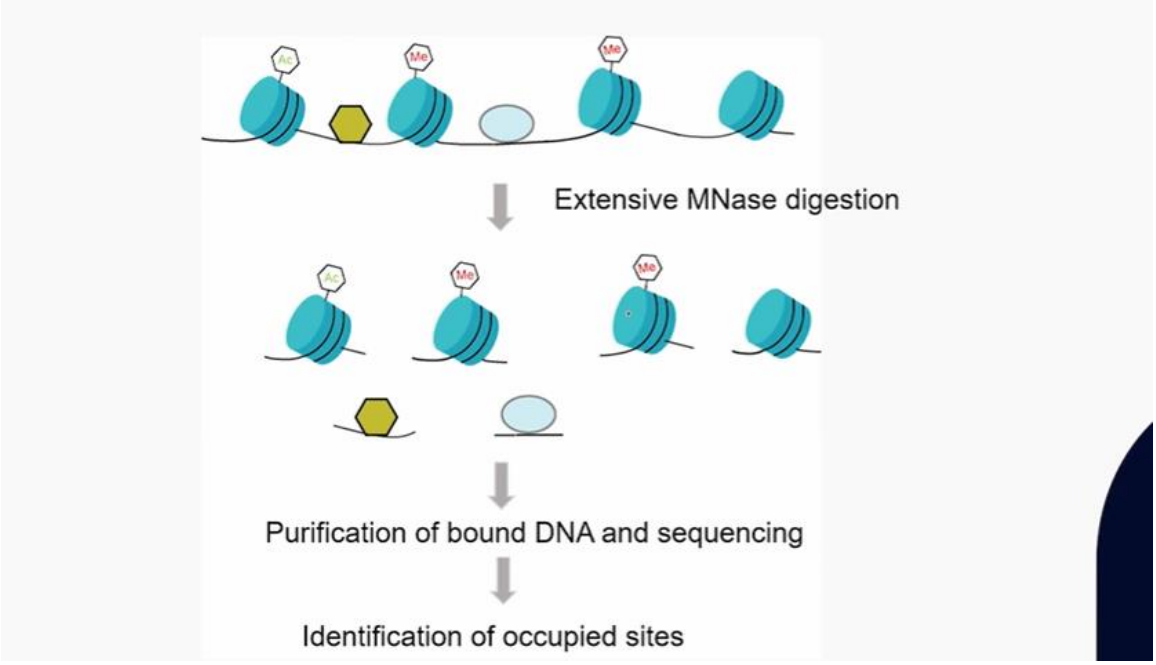
So, one major difference from a chip is that it will tell us where the regions where DNA is bound by proteins are, but it will not be able to identify those proteins. So, this is a major difference from a chip. So, here we are looking at a much more general or broader picture, but at the genome-wide scale. So, just a very brief background on this MNase enzyme. So, this is the first micrococcal nucleus that was derived from Staphylococcus aureus, and because of its properties, it was used to determine chromatin structure way back in 1975, and the term MNase-seq was coined in 2009.

So, MNase-seq was first employed for the determination of nucleosome occupancy in sea elegance in the worms. This was combined with NGS, okay? So, here is the reference that you can look at, ok? So, when we talk about nucleosome occupancy, we are looking at where nucleosomes are bound in the genome. Again, this relates to open- and close-chromatin conformation, and this method was then applied to the human genome, also looking for nucleosome occupancy, in 2008 in this paper that is mentioned here.

So, how does MNase-seq work? So, this is where we understand the process, right? So, what happens in the first step is that we digest the genome by the micrococcal nucleus. So, this micrococcal nucleus is endoexonuclease. So, as you see, this makes a difference in the process, ok? So, MNase-seq digests DNA in the presence of calcium, and it has a slight preference for its consensus sequence which is given here.

So, it is an AT-rich sequence, which it prefers. And what we do is actually, at the end, get the nucleosome-bound regions and identify those by NGS. So, here is the schematic of the animation for the whole thing, right? So, we have this region here, okay? Again, some regions bound by transcription factors and regions are bound to histones, and then these histones have some modifications, etcetera.

So, what we do is do extensive MNase digestion here, ok? So, what will happen because this is endo exonuclease? It will cut right, and it will also chew the ends. And what you will be left with are these bound regions, right? So, MNase cannot chew or digest these regions because they are protected by these proteins. Whether they are histones or some transcription factors, these regions will be intact. So, what we do is actually purify this bound DNA and then sequence it.



And at the end, again, what we will get is the sequences enriched for these occupied sites. So, some considerations for MNase-seq are that it requires many cells. So, we need about 1 to 10 million cells, which is quite a high number. You need to titrate this process very carefully, deciding how much enzyme you will add and how long you will incubate. So, all these things are very tedious because, with the wrong enzyme concentrations or if you do really long MNase digestions, some nucleases have been shown to be sensitive to MNase digestion.

So, even in these nucleosome-bound regions, they will be digested by MNase. So, there are some MNase-sensitive nucleosomes, okay? So, you have to titrate this very carefully. So, what we have seen is that MNase-seq will actually map the occupied positions. So, it will not directly identify the active regions, but it will actually be an indirect determination of the active regions.

So, you know the occupied position. So, the rest of the regions might be free. And what has been seen is that it requires about 150 to 200 million reads. So, that is quite a large number for accessibility studies in the human genome. Now, moving on to DNase-seq, this is the second method that we are going to talk about.

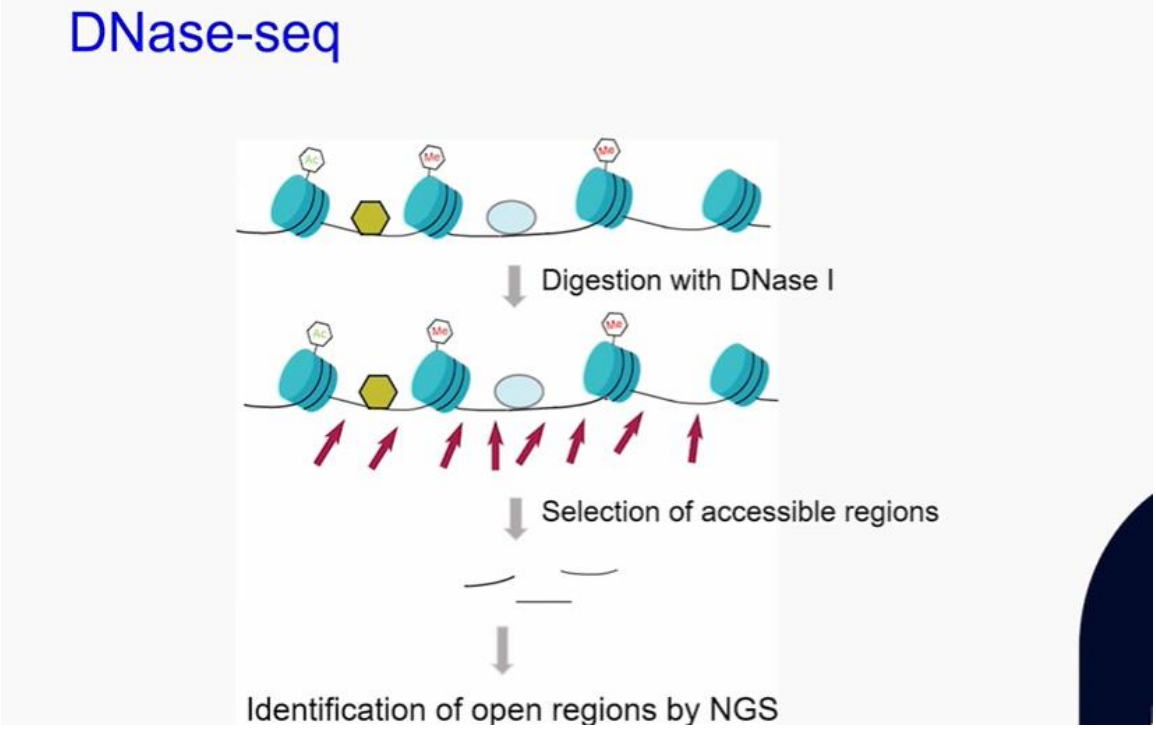
So, this actually helps us in the direct identification of the active region or open chromatin. So, this is in contrast with the earlier method of MNase-seq. So, here we can directly identify the open chromatin region, ok? So, the steps are very similar, but the principles are slightly different. So, what we do is digest DNA by enzyme.

So, this is a non-specific endonucleus. So, it will cut indiscriminately, and it will not be able to cut at occupied sites. So, sites that are bound by the stones or by other transcription factors will not be cut, but only the open sites will be cut. So, according to what researchers have seen, these active genes that have open chromatin conformation are more sensitive to digestion with DNase I. So, this means these open regions will be preferentially cut by this enzyme.

So, here are the steps. In the first step, we subject the DNA to mild DNase I digestion. And then what we want to do is size and separate these fragments, right? So, we want to select a certain size that will be suitable for sequencing. So, to do that, we actually take these DNA fragments and embed them in low-melting agarose plugs and then these digested DNAs are size-selected. And then we follow the library preparation and next-generation sequencing, ok?

Again, coming back to the animation, right? So, we have this DNA again mounted by nucleosomes or transcription factors, but then you have some regions that are free, and these are the sites where DNase 1 will work ok. So, you can see this size that will be attacked or digested with DNase 1, and you will end up with these fragments that are size-selected fragments, right? So, we will see that for some size, that is appropriate—not too short, not too long, right? So, we will select at the end of this site these regions that are

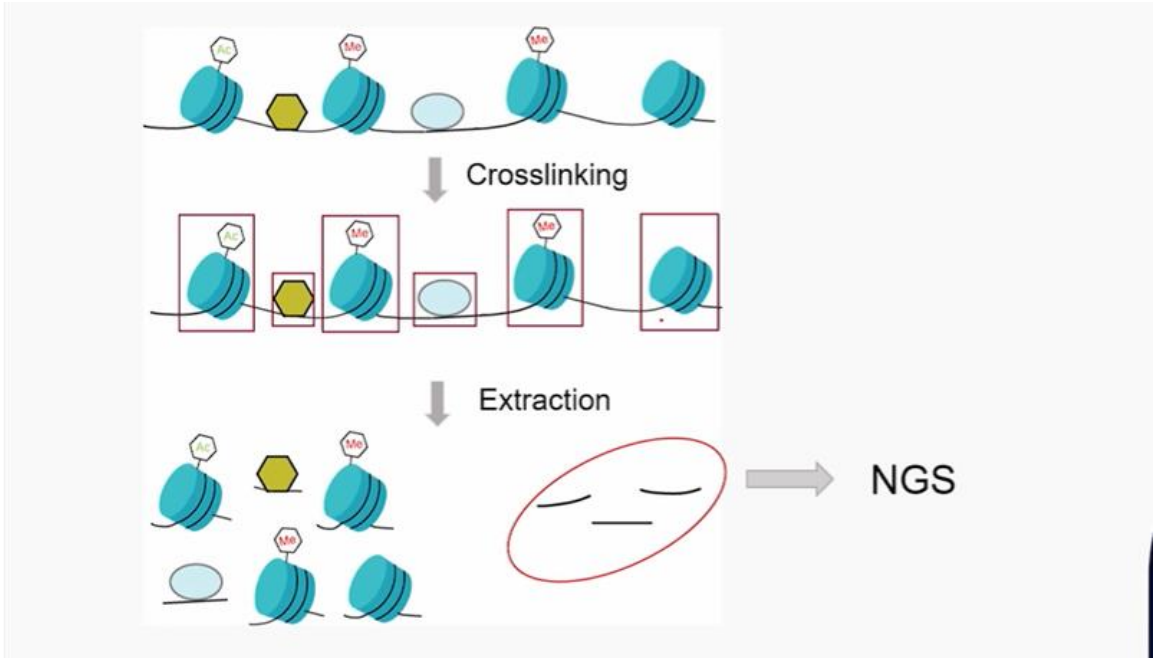
actually open right.



So, these are not occupied by nucleosomes or transcription factors, ok? And then we will sequence these size-selected fragments, and they will be sequenced by NGS, which will help us identify the open regions or open chromatins. Again, coming to considerations for DNase-seq, it again requires a lot of cells like MNase-seq and you can then compare it with MNase-seq, right? So, this is also a time-consuming process because you need enzyme digestion, etcetera. Again, we have to do the enzyme titration to determine how long you want to digest, what the concentration of enzymes is, etcetera, and you also need about 20 to 50 million reaches for standard accessibility studies of the human genome.

So, this number is slightly less than MNase-seq. Coming to the third method, which is called the FAIRE-seq method, this method relies on cross-linking. So, here you see formaldehyde-assisted isolation of regulatory elements, and this is combined with sequencing. And this method also maps the open chromatin region, and here is the original paper right here that is describing this method. So, what are the steps in the FAIRE-seq method? So, we have cross-linked chromatin with formaldehyde again, and this will actually capture the protein-DNA interactions that are happening right now.

And the next step is the sharing of chromatin with sonication, followed by phenol-chloroform transformation. So, you have separation of aqueous and the other layers, right? So, the DNA-free DNA will be in the aqueous layer. So, which you can take out and sequence, and then the sequencing of the open chromatin regions right? So, why am I saying we will sequence chromatin regions? Because when you do the phenol-chloroform separation you will take the aqueous layer and get the free DNA, which is not bound by proteins, and once you isolate those DNA fragments and sequence them, you will get the open chromatin regions.



So, coming to the similar diagram or animation again, we have the DNA with nucleosome binding and transcription factor. So, we will illustrate that the first step is cross-linking. So, again, this is preserving this DNA-protein interaction because we are not doing any digestion here right? So, we have this cross-linking here, okay? So, this will preserve these DNA-protein interactions, and then we can extract them.

So, we use phenol-chloroform extraction, and then we can separate out these two different types of genomic segments: the free region that is not bound by proteins and the region or segments that are bound by some protein, whether transcription factor or histones, and we can separate out this fragment that is free, just DNA, and we can sequence ok. So, the sequencing result will give us only the regions that are open or free; they are not bound by

any transcription factor or occupied by nucleosomes. Coming to the final method, which is the attack, So, this is the most recent method for assessing accessibility. So, the full form is an assay for transpose accessible chromatin using sequencing, and here are two papers that actually describe this method.

And as I mentioned, this is the most recent method, and one of the key steps in this attack-sake method is to fragment and tag the genome simultaneously by using an enzyme called Tn5. So, what you have is that this is part of this transposition, right? So, they actually insert DNA into the genome, okay? So, you can utilise that property by inserting adapters. These are the NGS adapters that you can insert into the genome. So, this transposes them to actually insert these adapters and also fragment the DNA.

So, that is why this is called tagmentation, right? So, we have fragmentation plus tagging. So, that actually is tagmentation, ok? So, what researchers have seen is that because of steric hindrance, So, this is an enzyme big enzyme right, and it can only access the sites that are open right. So, if you have occupied sites that are occupied by nucleosomes or transcription factors and open sites, these proteins can only access the sites that are open.

So, because of this steric hindrance, So, all these adapters will now be integrated into the regions that are accessible, because of the property of these transposes. So, it will integrate these adapters into the genomic regions, but where can it get access? And then what we can do is amplify these regions that contain adapters. So, we can use this adapter sequence as a primer binding site; we can amplify these fragments and sequence them using NGS. And this actually then identifies the open chromatin region, as we have just seen, because this integration of adapters will happen only in the open or accessible regions.

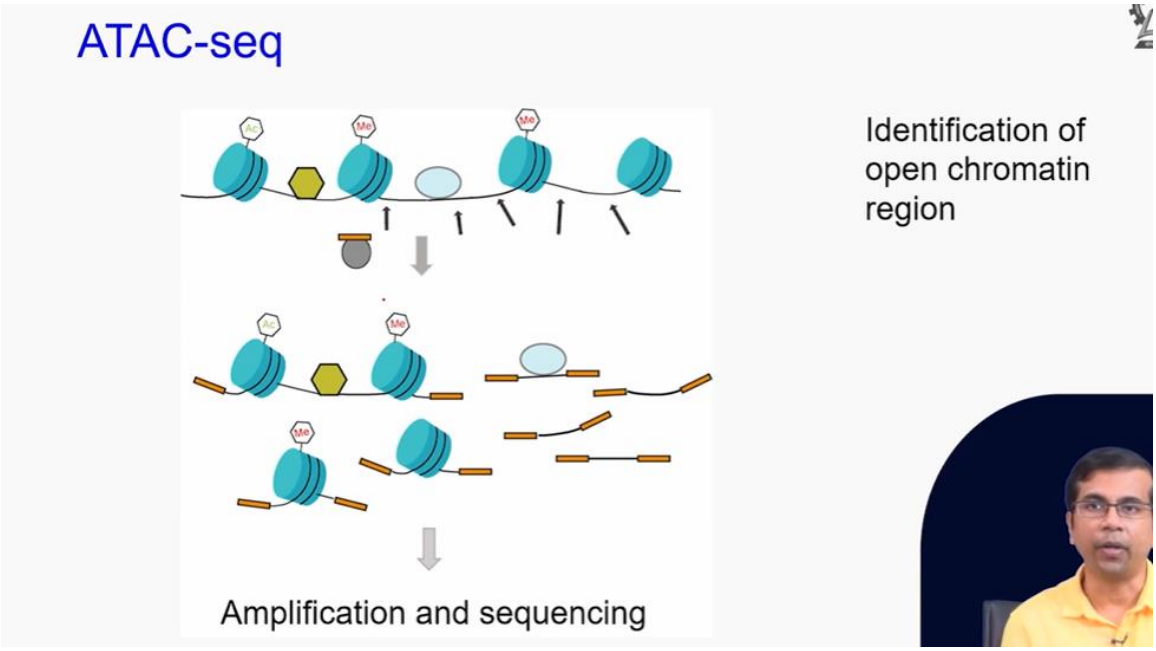
So, here is the schematic again, right? We have the DNA bound to nucleosomes and proteins. So, imagine this is the transpose right, and the orange is the adapter sequence right that will be integrated into the genome. So, you can see by the arrows that I am showing the sites that are accessible to this enzyme. These are open regions, but the rest of them are not right; they are occupied by nucleosomes or by other proteins. So, what happens here is that we get these fragments right, which are also integrated with adapters.

So, you get these fragments here, right here, you have this transcription factor that is there, and then at the end, at two sites right, you also get integration of this adapter sequence. Then you also have some fragments coming from these free regions, as you can see right here, and maybe this region is here. These are free regions that are here; they are coming into the picture. But here also, like some of these regions that are containing only, let us say, one histone, they will also be isolated, and they will also come in these fragments. And then, of course, you will get a bigger fragment because, in this region, there is no accessibility for this enzyme because of steric hindrance. So, you will get this big fragment with the adapters at the end.

Now what you can do is actually site select some of these fragments, right? So, of course, the amplification will not work for very long fragments, right? So, you can amplify only a certain range of fragments of a certain size. So, you can do a site selection also and then amplify and then sequence, ok? So, once you sequence, you will get those fragments whose size is within a reasonable range, and these will mostly cover these open regions.

So, open regions will be quite small because they are accessible for this enzyme, and it will integrate this adapter across the system. The other regions are right, and these ones again will depend on how long these fragments are and whether the amplification actually works. So, for now, we can now imagine right if you have genomic regions that are really dense. So, these are the closed regions with a lot of nucleosomes there; this enzyme will not have access or very rare access, and these integrations will happen at longer distances, which

will not be amplified by PCR, or you can discard those fragments by site selection.



So, I hope this principle is clear. So, in this process, we will identify the open chromatin regions that are accessible to the transposes. Now, here are some considerations for ATAC-Seq. So, what researchers have seen is that ATAC-Seq can work on 500 to 50000 cells. So, that is a very low number compared to the other methods that we have discussed so far and this has recently been extended to single-cell ATAC-Seq. So, it means we can now take out genomic DNA from one cell, and you can do this ATAC-Seq ok?

So, because this process requires a very low amount of DNA, you can even work with a single cell. This is not possible with other methods, right? So, other methods rely on some sort of digestion or chemical treatment, and one of the steps that you have with chemical treatment or restriction digestion is that you would have to purify the sample after those treatments. So, you need to start with more samples so that you have enough after those purification steps.

Here, you do not have to do any of these steps, right? So, there is no chemical process like chemical conversion or restriction digestion. So, you can work with even a single cell and just take its genomic DNA and do the ATAC-Seq. And the library preparation step is very

fast here, right? So, this is a two-step library preparation compared to the earlier ones, which were quite laborious. We we have to titrate to determine which enzyme concentration you should use, etcetera. Here, this is not really required because this integration of adapters will happen only in the free regions, open regions right that are not bound by any protein, and again, this is not very highly dependent on the concentration of the transposons.

So, this process is very fast, and it is also quite easy compared to the other methods. And it requires about 60 to 100 million reads for standard accessibility studies of the human genome. So, this number of reads is comparable to the other methods, but the advantage lies in the first two-step process, and we can work with a very small number of samples. So, this method has now been extended to single-cell studies. So, we want to compare these four methods, and this is what I am going to do here.

So, what we have seen is that MNase-seq, DNase-seq, and FAIRE-seq require chemical or enzymatic treatment of DNA for sample preparation. This is something we have talked about, right? So, MNFAIRE-seq utilises this MNase enzyme, which has endo exonuclease activity, and we map only the bound sites or the occupied sites. DNase-seq requires this DNase I enzyme, and FAIRE-seq requires the formaldehyde cross-linking.

So, this actually requires chemical or enzymatic treatment. So, this could require optimisation, and they increase complexity throughout the whole process. In comparison, ATAC-Seq requires a very fast two-step sample preparation protocol; it utilizes these Tn5 transposes, which actually integrate these adapter sequences. These are from NGS adapters that are directly inserted into the genome and prepared for library preparation and sequencing. So, another major difference is that MNase-seq, DNase-seq, and FAIRE-seq require much more samples compared to ATAC-seq, as I mentioned right, because in chemical treatment and enzymatic treatment, you need to purify or deactivate the enzyme.

So, this purification step means you lose some samples. So, you have to start with more samples, and as I have told you right, ATAC-Seq can be done on even a single cell. So, it requires a much smaller sample, and you can do this on even single-cell DNA. So, you will

see this study is now coming out called single-cell ATAC-Seq or scATAC-Seq. Another difference we have seen this time is with MNase-seq.

So, MNase-seq maps occupied or closed chromatin regions mostly right. You have seen because of the properties of the MNase I enzyme that this enzyme chews up the DNA that is not bound by any protein, for example, histones or transcription factors. So, at the end, we are left with DNA that is bound. So, we actually sequence the occupied or closed region. Whereas the other three methods, DNase-seq, FAIRE-seq, and ATAC-seq, actually map open chromatin regions.

Of course, they are not exactly identical. So, if you look at the results of these methods, you will see there are slight differences. Even though they map the open chromatin regions, you will see some differences across these methods. And also, you can compare these results with MNase-seq, and there will be some differences because of the properties and the protocols that will follow. So, these are the references for this class. So, to summarise, we have talked about different methods for assessing chromatin accessibility, and they do not require any antibodies.

So, we started with the limitations of the ChIP-Seq experiment or the ChIP-exo experiment. So, they required antibodies, and these antibodies would have to be raised against each protein or each type of modification that you want to study. So, that makes the whole process expensive. So, we wanted alternative methods that could allow us to study these binding sites or their accessibility. However, these methods that we discussed are not exact alternates because they can give us an overall view of the chromosome, including the open chromatin and the closed chromatin.

But if they cannot tell us about the transcription factor bindings, right, which transcription factor is bound where, or which modifications are present at which locations, For that, we still need to depend on the ChIP-Seq experiments. So, we have discussed these methods. Mase-seq, DNase-seq, and FAIRE-seq require chemical learning enzymatic treatment before library preparation. ATAC-Seq relies on something called the tagmentation process, which is tagging plus fragmentation, and this enables the insertion of these adapter sequences directly in the DNA. This happens specifically or preferentially in the open

chromatin region because the closed chromatin region is not accessible to the transposes enzyme. Then we can also see that ATAC-Seq can work with smaller amounts of sample, and it works even on single cells.

So, we have these single-cell ATAC-Seq datasets. And what I like to say at the end is that we can actually, by looking at this chromatin accessibility, start to connect the epigenome with the transcriptome. So, we know which regions are open, whether they are available for binding by transcription factors or where anion polymers can come and bind. So, this can be associated with the transcription state of the cell. So, what we have so far discussed are three different types of methods, right? So, we have talked about the methylation studies, where we are looking at the DNA methylation patterns.

And then we talked about the chromatin immunoprecipitation method, where we are looking at binding sites of transcription factors or sites of specific histone modifications. And finally, we talked about nucleosome occupancy. So, when we apply these methods to actually look at chromatin accessibility, or, in other words, nucleosome occupancy, we can differentiate between open chromatin and closed chromatin regions. And when we combine all these three together, we can connect the epigenome to the transcriptome.

So, we have talked about these three different methods, right? There are actually three different types of methods that can help us identify or study three different types of epigenetic modifications, and these are all important for the transcriptome status of the cell. In the next class, we will be talking about another aspect of the epigenome, which is the 3D genome configuration. This is not any modification on the genome; this is actually the configuration of the genome itself and how it can affect transcription or expression of the genes. So, in the next class, we will talk about some of the methods that will actually help us identify the genomic configuration in 3D space, and then we will look at the different structural elements that are present in the genome. Thank you very much.