Next Generation Sequencing Technologies: Data Analysis and Applications

Genome Organization in 3D

Dr. Riddhiman Dhar, Department of Biotechnology

Indian Institute of Technology Kharagpur

Good day, everyone. Welcome to the course on Negation Sequencing Technologies, Data Analysis, and Applications. In this lecture, we will be talking about genome organisation in three-dimensional space. In the last few classes, we have talked about different types of epigenetic modifications and how we can identify them using nucleotide sequencing technology. So, we have talked about DNA methylation, transcription factor binding site identification, or histone modification identification, and then we also talked about chromatin accessibility. So, those are the modifications on the DNA or some sort of binding of the proteins to the DNA, and they actually influence how they actually influence gene expression.

In this class, we will be talking about the three-dimensional configuration of the genome itself. There is no other protein that is actually modifying expression patterns or directing expression patterns; it is the three-dimensional configuration itself and how it can actually influence the expression of genes. So, of course, these three-dimensional configurations are associated with some sort of modification, which could be nucleosome occupancy or some other interaction between different elements in the genome, which can then drive the expression. So, let us start with that. So, this is the agenda for this class.

So, we will be talking about the organisation of the genome in three-dimensional space. Here are the keywords for this class: the first one is TADs, and the second one is LADs, and then we will talk about high. So, we will explain this term as we go along. So, the question that we are interested in is: how is genomic material organised in 3D space inside the nucleus? So, the genome is not compressed in one location and everything is all mixed up together, ok?

So, it also has a three-dimensional configuration and this is what this study is all about, ok? And this three-dimensional configuration is very important for long-range interactions, right? So, there are different elements of gene regulation. So, you can have, for example,

an enhancer that interacts with the promoter and can influence the downstream expression of the downstream gene. So, this also has regulatory links, right?

So, we can have this long-range interaction influencing gene regulation, and one example that I gave is the enhancer-promoter interactions. interactions and if we understand this genomic organisation in 3D, we can then map these enhancer-promoter interactions very accurately. So, with the earlier method, there has been a lot of interest in studying this three-dimensional configuration of the genome or this interaction between different genomic regions, and the traditional method utilised something called FISH or fluorescence in situ hybridization. So, very briefly, this method utilises a fluorescent probe for identifying this location of sequence in the genome. So, let us say you are interested in a specific sequence. So, you will design a probe that is complementary to that sequence, and of course, you also have the fluorescent molecule that is attached to this probe.

So, with microscopy, you can then identify the location in the genome where this probe is going and you can then say that this is where the probe is. So, you can utilise this method to actually study the interaction between two different regions of the genome. So, let us say your hypothesis is that this region A is interacting with region B, and then what you do is test this hypothesis. So, whether in region A or region B, they are interacting with each other. So, you design probes against these regions, region A and region B, with different fluorophores.

So, one could be a green fluorophore, and there could be a red fluorophore, and then you can see their localization in the genome. If you see that these probes are fluorescent molecules under a microscope, they are localised to the same region of the genome, so it is likely that they are interacting. Now, as you can realise, this is probably a quite time-consuming method. So, you have to design probes for each sequence, and then, of course, you cannot actually scale them up for studying all sorts of interactions that are happening in the genome. So, you cannot apply it on a large scale across the genome, OK?

So, as I mentioned, this is low throughput, but if you are interested in looking at interactions between two specific genomic regions right now and these are of interest to you, then you can, of course, apply this method and you will get some sort of idea about the distance in

three-dimensional space in the genome. So, before we actually go into the methods that allow us to do this in high-throughput manner, What I will do is introduce certain terms that are commonly used, and then I will discuss the methods. This way, it will be easier for you to understand because, without understanding these terms, if we use them later on, it will be very confusing for us. So, let us introduce these terms first, and then we will discuss the methods, but actually, in reality, this happened the other way around.

So, researchers applied those methods first and they identified these different regions of the genome and used these terms correctly, but we will go in another direction for a better understanding, ok? So, the first term that will come across is called chromosome territories. I will explain in a moment what chromosome territories are, and I will also have a schematic diagram for a better illustration. And then the next term we will use is called the A/B compartment. So, in A compartment or B compartment, we will use the term topologically associating domains, or TADs.

This is a term that you have seen in the keywords. I will explain what these TADs are. Then you have lamina-associated domains, or LADs; again, you have seen this in the keyword. Similarly, you have something called nucleolus associating domains, or NAGs, and now we will discuss each of these terms carefully. So, the first term is chromosome territories, okay?

So, it turns out that in a genome, let us say the human genome, the chromosomes actually occupy specific regions in the nucleus. It is not like all chromosomes are jumbled up or mixed together. So, each chromosome occupies a specific region in the nucleus, and these are called chromosome territories, right? So, you have chromosome 1 located in a specific location; chromosome 2 may be in another location; chromosome 3; and so on. And that means there is limited spatial interaction between two chromosomes.

There is some interaction, of course. You can look at some of the data that I will also mention in the reference. And you will see that there is some interaction between certain chromosomes, but these interactions are quite limited. The next term is A/B compartments, ok? So, what researchers have seen is that chromosomes are organised in two compartments: one is called the A compartment, and the other is called the B compartment.

So, a compartment is associated with open chromatin, and this is a transcriptionally active region, right?

So, we have now explained or come across this term, open chromatin, many times. And you understand that this region is open for transcription factor binding and RNA polymerase. Transcription can happen, right? So, this region is transcriptionally active. On the other hand, you have the B compartment, which is associated with closed chromatin.

So, these are all so; this region is condensed, right? So, it is not nucleosome-occupied and mostly transcriptionally inactive. The next term is topologically associating domains, or tags. So, these are regions in the genome that preferentially interact with each other. So, if you look at the three-dimensional structure, you will see that these regions have a lot of interactions within themselves.

And then they have much less interaction with other regions, ok? So, these regions are called tags, and they might have a regulatory role, which can impact the gene expression pattern. One interesting point is that tag boundaries are marked by CTCF binding sites. So, this is a transcription factor that recognises a specific sequence. And it has been seen that these boundaries, or the ends of tags, are marked by this CTCF binding site.

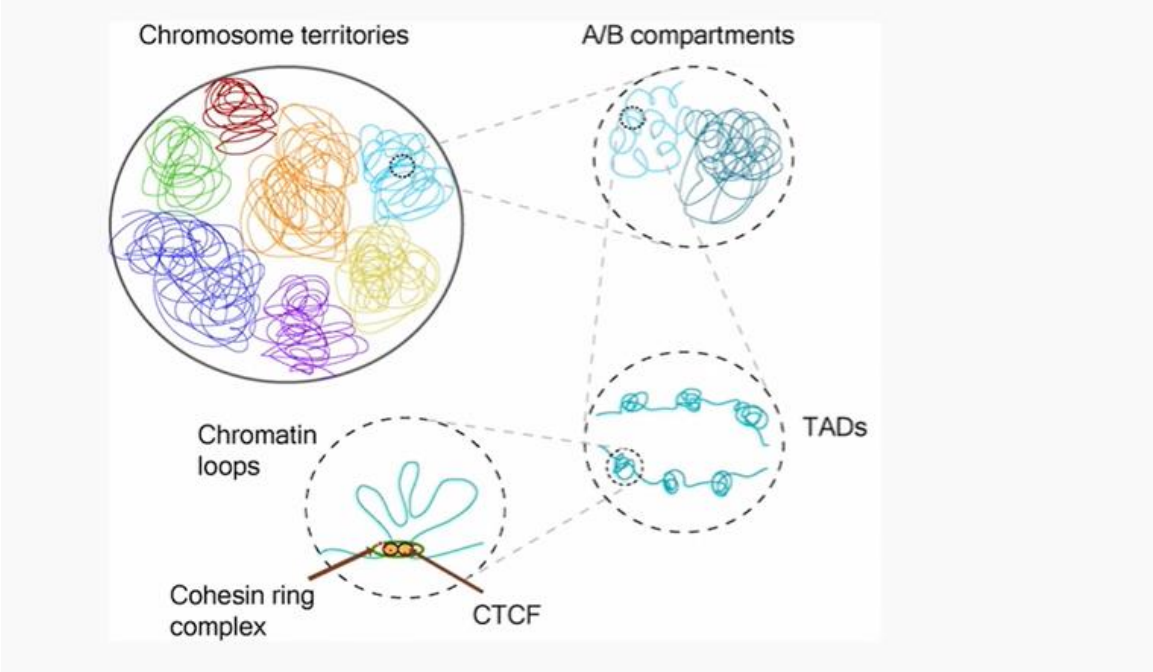
We will see the figures in a moment, and some sort of illustration will be better for you. And also, there is the presence of cohesion; it forms like a ring-like structure and that kind of keeps the kind of tag intact. Then you have lamina-associated domains, or lads. So, these are regions of the genome that actually interact with the nuclear lamina, or specifically with lamin protein. So, what it means is that it is mostly associated with the nuclear lamina, and this is mostly transcriptionally inactive.

Now, there are two parts to these lads. So, you have some constitutive LADs or cLADs. So, these are in contact with the lamina across cell types. So, if you study these regions across different cell types, you will see that in all cell types, these regions are in contact with the lamina. And these are mostly closed chromatin, and they are transcriptionally inactive.

And then you have something called facultative LADs, or fLADs. So, these are regions that are in contact with lamina in some cell types, and they are not in contact in some other cell types, right? So, in this is kind of not like constitutive lads, ok? Then you have the final term, which is the nucleolus associating domains, or nads. Like lads, these are regions that interact with the nucleolus.

So, the nucleolus is the region of the nucleus where hyposomes are formed. So, this is where the nucleolus interacts, right? So, these are also mostly heterochromatin in condensed form, and they are transcriptionally inactive. So, as you have seen, cLADs and nads are mostly transcriptionally inactive; F lads can be transcriptionally active or maybe inactive. So, just to illustrate better, I will use a schematic diagram and some animations.

so that you understand this better. So, imagine this is the nucleus, right? The circle is the nucleus, ok? And inside those inside the circle you have those different colours right ribbons of different colours; these are the chromosomes, ok? So, each chromosome is occupying its own space. So, these are the chromosome territories, and you can see that there is limited interaction between the chromosomes.



So, you can see some interactions, perhaps here or maybe here, but in total, this is quite limited. Now, if you zoom in right on one of the chromosomes, you will see things like A

and B compartments, just to remind you again. So, the A compartments are transcriptionally active, which means they are mostly open, and then you have the B compartments, which are in condensed form and are transcriptionally inactive. Now, if you zoom in further right in the small region here what you will see are the TADs now, ok? As I mentioned in TADs, these are topologically associating domains right?

So, they interact preferentially within themselves. So, so here is a TAD, here is another TAD, right, and you see this, they are interacting within themselves more and they are not interacting much with other TADs, ok? So, this is actually the definition of TADs, ok. And if you zoom in further, you will see these chromatin loops and these TAD boundaries right where you have these CTCF proteins. So, you can see this protein X-shaped protein, like this is CTCF, and then you have a ring-like structure that is formed by cohesive ok.

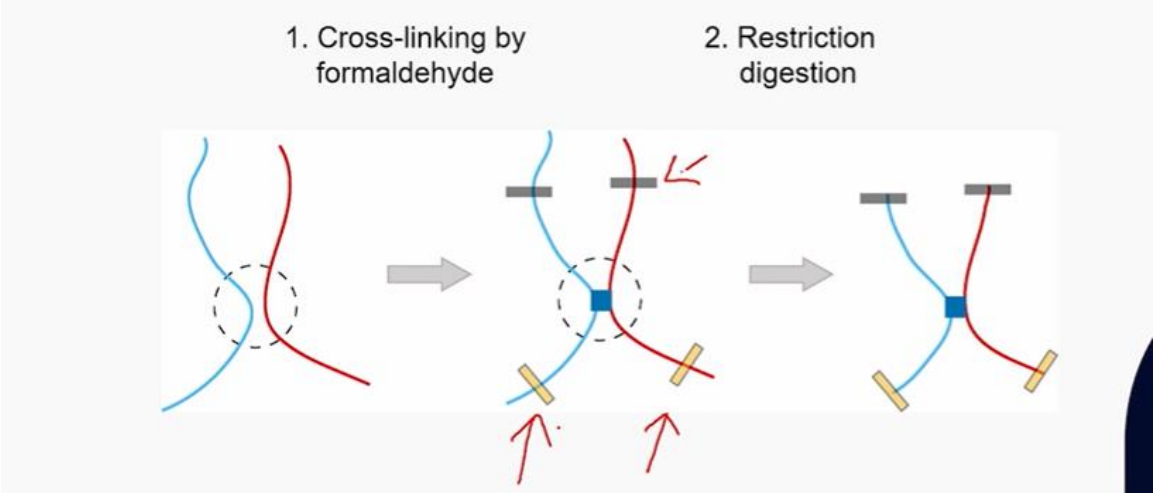
So, this actually kind of keeps the TAD in place or maintains the structure of the TAD. Of course, there are a lot of studies going on, but this is kind of the working model right now, ok? So, how do you study this genome organisation in 3D space, or how did we come to know about all these terms and aspects of genome organisation? So, TADs, LADS, etcetera, ok? So, what kind of method did you use or did researchers use to actually understand that? This is something we will discuss now.

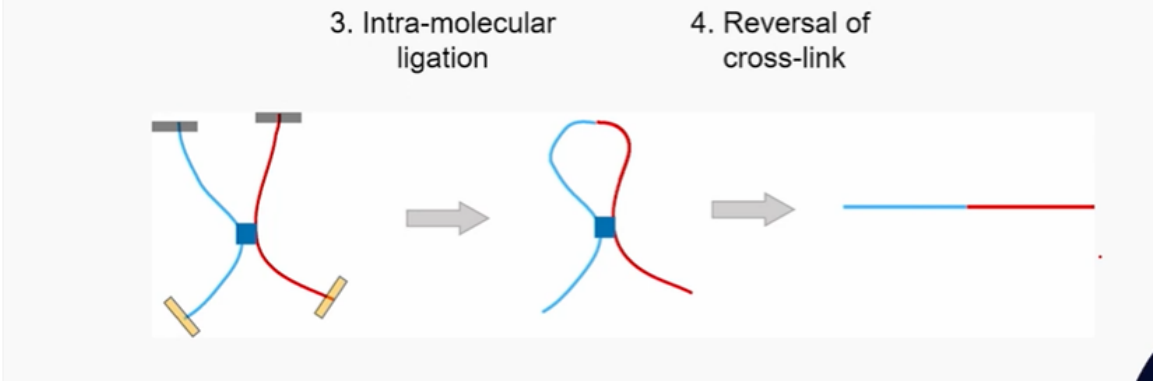
Here are several methods that I will discuss very briefly. So, we will start with the 3C and then the other methods that are based on the 3C, and then there is the final method, which is the ChIA-PET method. Again, we will discuss very briefly all these methods. So, 3C is the first method that was actually devised to study this three-dimensional genome configuration. So, the full form of 3C is the capturing chromosome confirmation, as you can see.

So, it is called 3C, and this is the reference paper that appeared in 2002. Very briefly, we will discuss the steps of how this works, and then, of course, we will look at the different variants or the different 3C-derived methods, those 4C, 5C, and IC, and we will talk about their advantages and disadvantages, and finally, we will talk about the ChIA-PET or ChIA-PET method. So, the first step is cross-linking by formaldehyde. So, we have come across this before, right? So, imagine these two regions: the blue region and the red region right.

They are interacting here, as shown in this circle right. This is the region where the interaction is happening, and before we prepare the sample for sequencing, we want to preserve this interaction.

So, to do that, we do the cross-linking by formaldehyde, and this cross-linking actually preserves that interaction, and this cross-linking happens mostly due to proteins. So, this is protein-mediated cross-linking. The proteins that are present in this region will form this cross-link. Once you have done the cross-linking, the next step is restriction digestion. So, you have seen that I have highlighted some of these restriction sites, and this is where you can do restriction digestion. So, these sites can be the same or different again, depending on your experimental design.





So, the next step is actually intramolecular ligation. So, once you have isolated this fragment by restriction digestion, the next step is intramolecular ligation. So, what we will do is join one end here, ok? Now, once this end is joined, we actually reverse the cross link, remove this cross link, and then the DNA is linearized. Now, the next step, sorry, is that once we linearize the DNA we can then amplify it. We can use these primers in black, and we can detect it by PCR amplification or by quantitative PCR.

So, if two regions are interacting, you will see a PCR band or a signal in quantitative PCR, and then you will infer that there is an interaction between these two genomic regions. Now, as you realise, this method can be applied when you have a hypothesis. So, you want

to investigate the interaction between two genomic segments, you know? Let us say you hypothesise that region X and region Y are interacting okay. And then to test that, you apply this 3C method, you design primers that bind to regions X and Y, and then you do the 3C, and when you see whether there is some sort of signal in the quantitative PCR at the end, ok.

And based on that, you draw inferences about whether there is interaction or not. So, as you also realise, this is mostly a one-to-one interaction right? So, we are testing the interaction of one genomic region with another genomic region. So, this is a one-to-one interaction, and the scalability is poor. So, we cannot scale this up very easily across a large number of genomic regions.

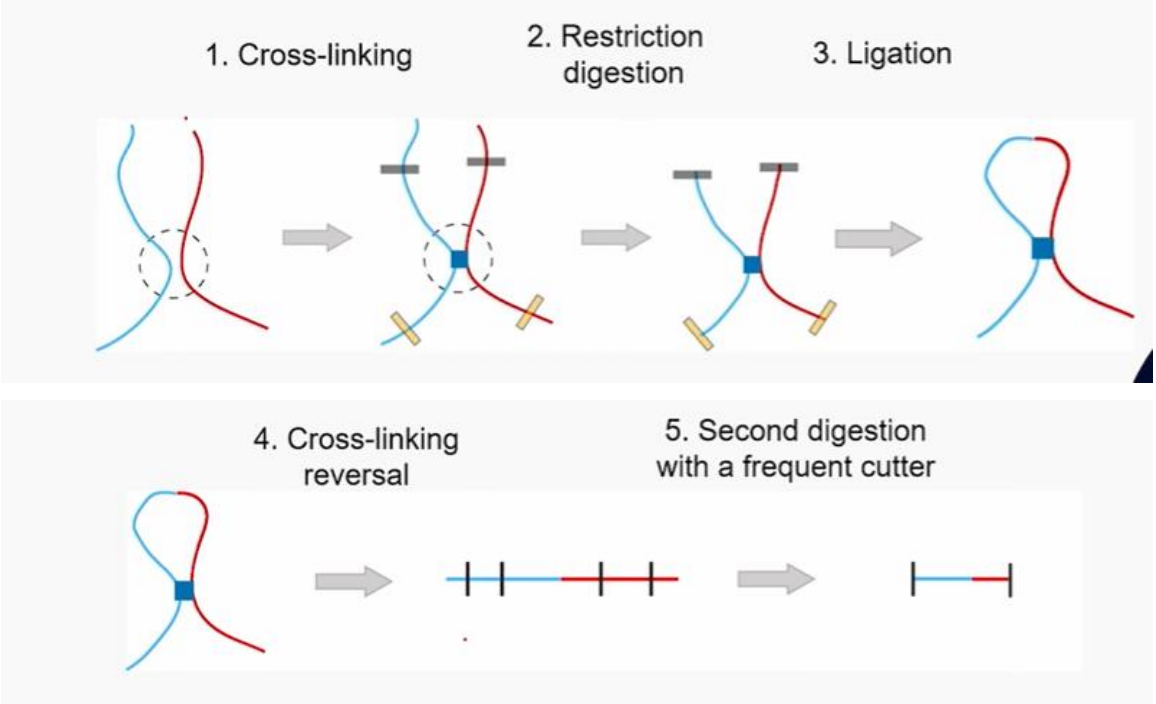
So, if you want to test this for 1000 regions, So, it may be just one region with 1000 other regions; this becomes very difficult to do. So, you have to design these primers; you have to do all the quantitative PCRs and that becomes really tedious, ok? And 1000 is already a small number, right? So, for a big genome like the human genome, you can have millions of these interactions, and you cannot scale this up. And then also, the results are dependent on the restriction enzymes that we use to some extent, but you can counter this by using different restriction enzymes and looking at the interaction patterns.

So, to counter this scalability issue and this one-to-one interaction. So, the question is: can we then identify genomic regions that interact with a specific genomic region? So, specifically, the example that I mentioned is right. So, we take one genomic region, and we want to test whether there are 1000 other genomic regions that interact with this region. So, this is something that would be very interesting. Let us say we have a gene that is very important for some purpose, and then we have the promoter of the gene, and we want to see whether the promoter region of the gene interacts with other genomic regions, which might influence the expression of this gene.

So, maybe instead of testing this 1000 other genomic regions, what we can do is ask whether we can explore right or unbiasedly discover regions that are interacting with that genomic region. So, this target region or this promoter that I just mentioned would be the bait or the viewpoint region, ok? These are the terms that are used in this analysis, okay?

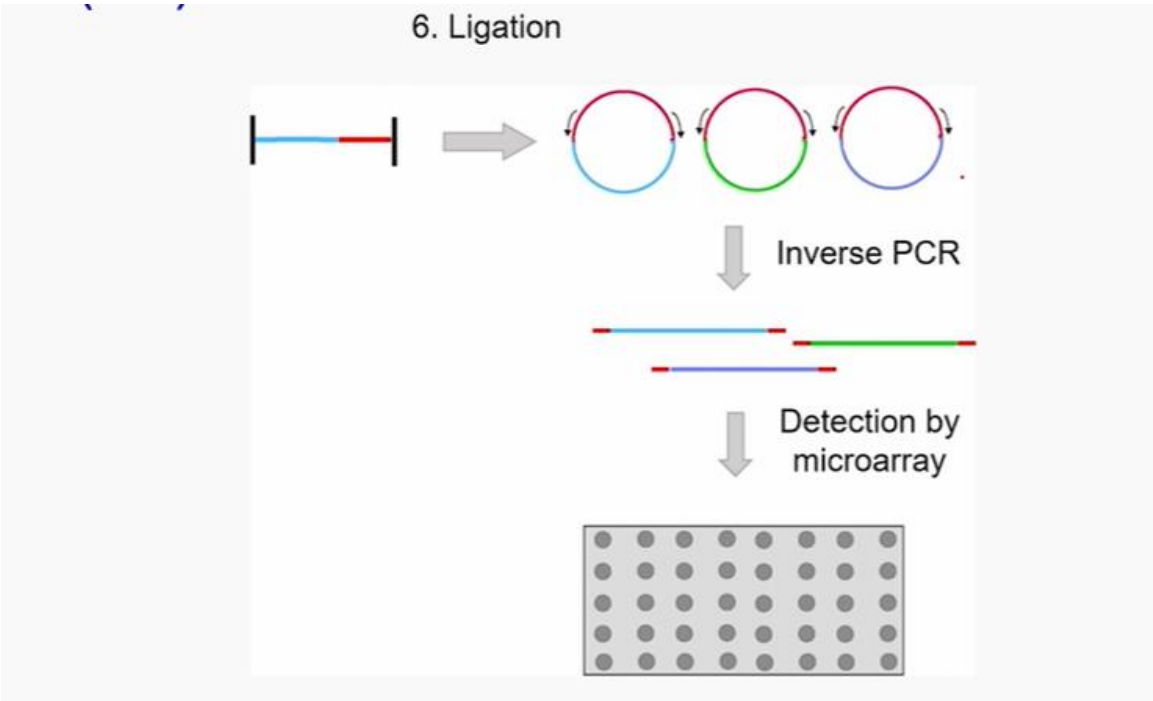
And there is a method that actually can enable us to do that, called chromosome confirmation capture on chip, or 4C. And this is the paper that actually described this method in 2006, and this method allows for an unbiased search for genomic regions that interact with a specific locus.

So, this specific locus is termed a bait or viewpoint. So, here are the steps for this 4C method. So, the first few steps are exactly identical with 3C, right? So, the first step is cross-linking. So, we want to preserve this interaction. The second step is restriction digestion. We are taking this fragment out of the genome, generating these fragments, and then we have ligation.



So, we are generating this or ligating these fragments back, but then the difference starts, ok? So, after this ligation, we cross-link and reverse that. So, we remove the cross link, we linearize, and then we have a secondary digestion with another restriction enzyme. So, I have marked this restriction site by these black parts. The difference from the first restriction digestion is that this second enzyme is a more frequent cutter, which means it

will cut more frequently in the genome.



So, usually the first restriction is an enzyme, which is a 6-base pair cutter right, which is rarer in the genome, and the second one is a 4-base pair cutter. So, these are much more common in the genome, ok? So, we take a second restriction enzyme to actually do the digestion. So, the idea is to generate smaller fragments after this restriction digestion. Now, once we have the smaller fragments, we can then ligate back and circularise this fragment.

So, we can ligate these ends, and we can circularise this fragment. Now, what one thing do you probably realise right now?So, we are doing this in parallel for different genomic regions, right? So, instead of just starting with this one interaction we have multiple interactions, and for each of them, we will have this kind of circularization this digestion, etcetera. So, at the end of circularization, we will get these different types of fragments, right? I am showing them in different colours.

So, these are the regions that interact with the red region, okay? So, here, the red region is the bed or the viewpoint. Now, once we have generated it, we know the bed sequence, right? So, this is the specific locus that we are interested in, ok? So, we can then design primers for the bed. So, here is the bed sequence, right? I can see these primers here in the black arrow, and we can do something called inverse PCR. This is not normal PCR; you

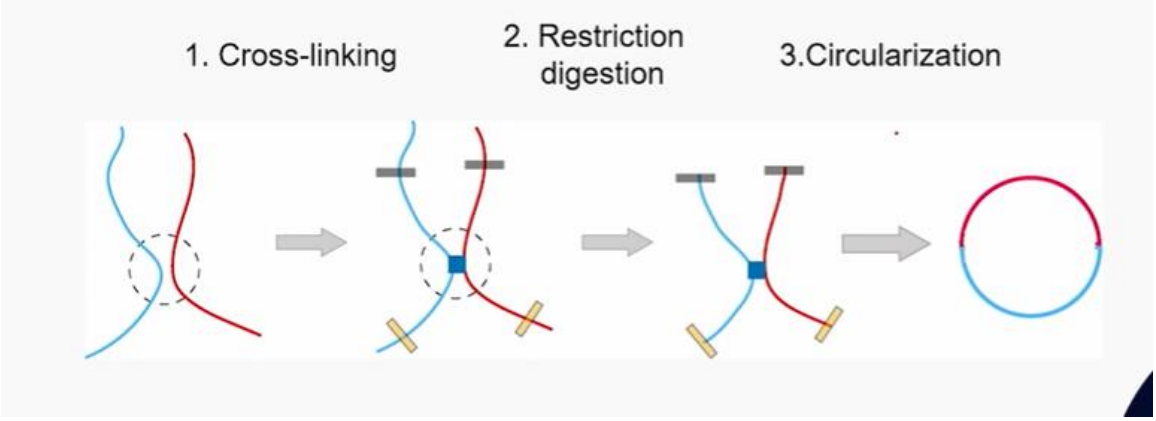
can see the primers are pointing away from each other, and this is inverse PCR, which will actually generate these amplified fragments containing the other fragments.

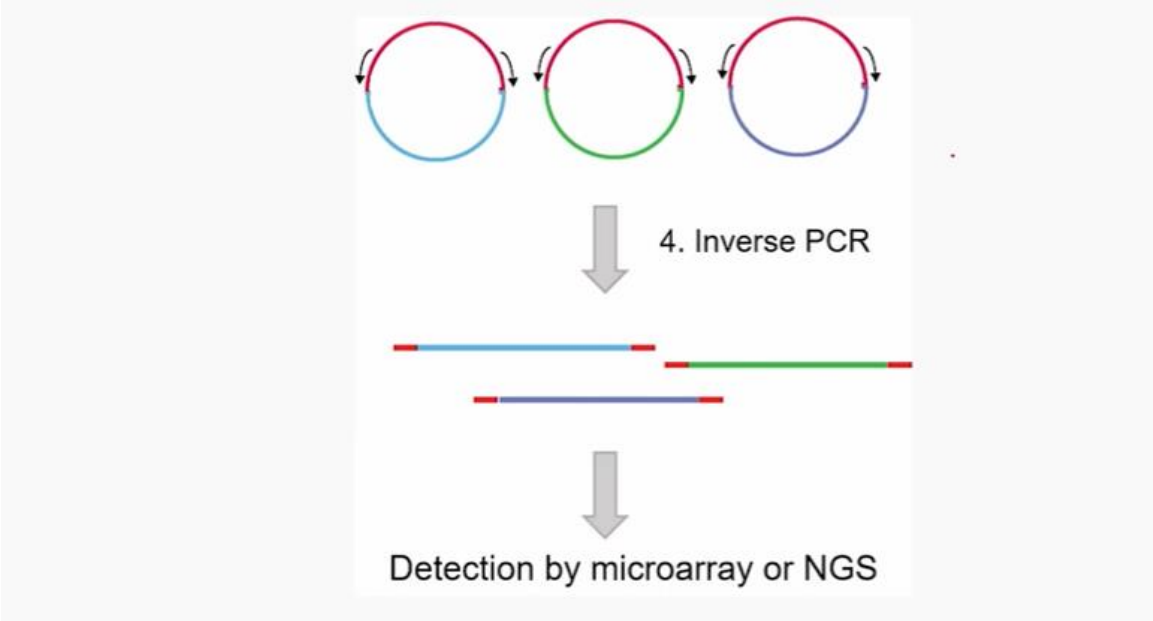
So, other than these blue regions or green regions, So, these are the interacting regions with the bed, okay? So, once we have generated this, we can then simply detect it by microarray. So, earlier, we used to detect by microarray; now we can also say we can apply exaltation sequencing here, ok? So, there is a variant of this 4C, which we will discuss again.

So, just to summarise what we have discussed, So, we have to again compare with a control because this is a microarray experiment. So, we will get some sort of enrichment, which we will have to compare with a control. So, in this case, the control is restriction enzyme-digested genomic DNA, and then we see enrichment of specific regions compared to the control that indicate interaction with the bed region. There is a variant that, as I mentioned, is called 4C or 4C-Seq. The full form is slightly different and is called circular chromosome confirmation capture, or sometimes also referred to as circular 3C.

And this is the paper that actually proposed this method. So, the protocol is slightly different. So, let us look at that. So, the first step is cross-linking, and as before, the second step is restriction digestion. And the third step is circularization, ok?

So, there is no linearization, then a second restriction digestion. So, this is actually a much simpler process; here we have circularization. So, we circularise this frame. Now, you can again imagine that this is happening for multiple interactions on the bed, and we get these different circular fragments, and we can again do the inverse PCR directly. And once we have done the inverse PCR, we can then detect it by macroarray or we can apply NGS methods to actually identify these regions. So, this will tell us which regions are interacting with the red locus or red region.

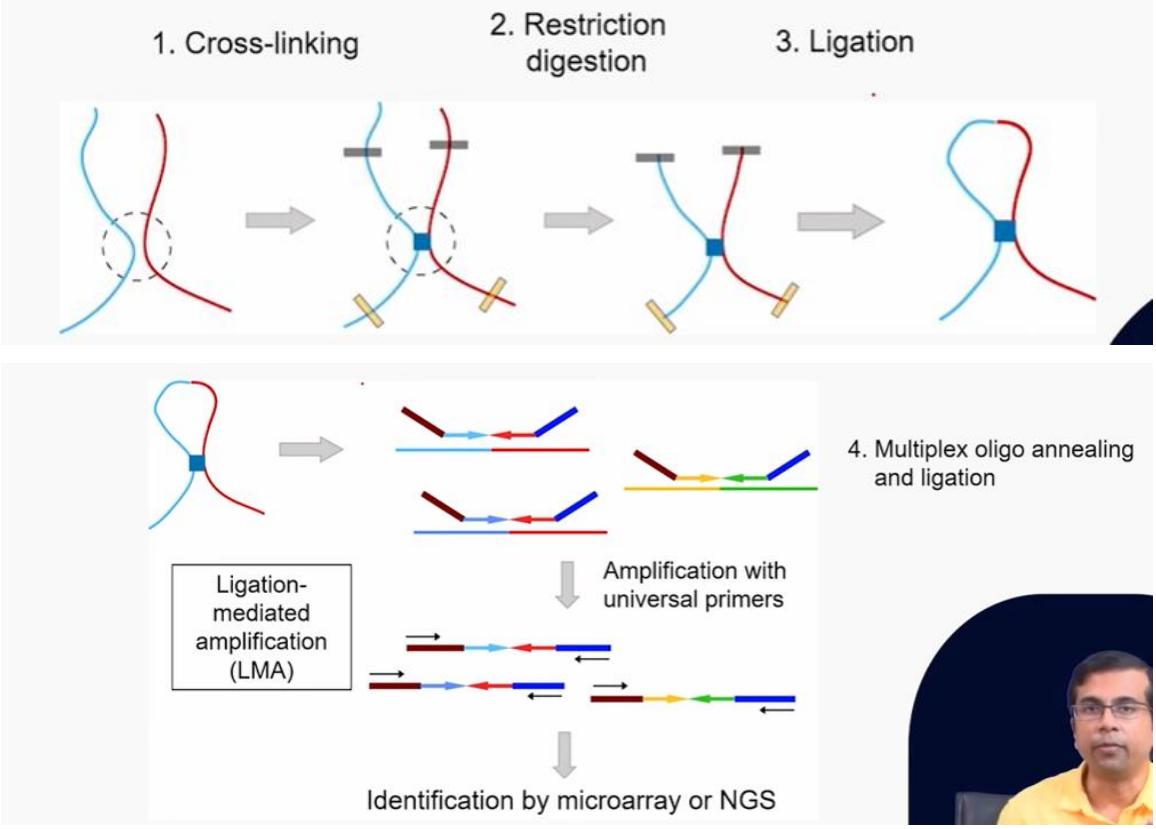




So, what you see from this discussion is that this is 4C, which is one-to-all mapping. So, we have one bed or viewpoint sequence, and we are looking at the interaction of this region against all genomic regions. So, we are identifying genomic regions that interact with this bed region. So, this is a one-to-all mapping.

Now, the idea is, can we extend it to many rights? So, instead of just having one bed, can we apply this kind of method to many beds? So, can we do this in parallel for multiple loci, not just one locus? And this actually led to this method called chromosome confirmation capture carbon copy, or 5C method. So, as you can see, we are getting 3C, 4C, and 5C, and then finally, we will get to its high C. So, this is a many-to-many mapping, and this is the paper that actually suggested or proposed this method.

So, let us go into the process of how it is done. The first few steps are exactly identical to 3C, right? So, you can see the cross-linking of restriction, digestion, and ligation. So, you can start with the 3C process, and then you can try 3C or 5C, depending on your needs. So, here, the difference is that instead of linearization, etcetera,



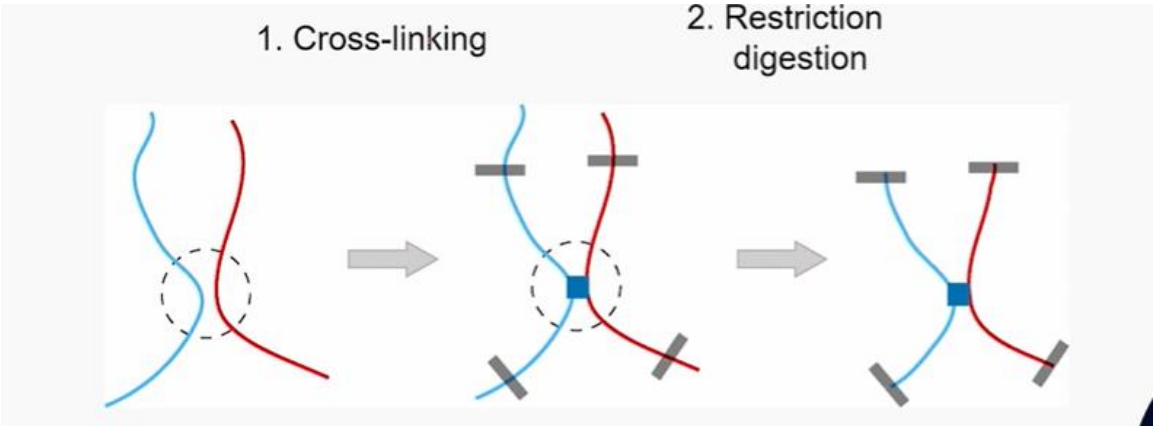
So, linearization happens, right? So, after the linearization what happens is something called multiplex oligo annealing and ligation. So, what is this multiplex oligo? So, you have this pair of oligos right that can bind or recognise different regions of the genome, and then they can anneal to those fragments right of the genomic regions and also ligate with each other right. So, there is this ligation between the oligos, ok? So, what will happen because these are multiplex oligos means they contain a combination of different oligos that can bind to different genomic regions, and so, they will anneal right based on their complementarity to different genomic segments, and they will also anneal anneal, and then they will also ligate. They are designed in such a way that they will ligate with each other.

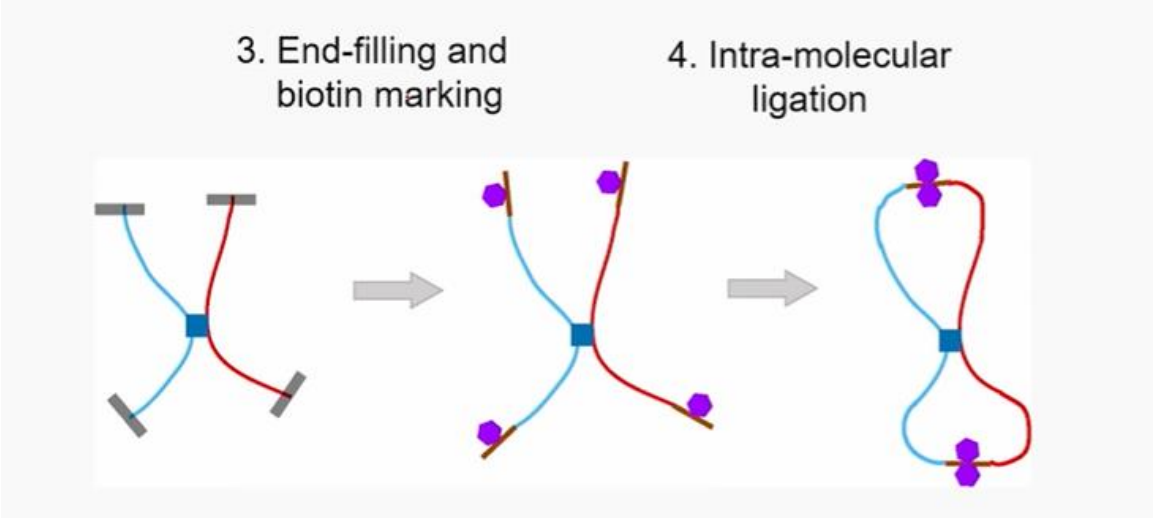
So, once they ligate, right? So, then, what you can do is amplify with universal primers. So, what you see is that even though they contain these multiple oligos, these n regions are

identical for each of them. So, these are the regions where we can use universal primers that will bind to all of these combinations. And then we get these amplifications, which we can then sequence by negotiation sequencing or identify by microarray.

So, what you have seen here is many-to-many mapping, right? Why are we not saying this is all-to-all because we are limited by the multiplex oligos? We can contain only a certain number of combinations in the oligo library, okay? So, we cannot have all oligos in the library. So, this process is limited by this oligo design.

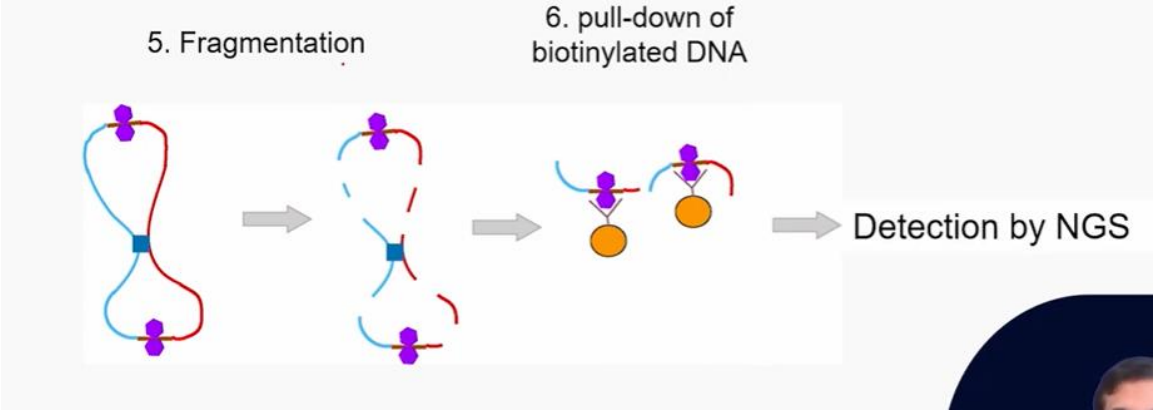
The final method that we will talk about is Hi-C in detail, right? So, this is an extension of the earlier methods; this is all-to-all mapping, and this is the paper that actually proposed this method. So, here is the protocol. So, we start with cross-linking, then do the restriction digestion. The next step is something called end filling that happens at the end of these fragments and a biotin marking is OK. So, this is a chemical molecule that is attached to the DNA, because this will help in further purification of the DNA fragment.



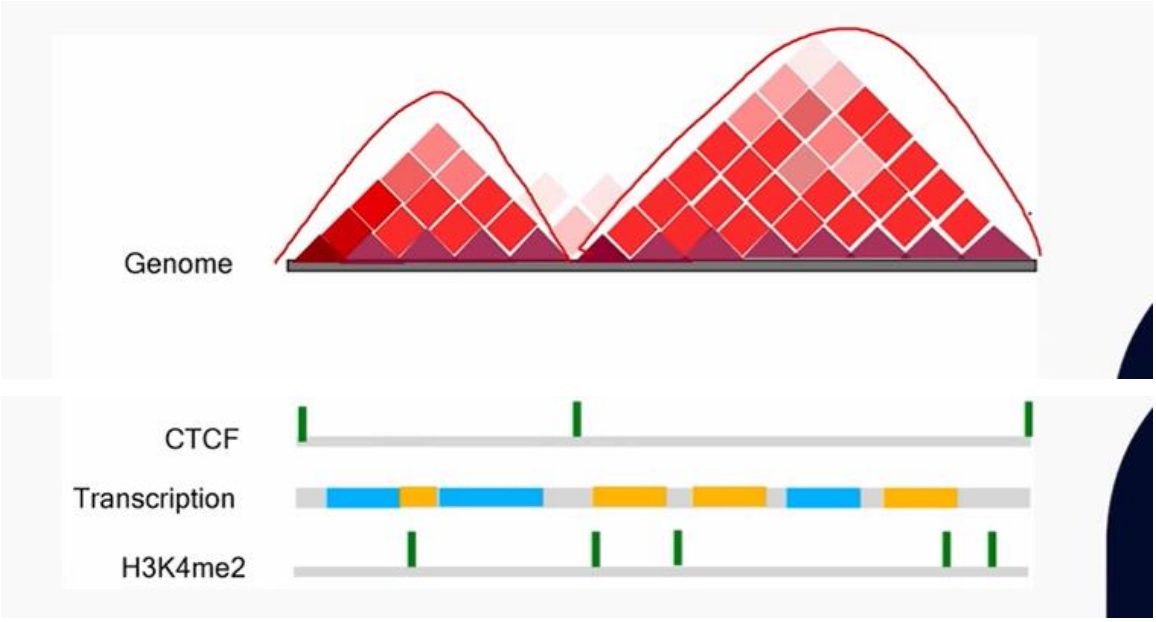


So, the next step is intramolecular ligation. We ligate these fragments using a ligation enzyme, we fragment, and then we pull down this biotinated DNA using a molecule called streptavidin. So, the biotin-streptavidin interaction is very strong. So, this will pull down this biotinated DNA. And once you have these fragments, we can then separate them out, and then we detect these fragments by NGS. As you can see, this is all-to-all mapping, Hi-

C, because we are not limited by any oligo design, etcetera.



We will get all sorts of fragments that are marked with biotin, and they will be sequenced by the negotiation sequencing method. But it is dependent on the restriction enzyme that is used, but this is easy to overcome because you can do the experiments with different restriction enzymes, and then you can compare the results and see if you get similar results. And this has been applied for mapping interaction within chromosomes and between chromosomes in the human genome, and here is the reference you can look up. Now, one of the things that we will get from the Hi-C map is the strength or frequency of interaction between genomic regions, and this will look something like this. If you have the genome for different genomic regions, you can see the interaction in different colors; these interaction strengths will be represented by different colours.



So, darker colour means stronger interaction lighter colour means weaker interaction. So, as you can see, there is this region that shows interaction within itself. So, this is likely to be a TAD. Here also, you have this other region that shows very strong interaction within itself, and this probably is a tad right. So, just to interpret this square here, this actually looks at the interaction between this genomic region and this genomic region, right? So, you can then interpret this across all the squares and look at the interaction between different genomic regions.

So, you can overlay this with chip-seq data. So, you can do chip-seq on the same sample and overlay with chip-seq for, let us say, the CTCF binding site, and maybe here are the peaks for CTCF, meaning we are looking at tads, right? So, remember that the tad boundaries are marked by CTCF. So, this actually aims to provide information about the tads. You can also look at the transcription profiles of these regions, which you can then classify into A compartments and B compartments. You can also look at histone modification patterns again by CHIP-seq or CHIP-exo, and you can then actually overlay these with a lot of things. So, this kind of helps us in understanding these interactions, and this kind of figure you will see across all the interaction maps that you come across in the references.

The final method I will just very briefly mention is called ChIA-PETor ChIA-PET. I will not discuss the protocol. So, this is the full form of chromatin interaction analysis by paired and tag sequencing. Here is the reference. So, what this method does is find interaction sites that are mediated by specific proteins or transcription factors. So, on top of these interaction maps, you also get the proteins or transcription factors that are mediating those interactions.

So, you can think of this as Hi-C coupled with chip-seq. So, this Hi-C method is so powerful that it has now been extended to single cells, and you can get single-cell Hi-C. So, this is the paper that actually shows this cell-to-cell variability in structure, chromosome structure, and this is something that you can actually now study with single-cell Hi-C. Another project I will mention is called the 4D nucleon project. So, this is looking at the 3D organisation of the genome and how it changes with time, which is the fourth dimension.

Here is the link to the project, and here is the paper for this 4D nucleon project. So, these are the references for this class. To summarise, we have seen that the genome has distinct 3D organisation, and some genomic regions interact more frequently within themselves than with other regions. We have seen that these are called TADs, and these structures actually impact transcriptional activity. We have discussed different methods (3C, 4C, 5C, and Hi-C) that enable us to detect the interactions between genomic regions in three-dimensional space. Thank you very much.