Next Generation Sequencing Technologies: Data Analysis and Applications

Applications of NGS in Epigenomics

Dr. Riddhiman Dhar, Department of Biotechnology

Indian Institute of Technology, Kharagpur

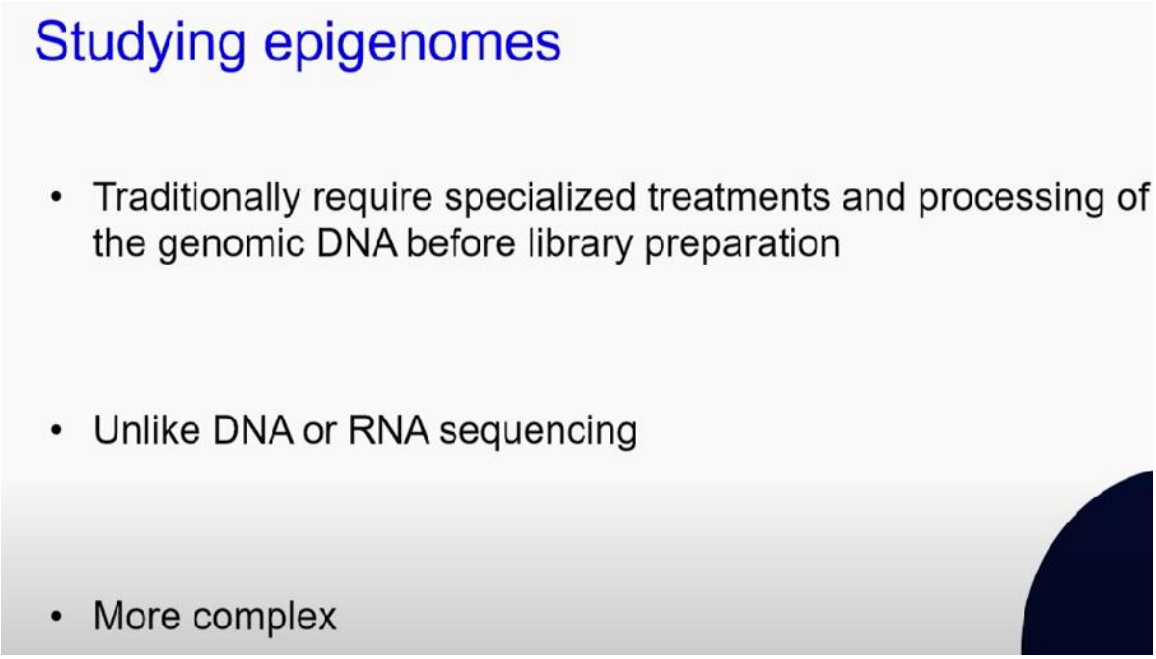
Good day everyone. Welcome to the course on Negation Sequencing Technologies, Data Analysis and Applications. We are into the final week of the course and we have discussed the genome sequencing part, identification of single nucleotide polymorphisms and transcriptome sequencing. What is remaining is the application of NGS technologies in epigenomic studies and this is what we will be discussing in this class and the last few classes. So these are the topics that we will be covering today. So we will mostly talk about epigenetic modifications because it is important to understand which what modifications we are looking at or what modifications we want to identify using the NGS methods.

We will talk about epigenomics and then we will talk about or briefly introduce this application of NGS for epigenomics and in the subsequent classes we will actually expand on this application part a lot more. So these are the keywords for this class DNA methylation, histone modifications and nucleosomes. So going back to the flowchart for NGS data analysis, we started with the read data from sequencer and this is common for all applications right. We looked at different data formats.

The first step is the quality control, we talked about that and then we talked about read mapping process and we applied that this read mapping process for identification of single nucleotide polymorphisms in their structural variants. We also talked about transcriptome analysis again mapping is the first step in that process. On the other side we talked about read assembly in last few classes. So last week we talked about the assembly algorithms, we talked about tools and we then again the new get a new reference genome after the assembly process. So we have discussed all this and what you what you see the epigenomic part is remaining and this is something what we are going to discuss in the next few classes.

So the question is why did I left this epigenomic studies for the last part right. Why did not we discuss after the transcriptome analysis? So I will start with that. That is because this studying epigenomes it actually traditionally requires specialized treatments and processing of the genomic DNA ok. So all types of applications that we talked about there is some sort of experimental step right. So we isolate DNA, we then fragment DNA, we add adapter, we then do amplification etc.

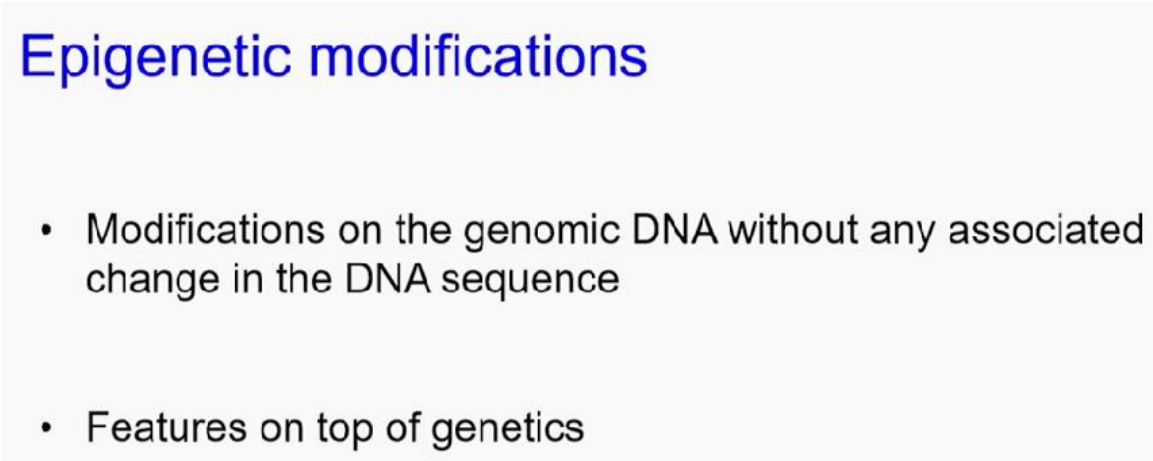
For RNA-seq we again convert to cDNA etc. Then prepare the library for sequencing through adapter ligation etc. And then do the sequencing. For epigenomic studies the DNA has to be treated chemically through different chemicals right. So this is something we will discuss again which means it adds to the complexity of the whole process ok.



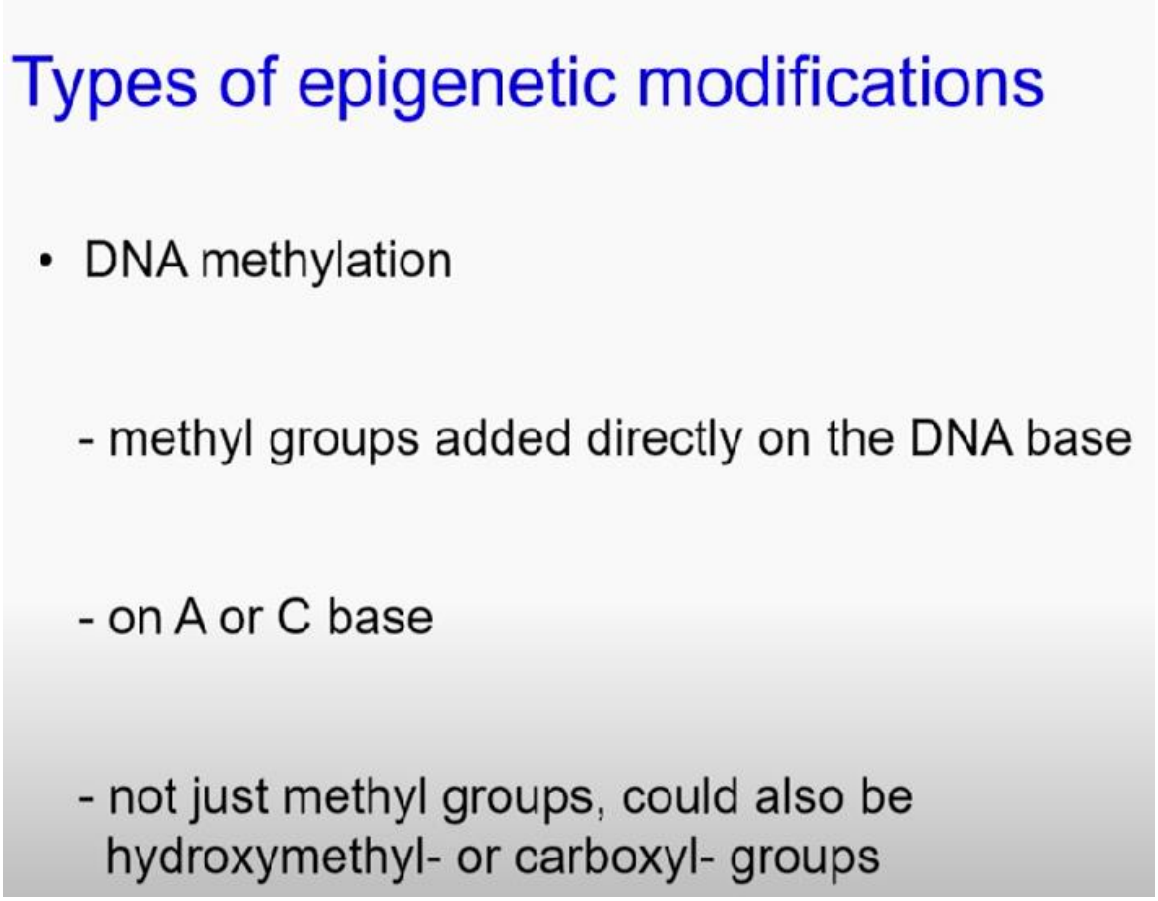
So this is unlike DNA or RNA processing which still have some experimental step, but these are simple steps. Epigenomic studies they require this specialized treatments which means we need to have controls and this makes the experimental process more complex, but also the data analysis process more complex because you need to have more control experiments now ok to account for these different biases that will be introduced during the experimental steps ok. So this is the reason why I chose to keep it at the end because we will also need to talk about the experimental steps and their implications on the data analysis. So let us now start with epigenetic modifications right and what are epigenetic

modifications? What are the different types of epigenetic modification? That is something I will introduce ok. So epigenetic modifications are modifications on the genomic DNA that are not associated with any change in the DNA sequence ok.

So these modifications would not change that DNA sequence right. It will not change this ATGC base. It will simply add on something to the genomic DNA right. So it will add something extra to the genomic DNA and that is why it is called epigenetic right. So it is features on top of genetics ok.



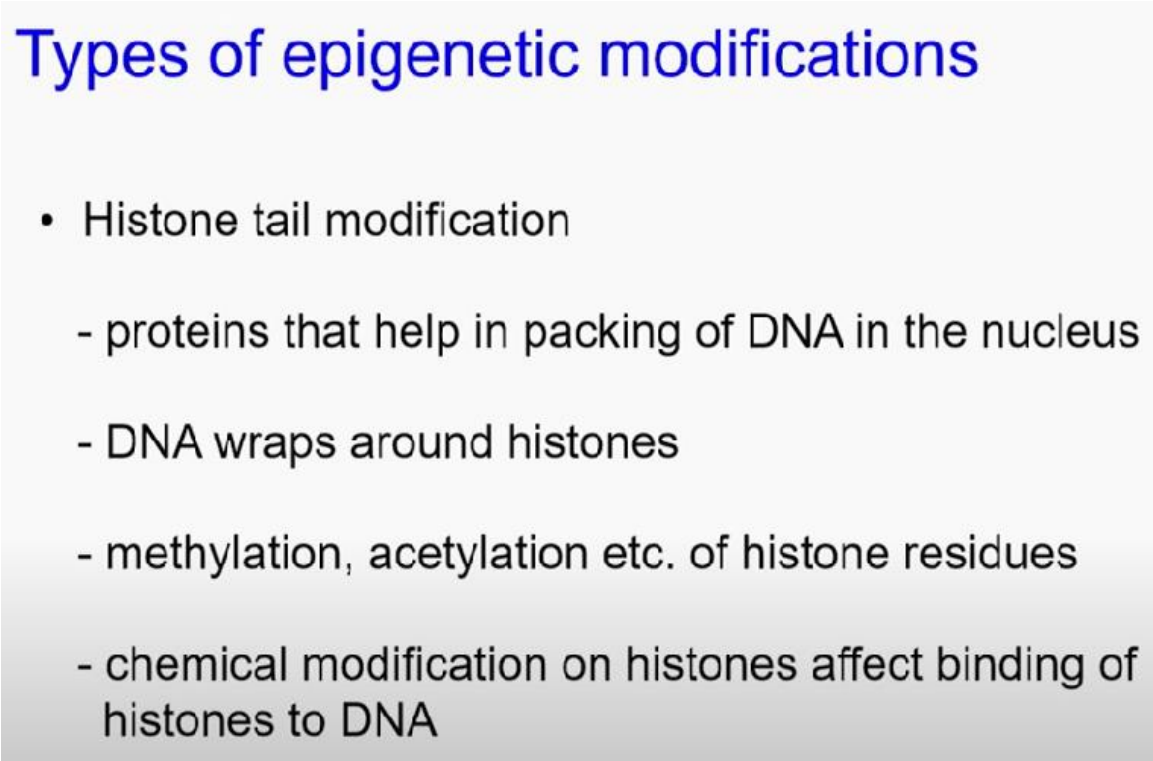
So it is not changing the sequence itself. It is adding something on top of the genetics on top of the genomes ok. So there are different types of epigenetic modifications and the first one is very common one is called DNA methylation ok and this is something you probably all are aware of right. So these are methyl groups that are added directly on the DNA base right. So if you have the DNA strand right you have this methyl groups added to the bases and most of these additions happen on A or C base right.



This is commonly studied and we have seen this in prokaryotes as well as in eukaryotic systems. And also what we see now is that these are not just methyl groups that could also be hydroxymethyl or carboxyl groups that are added on top of these bases ok. We will talk about this in much more detail in the subsequent class right where we will actually specifically focus on this DNA methylation. So it is important to know these modifications because then you can devise strategies to identify these modifications using the next generation sequencing methods ok. As you can see without chemical treatments or experiments right we cannot directly detect this ok at least with the traditional next generation sequencing methods.

But the newer techniques the single molecule sequencing methods they can actually help us identify them directly ok. So we will talk about this again in the next class where we will talk about how we will identify these modifications if we are using the Illumina platform or how we can identify these modifications if we go through the nanopore or the

SMRT platform. You can also have something called histone tail modifications ok. So histones are proteins that have been packing of DNA in the nucleus and these DNA actually wraps around histones without going into too much detail right all those nice figures you probably can see in the textbooks. And you can have these histones being modified at certain residues ok.

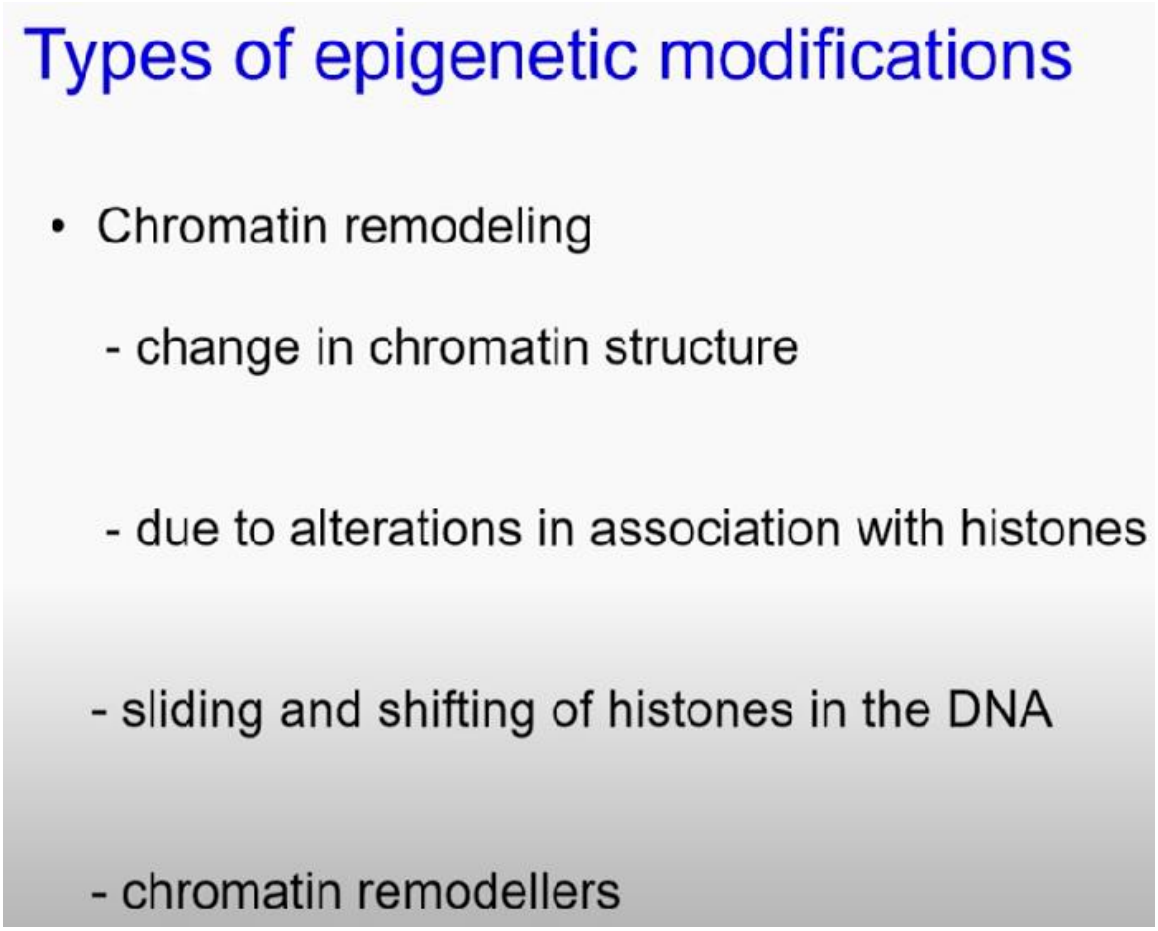


So you can have methylation, acetylation etcetera on the histone residues ok. And there are specific type of modifications which I will just mention one or two of them. For example, histone 3 lysine 4 residue this is modified histone 3 lysine 36 these are modified ok. So again there are specific bases specific places or specific residues of histones that are modified and you will see methylation or acetylation or even other types of modifications. Now what happens is that when you have this chemical modification of histones they affect binding of histones to DNA ok.

So this modifications are very important for controlling this binding of DNA to the histones ok. So how tightly this DNA will bind to the histones will be affected. You can also have something called chromatin remodelling ok. So chromatin remodelling ok, so this means there is change in the chromatin structure ok. And this is because the alterations

or changes in association of DNA with histones ok.

As I just told you right DNA wraps around histones. Now if you have changes in this association right if that histone is coming off from the DNA or the histone is going in and binding to DNA right that will actually lead to change in the chromatin structure ok. And this is what we call as chromatin remodelling ok. And this can also happen due to sliding and shifting of histones in the DNA right. So you can have association, dissociation, you can have sliding, shifting etc.

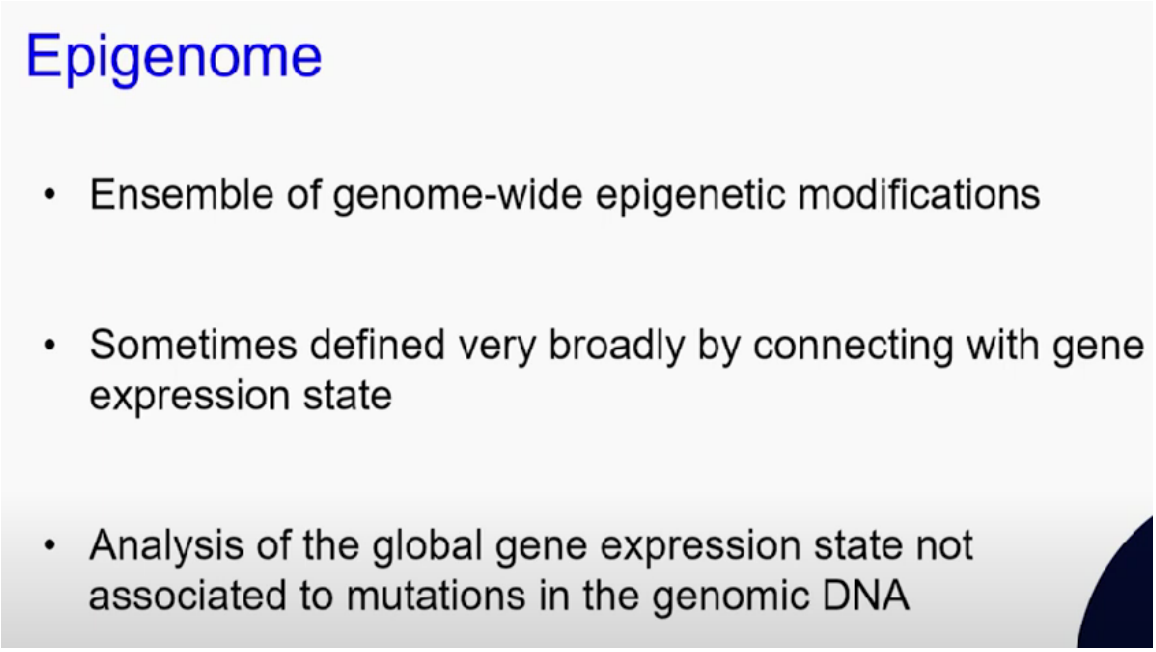


And this is usually done by some proteins which are called chromatin remodellers ok. So again without going into the details I will not mention the names of those, but then again these proteins exist in all organisms all higher equality organisms for example, in humans also ok. So now we understand these different types of epigenetic modifications and we would like to identify these modifications right. So that is why we discussed right what are the different types of modifications that would be targets for our NGS methods right.

Now as I said like we want to talk about epigenomes right.

So the question is what is epigenome? So epigenome is simply the ensemble of genome wide epigenetic modifications right. So we have these different types of epigenetic modifications and what you want to do when we are studying epigenome is we want to look at all these types of modifications across the whole genome ok. So if you take the full genome and you want to identify all these modifications that are present in there. Now one of the things you probably realize is that these modifications these are actually transient ok. These are not always permanent like that like DNA sequence that it will be there and that base will not change ok.

So these modifications are transient they can change as we just mentioned right you can have this chromatin remodeling. So again this is changing time and similarly you can have this histone modifications or DNA methylation that can also change ok. You can if you have proteins or enzymes that will add these groups and there are proteins and enzymes that will remove these groups ok. So again this actually complicates the full analysis right because there is a time component here. Now also epigenome is sometimes broadly defined.



So this is also connected to the gene expression state. So as you can realize and we will

discuss a bit later is that all these modifications on the DNA they have some impact ok. It is not like they are just sitting there doing nothing they have some impact on the expression state of the cell ok. So this means there is a connection between epigenome and the transcriptome ok. So by looking at the epigenetic modifications you can sometimes say which gene will express at high level or which gene will not be expressed ok.

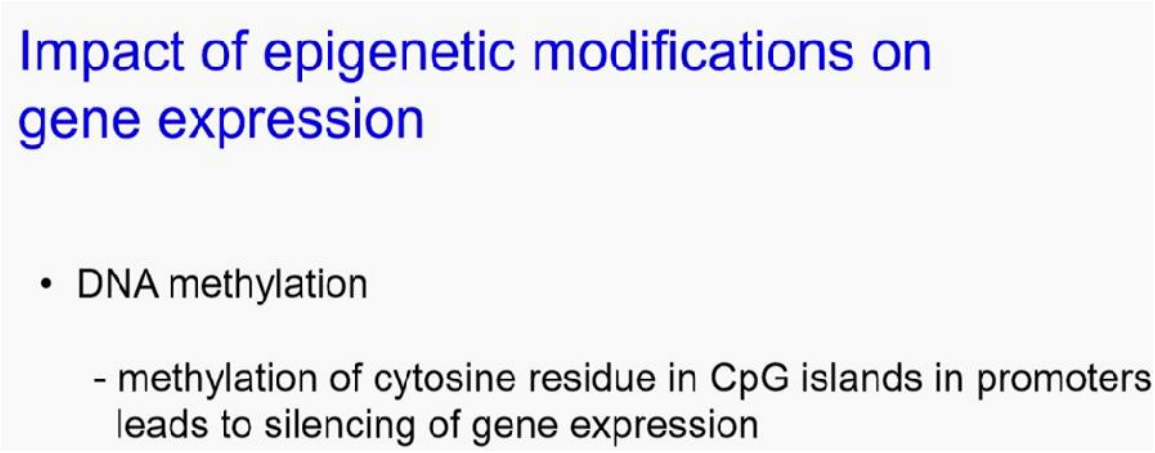
And we can actually study these connections if we look at the epigenomic data and transcriptomic data from the same cells ok. So that is why epigenome is actually connected with gene expression state and it can be simply the global gene expression state that is not associated to mutations in the genome. So if you take mutations out of the picture so because if you have mutations in the promoter regions of the gene this will affect the gene expression state. So if you take out the mutations whatever changes are happening in global gene expression state that will be because of the epigenome ok. So what is epigenomics is simply the study of this epigenetic modifications at the genome scale as I mentioned right if you look at human genome right 3 billion bases.

We want to look at all the epigenetic modifications across the full genome ok. So this is a genome wide study and not limited to just local regions or specific genomic segments right. So it has to be high throughput in nature and this is where the next generation sequencing technologies coming ok. So the question might ask right why do we study epigenomes ok. So there are two parts right so of course, why do we study epigenomes and then the next part is how ok.

So how part we will just introduce in this class and in subsequent classes I will discuss them in detail right. How do we apply next generation sequencing methods to actually understand or to study this epigenomic variations ok. So the question is why do we study epigenomes what is the impact or importance right. So as I have just mentioned epigenetic modifications impact gene expression state of a cell and this is we now know and these combinations of these different types of modifications can have different unique effects ok. So we talked about three different types of modifications right.

So we talked about DNA methylation, histone modification and chromatin remodeling. So you can take any combinations and these combinations will have unique impact on the expression level right. So this is what we want to understand right how epigenomes or how epigenetic modifications across the genome impact the expression levels of the cell ok. So this is something now we realize right this is actually very important for gene expression and I will just illustrate with some examples how actually these modifications impact gene expression.

So let us take DNA methylation first. So these are direct modifications on the DNA bases and one of the very common modifications that we know is on C base in something called CpG islands. So these are C-G bases that are present in promoter regions of genes in especially in mammalian systems and methylation of this C base in this context it actually leads to silencing of gene expression ok. So this is something that is well known. So if you can study this CpG methylation you might want to you might be able to predict right which genes should be silenced in a cell. Similarly if you look at histone modifications and this has been studied by researchers that they are also associated with different aspects ok.

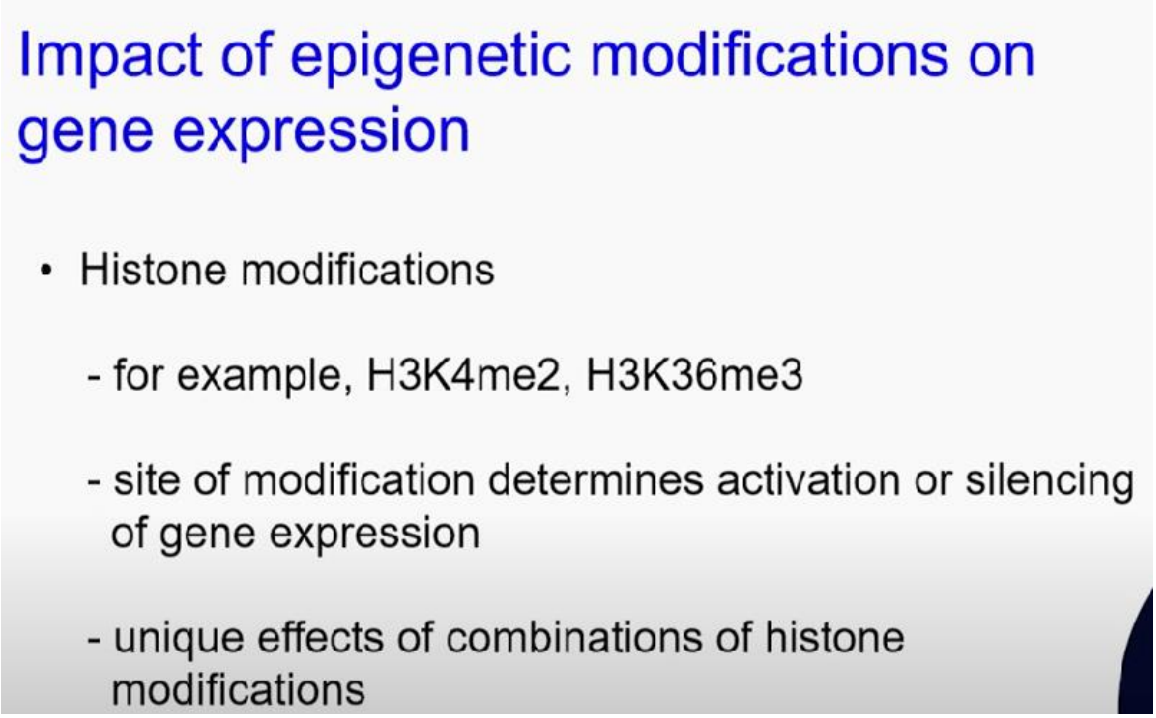


So I have just mentioned right there are different residues of histones that are methylated or acetylated. So here are two examples for example, you have HTK4Me2 or HTK36Me3 right. So HTK4 is the histone 3, lysine 4, Me2 means you have double methylation right. So dimethylation and K36Me3, 36th lysine you have triple methylation, 3 methylations ok. And you can have all sorts of combinations there are other residues which are also commonly modified and there they have been associated with specific gene expression

pattern.

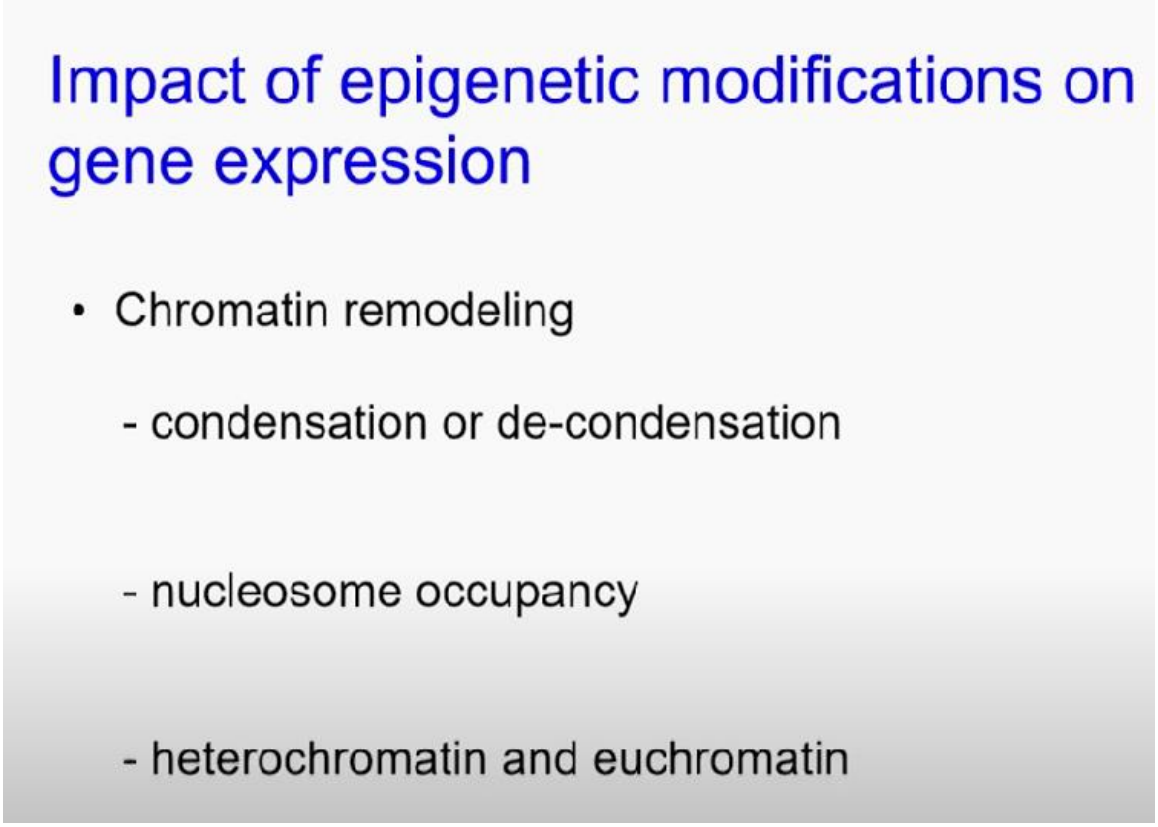
So some of these modifications associate are associated with activation of gene expression and some of them are associated with repression of gene expression ok. So again we do not have to remember all that all these combinations if you want to understand the biology or if you want to predict the biology right then you will remember this methylation pattern or this modification actually gives to activation gives rise to activation or this modification pattern leads to repression right. Our goal is to identify these modifications right and to see and across the genome right through the application of next generation sequencing technologies ok. So as I just mentioned site of modification determines activation of silencing of gene expression. So that is this is something we want to study across the genome and then perhaps relate to transcriptome or gene expression levels.

And what has been seen is that again you have this different combinations of histone modifications right. So you have HTK4Me2 with something else HTK36Me3 or some other combinations right. Again there are many sites which can be modified you can have dimethylation, trimethylation etcetera and you can generate these different combinations and there are unique effects of these combinations ok. And this is something that actually referred to as histone codes. So you can probably once you study this right once you associate these modifications with expression you can then perhaps predict with this combination what would be the expression state of a gene.

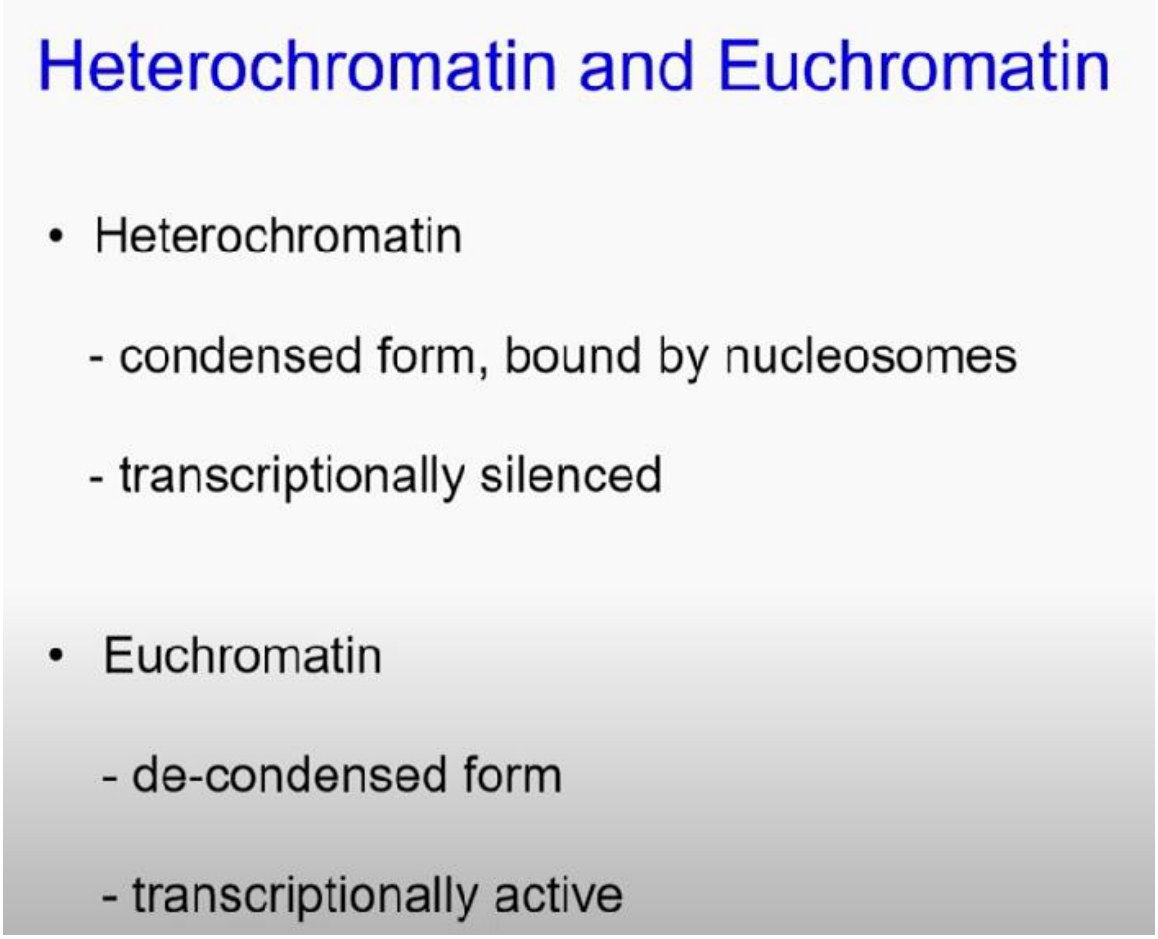


The third part which we talked about is the chromatin remodeling part ok and this is related to condensation or decondensation of chromatin. Again here this is because of the nucleosome. So nucleosome is the DNA bound with the histone so that gives rise to nucleosome and where you have nucleosome right. So that would be more condensed and this region is unlikely to be active right because for expression for transcription to happen a DNA has to be free right with the scope of the transcription factor binding RNA polymerase binding. But if it is kind of compact already because of the nucleosomes right that region is unlikely to be expressed ok.

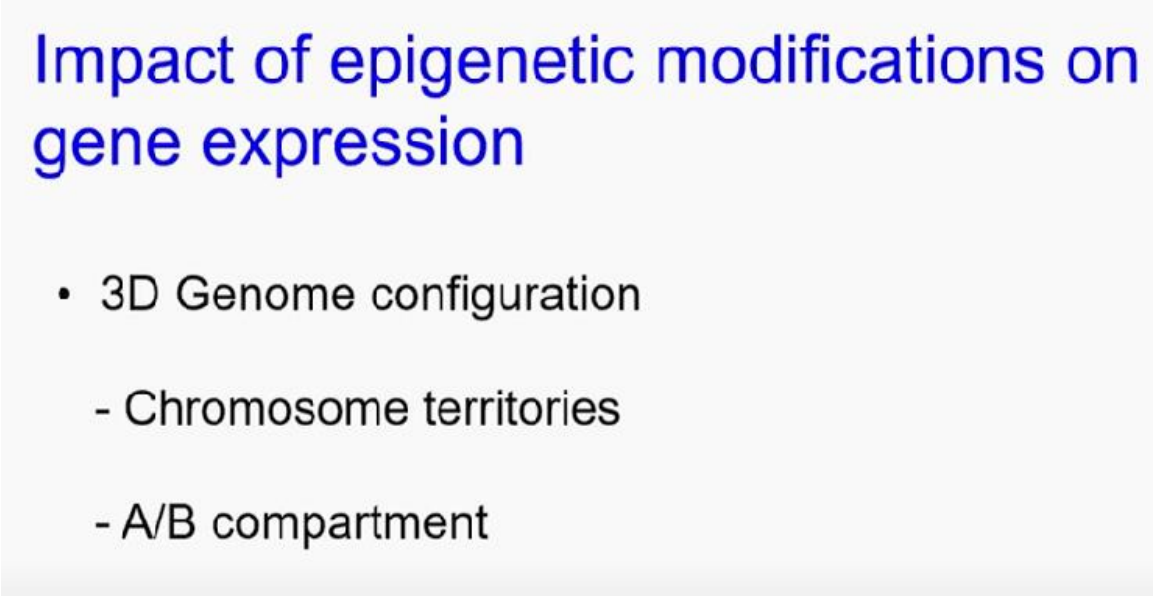
And that gives rise to two regions two different types of regions we call heterochromatin and euchromatin ok.



So what is heterochromatin? This is actually condensed form this is bound by nucleosome so you have all those proteins in there so this is in compact condensed form and this is transcriptionally silenced. And on the other hand we have the euchromatin which is in the decondensed form or sometimes we refer as open form ok and which means it is free to bind transcription factors and RNA polymerases and this is the transcriptionally active region ok. So as you can see we are learning a lot of concepts right so different types of epigenetic modifications and we are learning how these modifications right how they can impact gene expression ok. So this is the idea that we want to take forward from this class right so these modifications are important and they actually affect gene expression.



And I will also finally also give some examples of how the genetic modifications are also important for different phenotypes and behaviors of cells and organisms. There is another part another type of modification which we have not talked about this actually happens because of the 3D genome configuration. So this is something that is very new that has been studied being studied by researchers and we still do not have the complete picture of this right. So what do you mean by 3D genome configuration ok? So 3D genome configuration means you can imagine the DNA right so it is in the nucleus it also has a three dimensional structure ok. And it is not like everything is compact and everything sitting in one place etcetera this genome has three dimensional structure ok and this structure can impact the gene expression pattern ok.



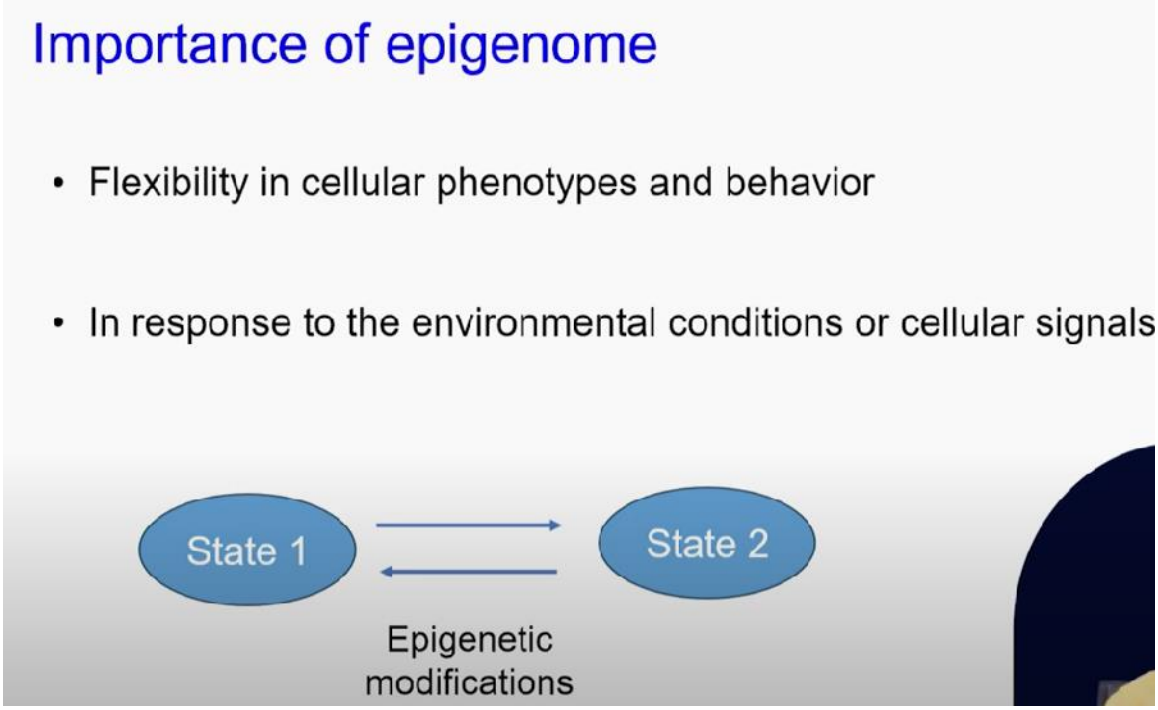
So this is what the researchers have seen now is that there are regions called chromosome territories. So if you are talking about an organism right where you have let us say 16 chromosomes right each chromosome will occupy a certain space inside the nucleus it is not like they are all jumbled up together they are separated and they take their own space ok. So this has been now studied so and there are also something called AB compartments there are other different components of this ok. And we will talk about them when you talk about this 3D genome configuration and how we can study this configuration through application of NGS technologies ok. So what you see now is these applications the epigenomics actually is a huge field it encompasses different aspects ok.

So it starts with the modifications which we talked about DNA methylation histone modifications, but it also is associated with the nucleosome occupancy it is not a chemical modification right it is actually something just association of histone with the DNA. And here we are simply talking about the three dimensional structure of the genome right and that can also have impact on gene expression. So why do you want to study epigenome right what is the importance of epigenome right and one of the things that we see is that this actually this epigenetic modifications they actually give flexibility in cellular phenotypes and behavior. Now if you think or compare against genetic mutations ok so the single neck polymerisms that we discussed those are mutations that actually change the base right so that is actually is going to happen mostly in one direction right. See it is

very unlikely that you are going to revert back to the original sequence that really happens in nature.

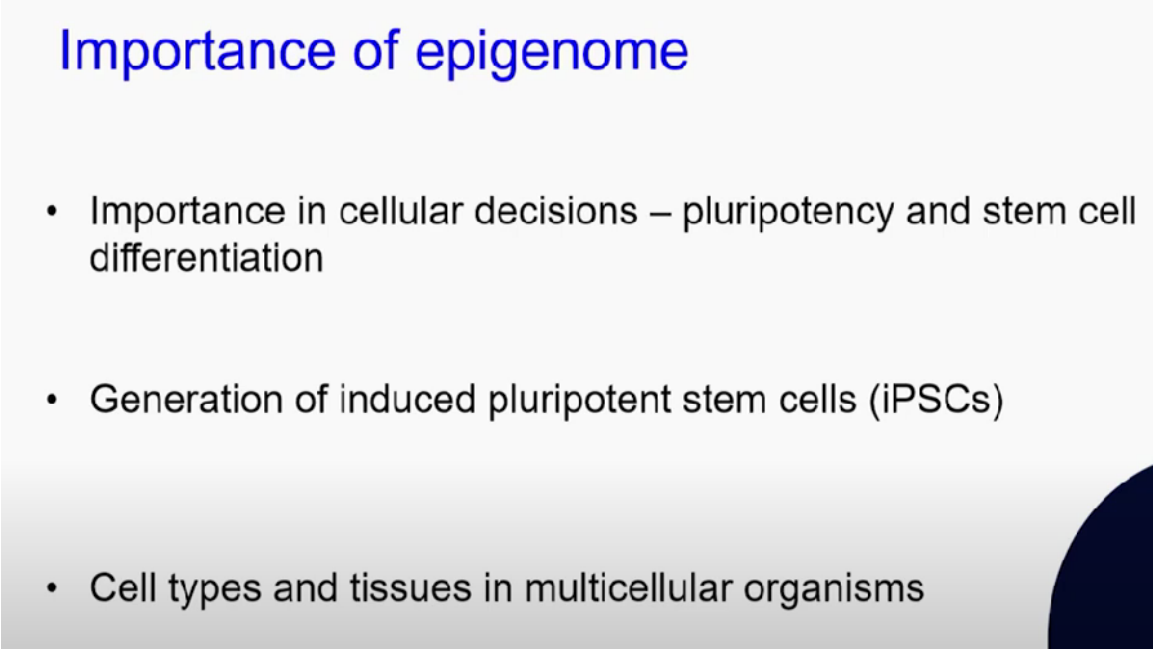
Whereas if you think about this epigenetic modifications these are reversible ok. So all these histone modifications or the DNA methylations that you see right those can be added and removed because you have all these machinery inside the cell that will do this ok. So you can add them and remove them right so it gives cells flexibility right. So if they want to express certain genes or if they request expression of certain genes they can remove certain marks and that can help in expression of that gene or if the cell does not require expression of certain gene it can also silence through this epigenetic modification right. So as you can see this gives a good flexibility in terms of cell the phenotypes and behavior so you can modify gene expression according to the need ok.

So and I mean clear enough right this what you see is that most of these epigenetic modifications they happen right they are actually changing in response to the environmental conditions or cellular signals right. So in case of humans right of course this is also affected by for example not just environmental conditions or diet right lifestyle etcetera. So everything actually plays a role in epigenetic modifications right. So this is again giving us the flexibility right so you have this flexibility in cell the phenotypes and behavior. So just to illustrate this with a figure so you can have a cell that can actually exist in two states right and this can be driven by this epigenetic modification.



So you can have state 1 state 2 this is reversible and this is driven mostly by epigenetic modification so this gives this give the cell this flexibility. Now this flexibility has very important implications throughout different processes in biological systems ok. So this flexibility is important for importance in cellular decisions right. So if you probably have heard about stem cell differentiation and pluripotency right. So here you have a important role of epigenetic modifications or epigenome then you have probably also heard about induced pluripotent stem cells and this has been seen there is epigenetic reprogramming in iPSCs and again indicating right.

So you need to have this epigenetic modifications that actually give this flexibility. And one question that you probably also would be wondering about right how do you get these different tissues right or different organs or different cell types within a body right. So if you take our examples right we have different organs right different tissues and different cell types right. So they actually contain the same genomic DNA right so the genetic sequence is mostly identical ok. So then the question is how do they actually become so different in terms of their phenotype their behavior or the job they are doing right inside the body.



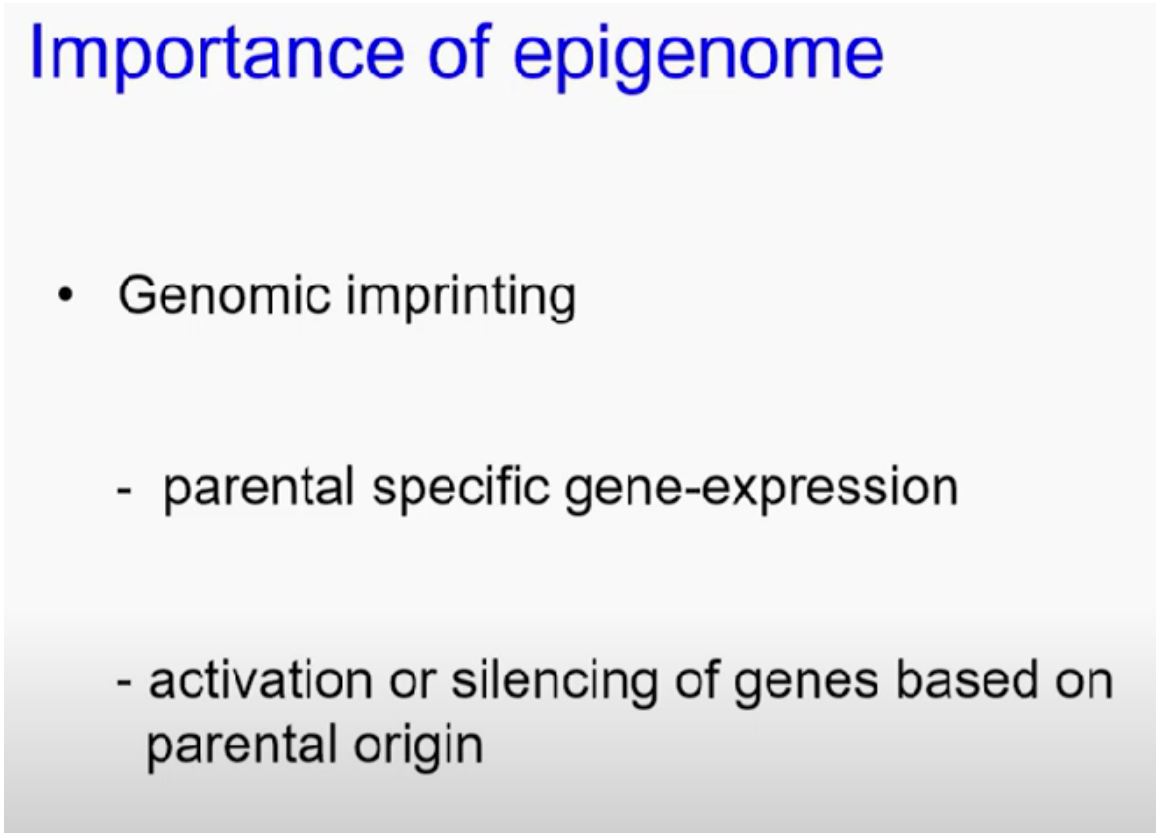
So this is again comes to the epigenetic modifications because then this can actually specify which genes will be expressed in which tissues and which genes will be expressed in which cell types or which genes will be silenced in which tissues and cell types ok. So this is again determined by the epigenome or the epigenetic modifications that we talked about. So as you probably now realize that this flexibility is also important for development of organisms right. So if you talk about development of multicellular organisms like us right so we start from a single cell and you have this development process right that goes on for quite a while right. So what researchers have seen that you have this epigenetic reprogramming during development right different marks are being added removed etcetera and that again controls expression of different genes right that actually are required for at different stages of development.

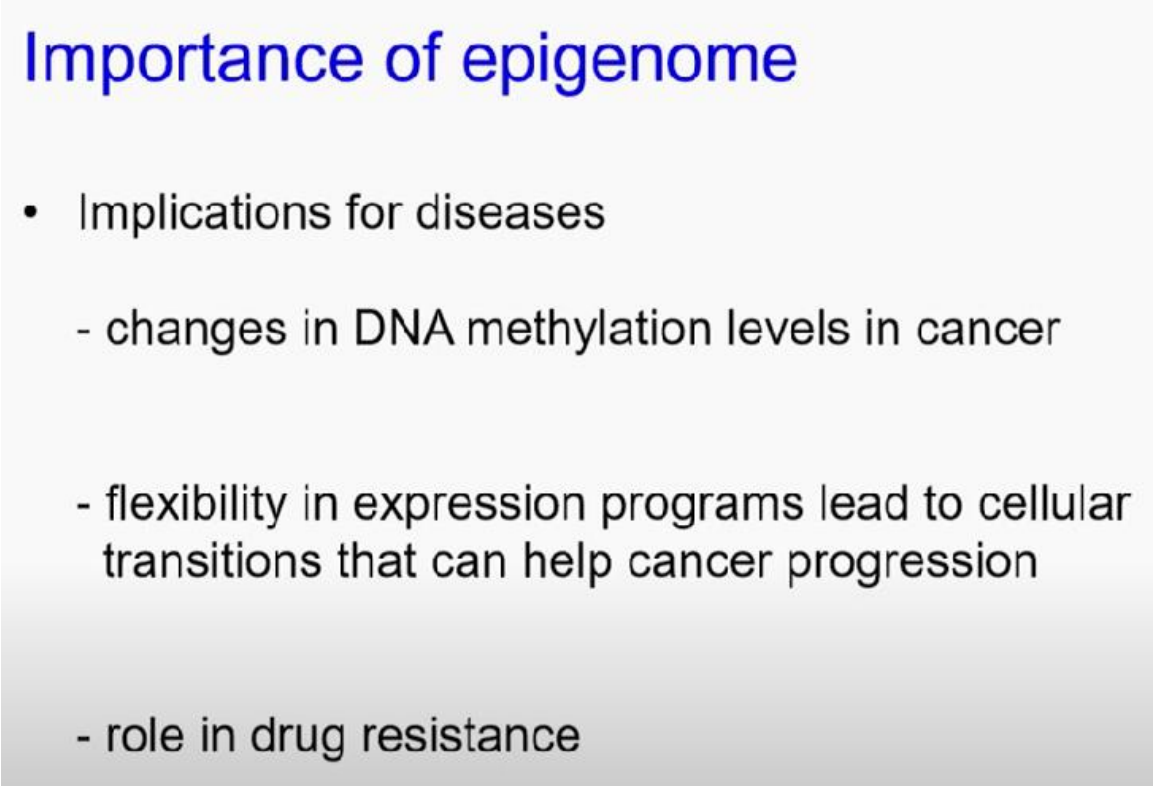
So you so again you can do this you can observe this very nicely if you study this epigenome along with the transcriptome. You can have something called genomic imprinting again this is through epigenetic modification. So genomic imprinting refers to parental specific gene expression. So this is relevant for diploid organisms like us ok. So where you have two copies one copy inherited from the mother one copy inherited from

the father.

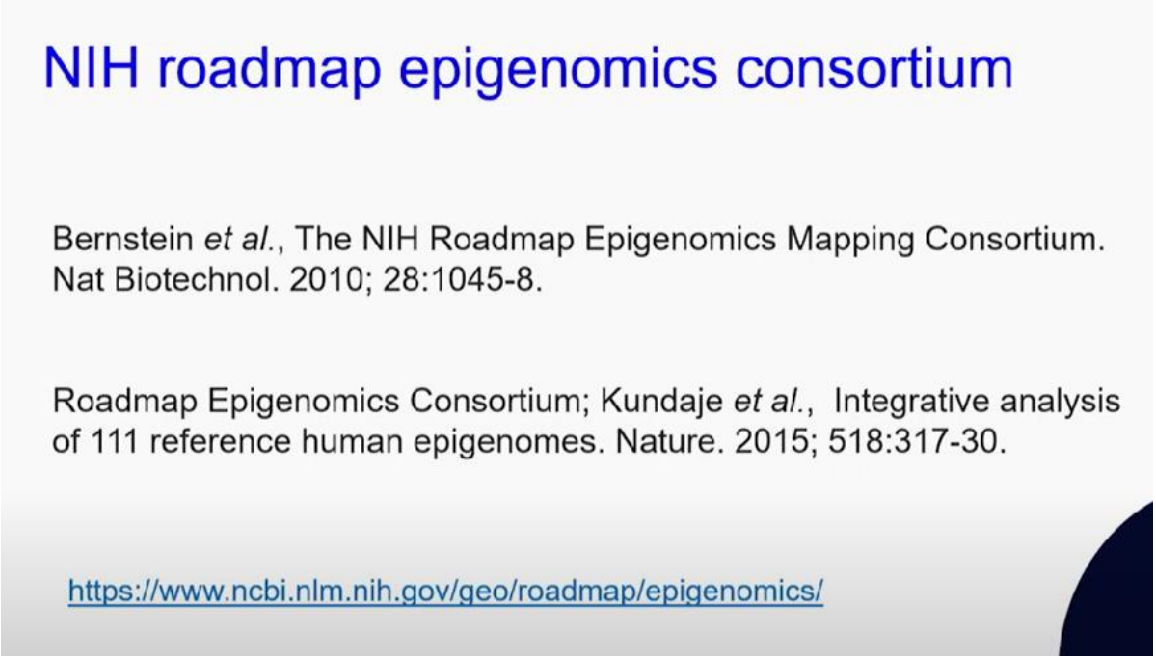
Now what happens in genomic imprinting is that only one copy probably can would be expressed ok. So out of these two copies only one copy would be expressed so this is parental specific gene expression and again this is specified by the epigenetic modifications ok. So this is activation of silencing of genes this is based on parental origin, but finally determined by epigenetic modifications or epigenetic marks that are present on these copies. Epigenome has a very important role in diseases so this has been studied now for many diseases but a particular good example would be cancer this has been well studied and what we have seen is that there is there are substantial levels in changes in DNA methylation levels in cancer and there are databases for that where you can actually get this data and observe these differences. And what researchers have also seen is that flexibility in expression programs actually lead to cellular transitions that can help in

cancer progression or also can give rise to drug resistance.

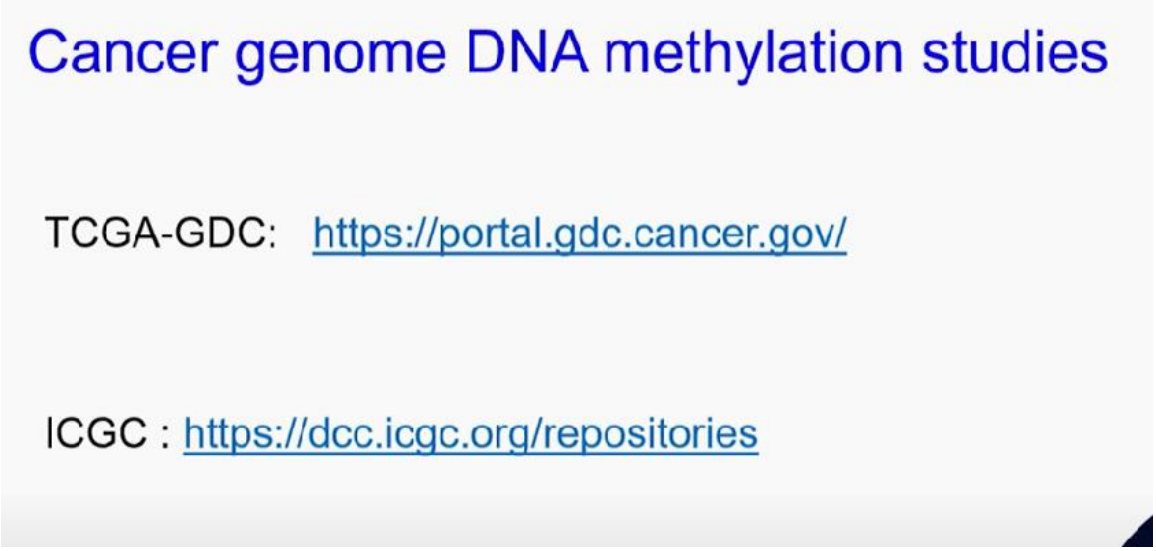




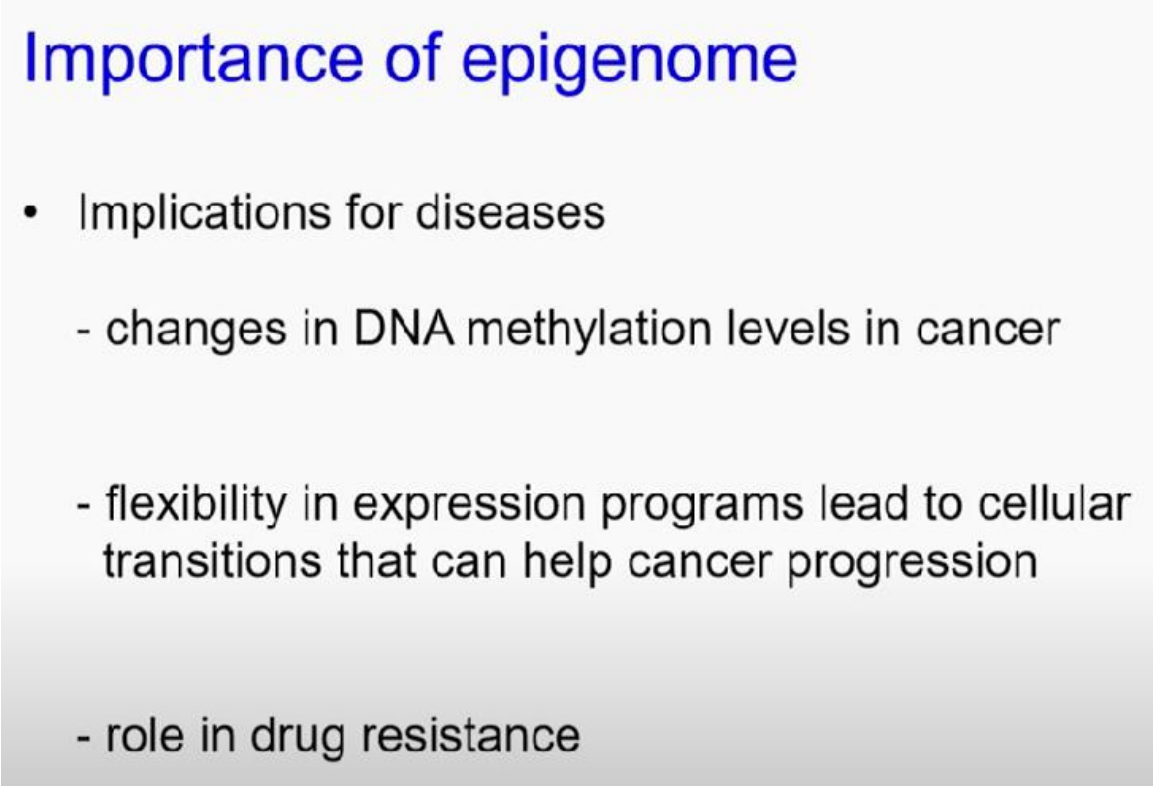
Now as I mentioned right you have the flexibility right so the cancer cell can exist in two different states one state can be expressing certain genes that actually give rise to drug resistance, but on the other hand you can have another state right and again this is reversible because of this epigenetic modifications and this state can then can be able to grow fast right. So, you can have these multiple states existing within a cell population because of these different epigenetic modifications. So, because of this importance you will find different initiatives for looking at the epigenomes and generating reference epigenomes. So, this is one such initiative and the goal is to generate reference human epigenomes ok and here are some papers and the link from that for the data sets that are coming out of this initiative ok. So, it is called NIH roadmap epigenetic consortium please look up and similarly you have another consortium which is looking at again human epigenomes.



So, international human epigenome consortium and one of the difficulties I just I mentioned right so unlike genomes right these epigenomes will be very different across different cell types and different tissues right. So, these it means there would be multiple reference epigenomes right. So, even within a single in even within one species right. Similarly there has been lot of studies on DNA methylation in cancer and here again links to two databases that actually contain data on these DNA methylation studies in different cancer types and also some normal samples and you can clearly identify these differences across the genome.



Now we come to the part to study the epigenomes right. So, how do you study epigenomes ok? The why part is I think is clear right and I hope I have convinced you that this is a very important part. So, the question is now how do you study epigenomes ok? So, earlier methods before NGS so they relied on mostly Sanger sequencing right to we will discuss again these methods when you go into subsequent classes also PCR amplification based detection. Again we will see some examples as we discuss different techniques in next class researchers also relied on restriction enzyme digestion ok. So, what you probably notice is that these are small scale methods right. So, you cannot apply all these methods to genome wide scale right and most of them are localized to certain genomic regions or segments ok.



And this all changes with the next generation sequencing methods. Now if you wanted to do the genome wide studies then of course, we had to rely on microarray based methods earlier right and this was genome wide. But then of course, we have discussed already about the limitations of microarray methods and again the sensitivity and specificity those are much less in case of microarray based methods. And this is where the next generation sequencing methods come in right. So, the current methods they rely on next generation sequencing based approaches these are genome wide and they are sensitive and specific as we will see as we go into different examples in subsequent classes. Here are the references for this class and some of these references actually talk about this impact of epigenome right for disease.

So, you can see some of these references talking about cancers right drug resistance in cancer or in or development of cancer right. So, you can see this here in these papers. To summarize so, we have discussed about epigenome. So, which is actually genome wide study of epigenetic modifications and we have talked about different types of epigenetic modifications.

So, we to include right. So, we talked about DNA methylations, histone modifications, chromatin remodeling and 3D genome configuration. We will talk about all these different types of modifications later on and how we actually identify these different types of epigenetic modifications through application of next generation sequencing technologies. We have also seen that epigenome plays key roles in development, differentiation and diseases for example, disease like cancer this has been well studied. And earlier methods for epigenetic studies right.

So, they were limited to specific segments right. So, if you talk about Sanger sequencing or amplification based methods they are limited to only small segments of the genome and not sensitive or and specific enough. If you took microarray this is again we know the limitations of microarray methods and this is where NGS technologies they provide high throughput approaches and these methods are sensitive and specific. So, in the subsequent classes we will actually see application of NGS methods for detection of these different types of epigenetic modifications at the genome wide scale. Thank you very much.