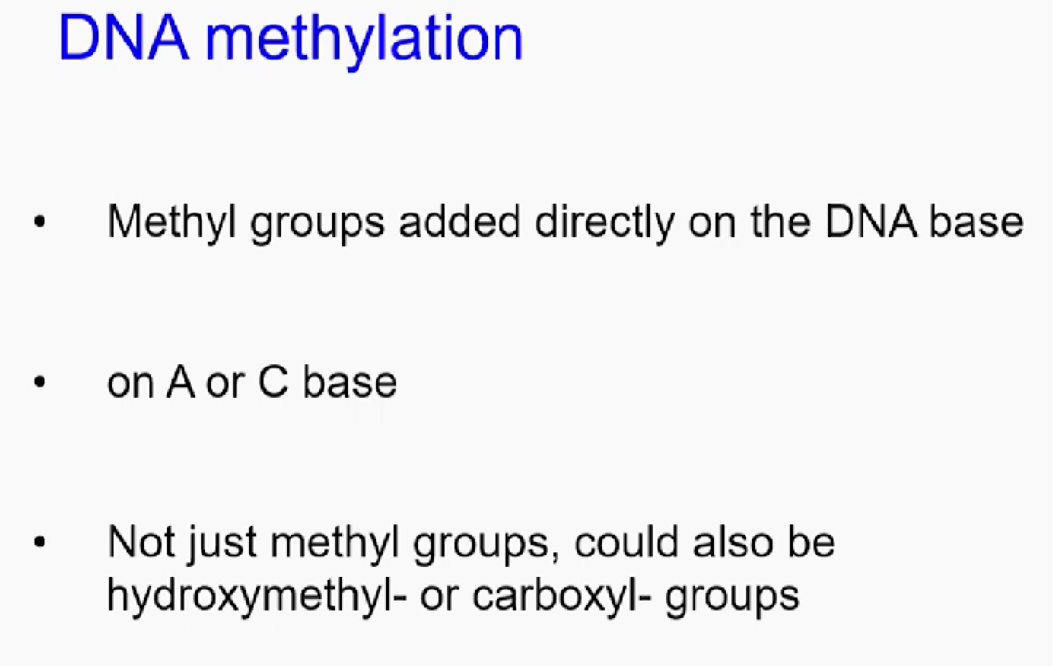
Next Generation Sequencing Technologies: Data Analysis and Applications

Detecting DNA Methylations

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 introduced different types of epigenetic modifications, namely DNA methylation, histone modifications, and chromatin remodeling. In this class and the remaining classes, we will be focusing on how we can apply next-generation sequencing technology methods to detect these epigenetic modifications. So, in this class, we will be talking about DNA methylation and how we can apply next-generation sequencing technologies to detect genome-wide DNA methylation patterns. These are the keywords for this class: methylation and bisulfite.

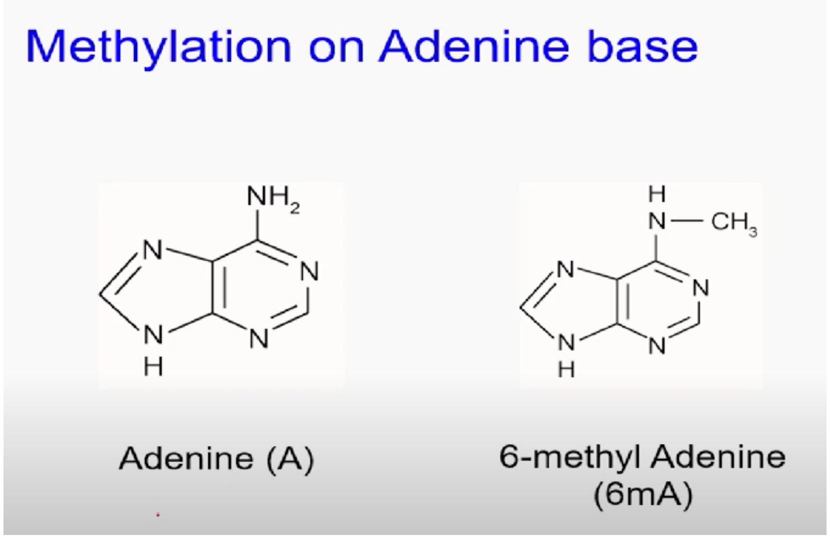
So, in the last class, we introduced DNA methylation. So, these are modifications that are directly added to the DNA base. The modification here is the methyl group, and this modification appears mostly on the A or C bases. So, another point is that these are not just methyl groups; they could also be other groups, for example, hydroxymethyl or carboxyl groups.



DNA methylation

- Methyl groups added directly on the DNA base
- on A or C base
- Not just methyl groups, could also be hydroxymethyl- or carboxyl- groups

We will see in a moment. I will show you some of these modifications, and we will see there are a variety of other groups that can also be added, although these are rare. The most common one is the methyl modification on A and C. So, let us talk about the methylation of the adenine base. So, the A base looks like this on the left. So, you see the A base, and we have the 6 methyl adenine, or 6MA in short.



Methylation on Adenine base

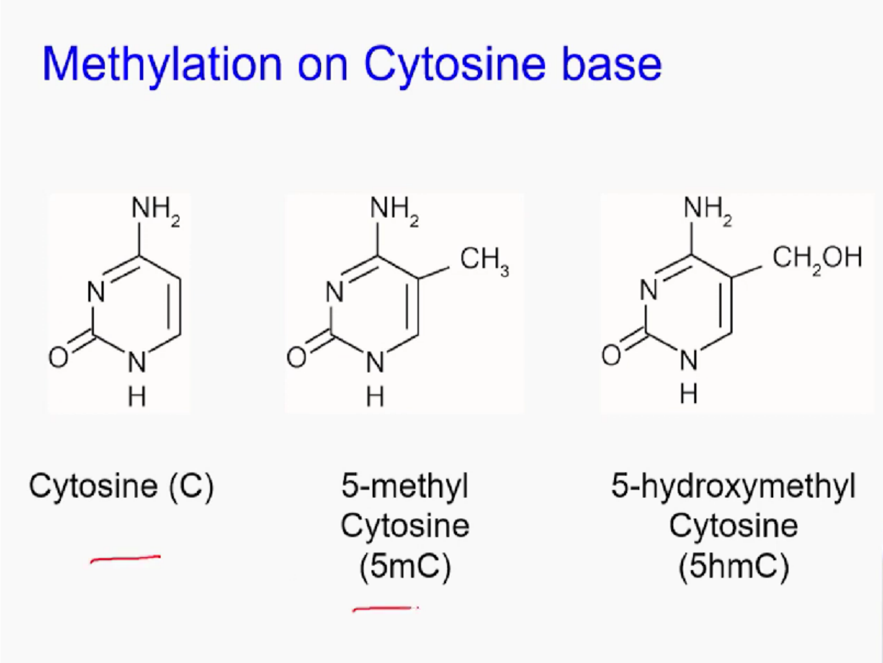
Adenine (A)

6-methyl Adenine (6mA)

So, you can see this methyl group added here on this nitrogen. Similarly, we have methylation on cytosine, and again, I have drawn these modifications for you. So, on the left, we have the cytosine, and in the middle, we have the 5-methyl cytosine. This is the most common modification of cytosine that we see. So, here is the methyl group, and you also see 5 hydroxymethyl cytosine.

So, here is the CH2OH. So, as you can see, we also see this modification quite commonly in the genomes, and again, these two types of modifications of cytosine are associated with different outcomes and affect gene expression. So, 5 hydroxymethyl cytosine will be denoted by 5 hmC. There are other modifications on the cytosine base as well, but these are rare, ok? So, compared to 5 mC or 5 hC, these two modifications are quite rare in the genome.

So, we will not be focusing on them today, but at least I should mention what these modifications modifications. So, one is formyl cytosine, or 5fC; you can see this CHO group, and then on the right, you have the 5 carboxyl cytosine, or what we refer to as 5caC, and you see the COOH group, ok? So, these two modifications are rare. So, we will not be talking about detecting these types of modifications; we will be talking about detecting the other ones, right? So, these will be three modifications now, ok.

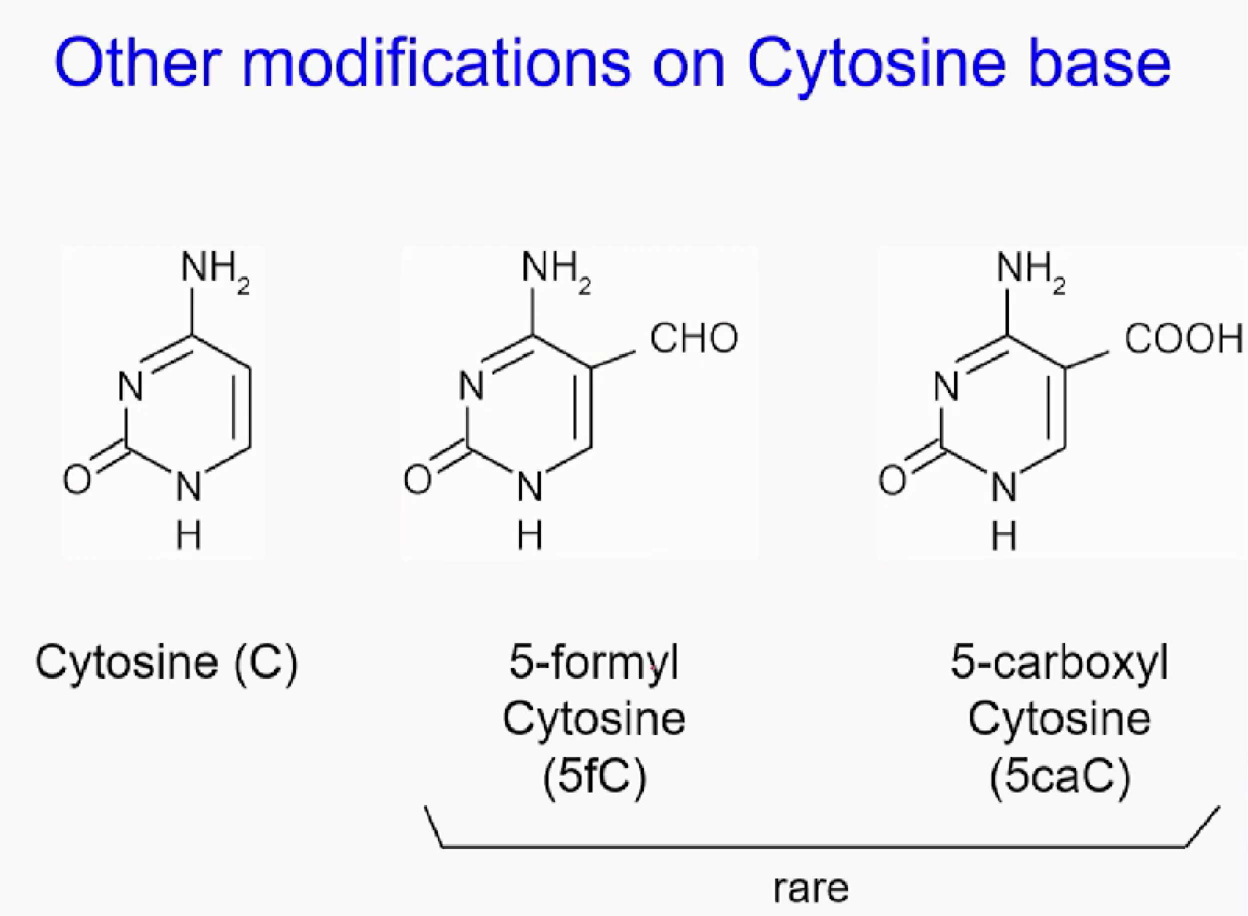


Methylation on Cytosine base

Cytosine (C)

5-methyl Cytosine (5mC)

5-hydroxymethyl Cytosine (5hmC)



Other modifications on Cytosine base

Cytosine (C)

5-formyl Cytosine (5fC)

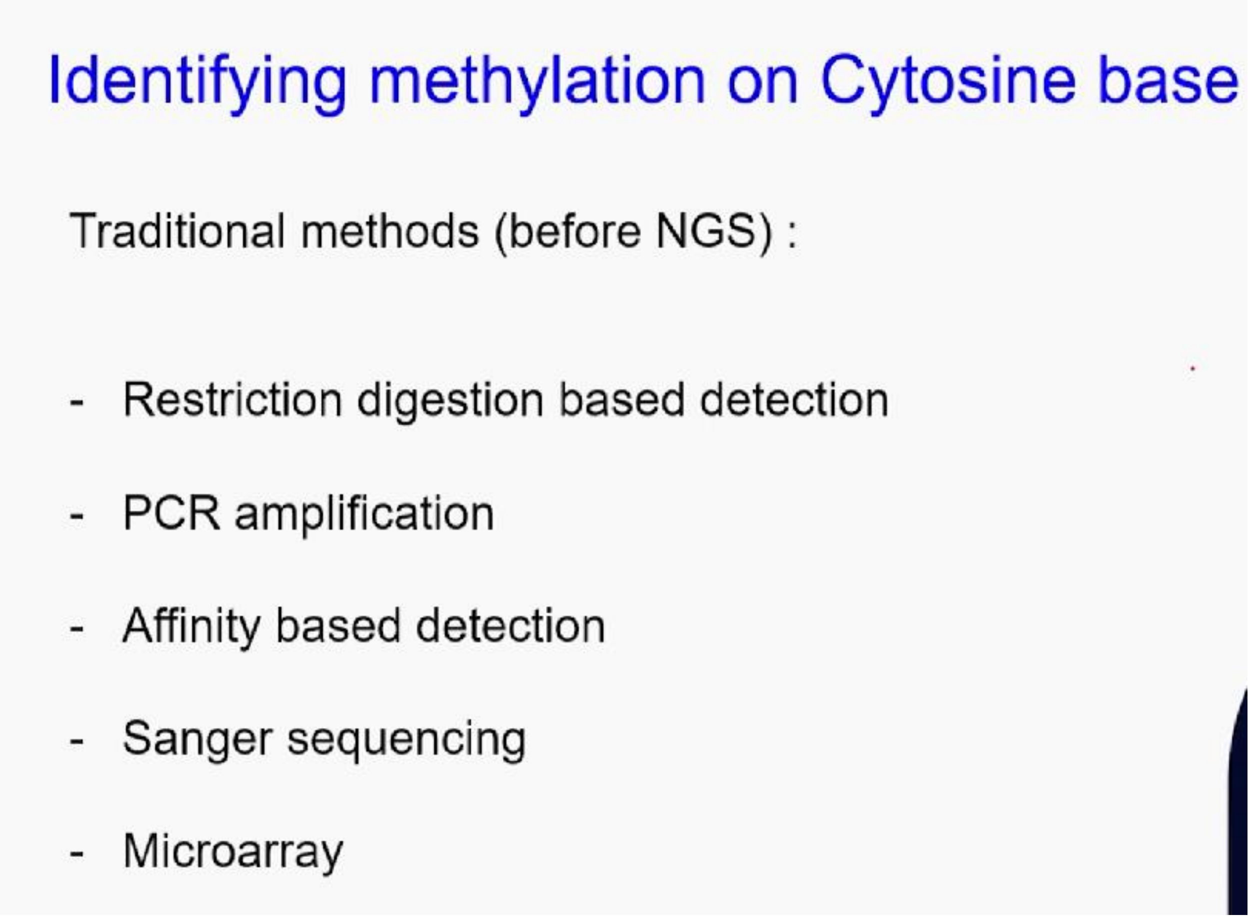
5-carboxyl Cytosine (5caC)

rare

So, we have 6 mA that is a modification on A, and we have 5 mC and 5 hmC, ok. So, let us first focus on this cytosine base and how we can identify methylation on the cytosine base, ok. So, this has been an area of interest because these modifications were quite common, ok, and they appeared in the CpG islands of the fomotoids in the mammalian genomes, ok. So, in the presence of these modifications, they actually had a silencing effect on the gene. So, that is why researchers wanted to study these modifications in the CpG islands mostly, ok.

So, what we will do is we will talk about some of the traditional methods that were employed to actually identify this methylation and this was before NGS era and this will actually give you some background, right, some of the strategies that can be applied and what you will see is that after the advent of NGS we can actually extend some of these techniques for identifying this methylation patterns on cytosine base, ok. So, these methods already existed and they simply been extended now with application of NGS and then we will talk about a method that actually applies

NGS liberally and then we can study this genome white patterns on cytosine bases, ok. So, let us start with the traditional methods, ok. So, these methods can be classified under different classes. I will be just touching upon them very briefly, and I will just mention the principle, ok? We will not have time to go into the details.



Identifying methylation on Cytosine base

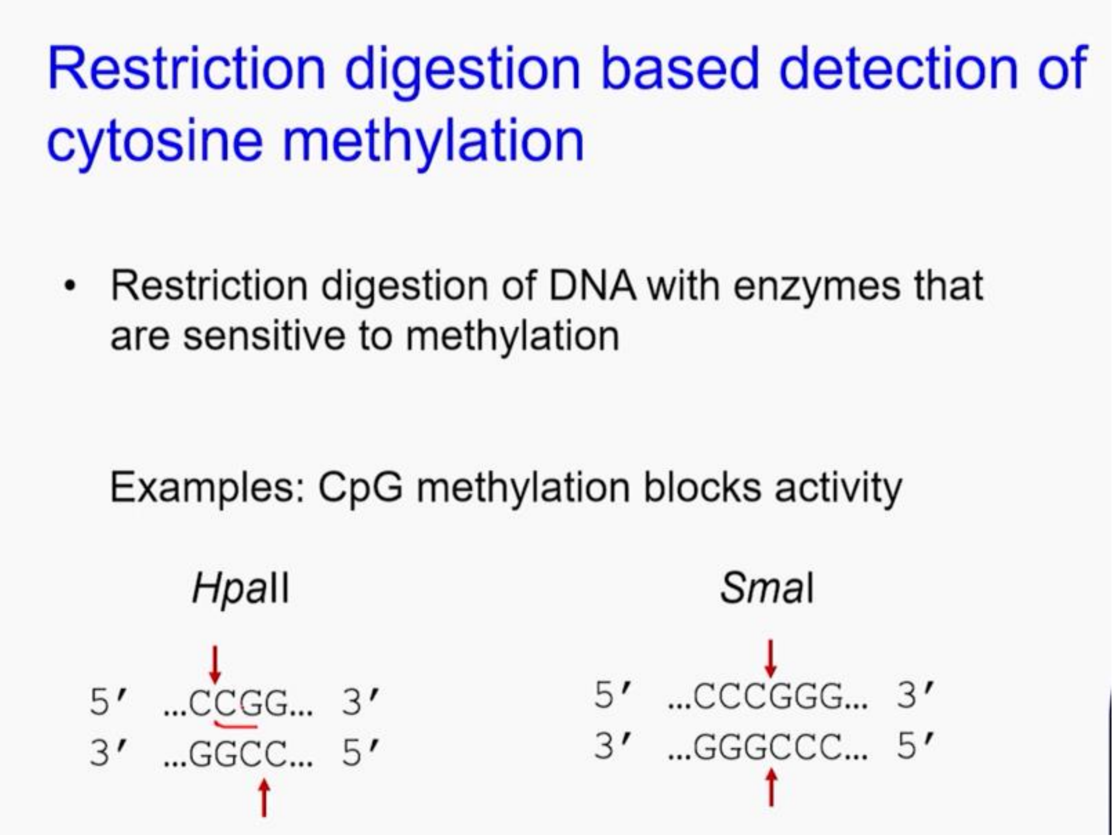
Traditional methods (before NGS) :

- Restriction digestion based detection
- PCR amplification
- Affinity based detection
- Sanger sequencing
- Microarray

If you are interested, of course, I will share all the references so you can go through them. So, the first type of method relied on something called restriction digestion-based detection. So, I am sure you are aware of restriction enzymes. So, these methods use those restriction enzymes to identify the methylation patterns.

In the next slide, I will actually discuss this in a bit more detail. Then we also had a PCR-amplification-based method for the detection of this methylation pattern. We also had affinity-based methods, and then, of course, we had Sanger sequencing, ok. So, Sanger sequencing could allow us to detect some of these regions, right, showing changes or showing this methylation on cytosine, and then there were microarray techniques that actually allowed some level of high throughput, ok. So, all the methods are DR methods.

They are mostly low throughput in nature, but with the appearance of microarray techniques, this actually allowed some high throughput analysis, and now this has been extended with next-generation sequencing technologies, ok? So, I will discuss this restriction digestion-based detection method very briefly. So, the principle is actually very simple, ok. So, we use restriction enzymes that can digest DNA, and they are also sensitive to methylation, ok. So, if you have an enzyme that can actually detect this or cuts around this CG base or CpG islands, ok, and if this enzyme is sensitive to methylation, ok. So, if methylation is present, the enzyme will not cut, ok, but if there is no methylation, the enzyme will cut, right? So, by looking at the restriction pattern, you can differentiate, right, between unmethylated and methylated bases, ok? So, that is the idea, and there are some examples of such enzymes where you see the restriction sites or the recognition sites for this restriction enzyme and then the CPG island, or here is you have this CPG methylation, right? So, you have the CG bases. So, any methylation on C would block activity of these enzymes, ok.



Restriction digestion based detection of cytosine methylation

- Restriction digestion of DNA with enzymes that are sensitive to methylation

Examples: CpG methylation blocks activity

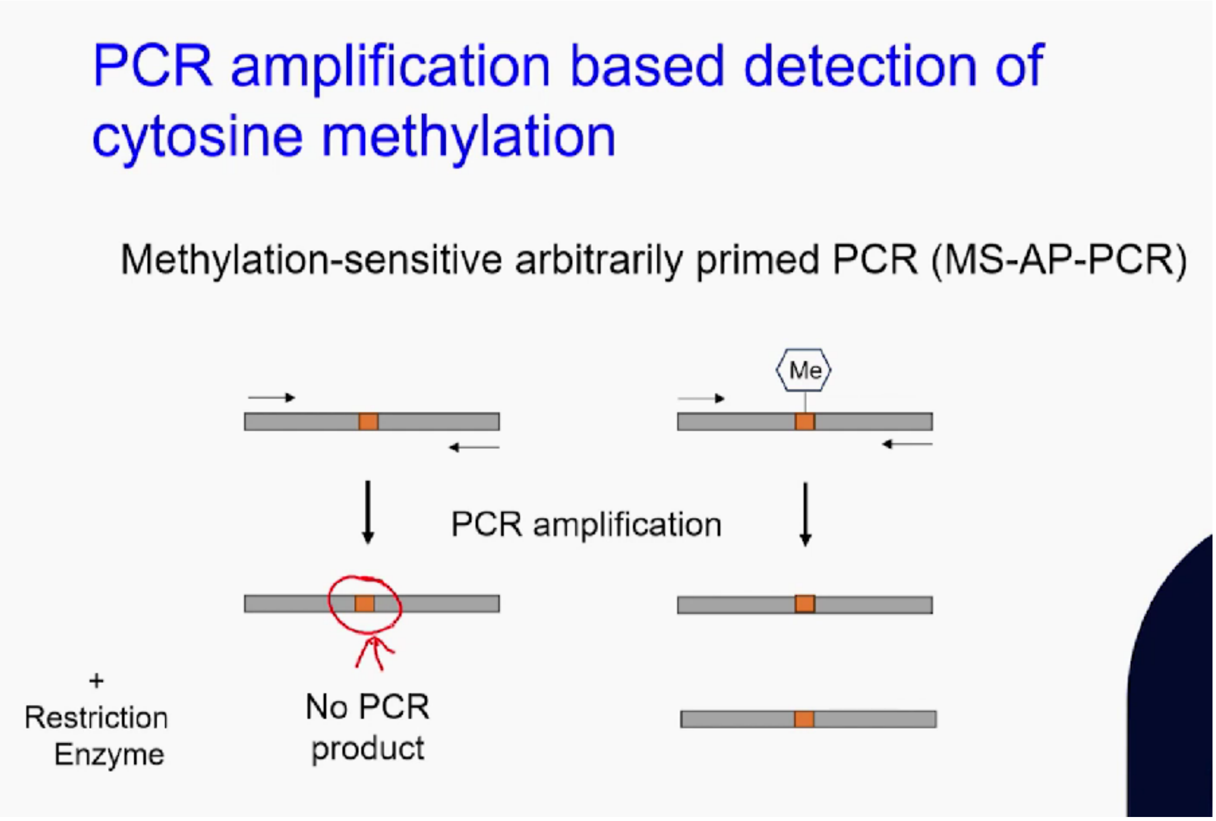
*Hpa*II

5′ ...CCGG... 3′
3′ ...GGCC... 5′

*Sma*I

5′ ...CCCGGG... 3′
3′ ...GGGCCC... 5′

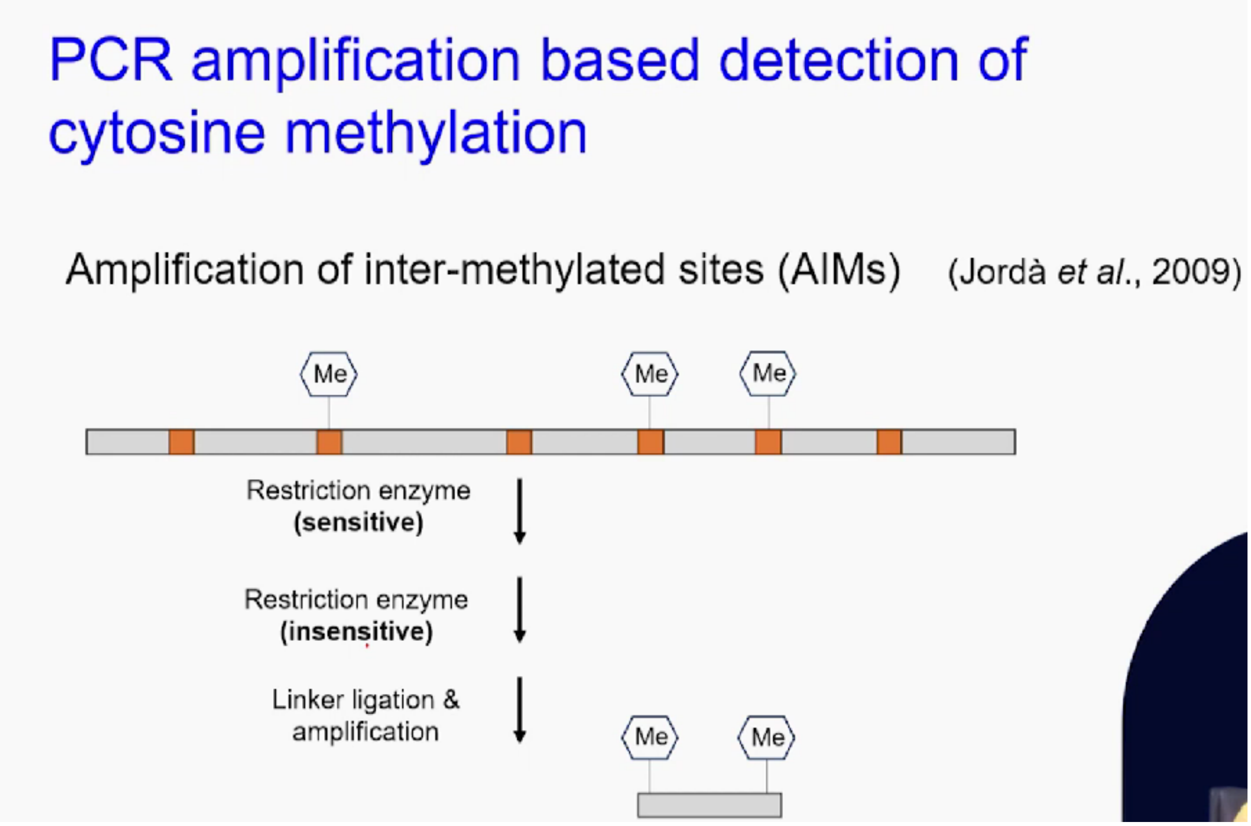
So, using these enzymes, you can then differentiate, ok, which DNA fragment has methylation and which DNA fragment does not, ok. So, there is a specific method that actually employs this technique called restriction landmark genome scanning, or RLGS. So, what happens is that these restriction sites act as landmarks in the genome, right? They are scattered around the genome and in different locations, and they act as landmarks that you can use to identify, ok, whether there is methylation or not. So, this method actually applied this restriction digestion following the same principle that I just mentioned, and then it separated these fragments using 2D gel electrophoresis, ok? So, very briefly, what will happen if you have methylation, and if you do not, the pattern on 2D gel electrophoresis will be different, right? You can then separate out fragments, and then from that you could deduce, ok, where you have the methylation and where you do not, ok. But again, you probably realize, right, because genomes are big, you cannot do this for a huge number of fragments. So, this means there would be limited throughput, right? So, you can do it only for a certain region or size, ok? So, this actually limited the throughput. Then we have the PCR amplification-based detection methods, and very briefly, we had something called methylation-sensitive arbitrarily prime PCR. I will again mention the principle very briefly, ok? So, here is an example, right? So, you have this DNA fragment, the same DNA fragment, ok? And imagine this is the site of methylation, or C, right? This is the C base, which can be methylated, ok? So, in one sample, this site is not methylated; there is no modification, and on this sample on the right, you have this methyl group. You can see this methyl group, right? So, I have added a cartoon for this, ok? Now what will happen? So, we want to amplify this fragment, right, using these two primers on the end. So, you have these primers, and then we amplify, and both of them will get amplified, right? So, you will get a PCR product on the gel. So, under normal circumstances, the methylation would not cause any problems for the PCR product. Now, what did this method do? It actually applied a restriction enzyme that will actually recognize this site here around this methylation site and it will cut if there is no methylation, ok.



PCR amplification based detection of cytosine methylation

Methylation-sensitive arbitrarily primed PCR (MS-AP-PCR)

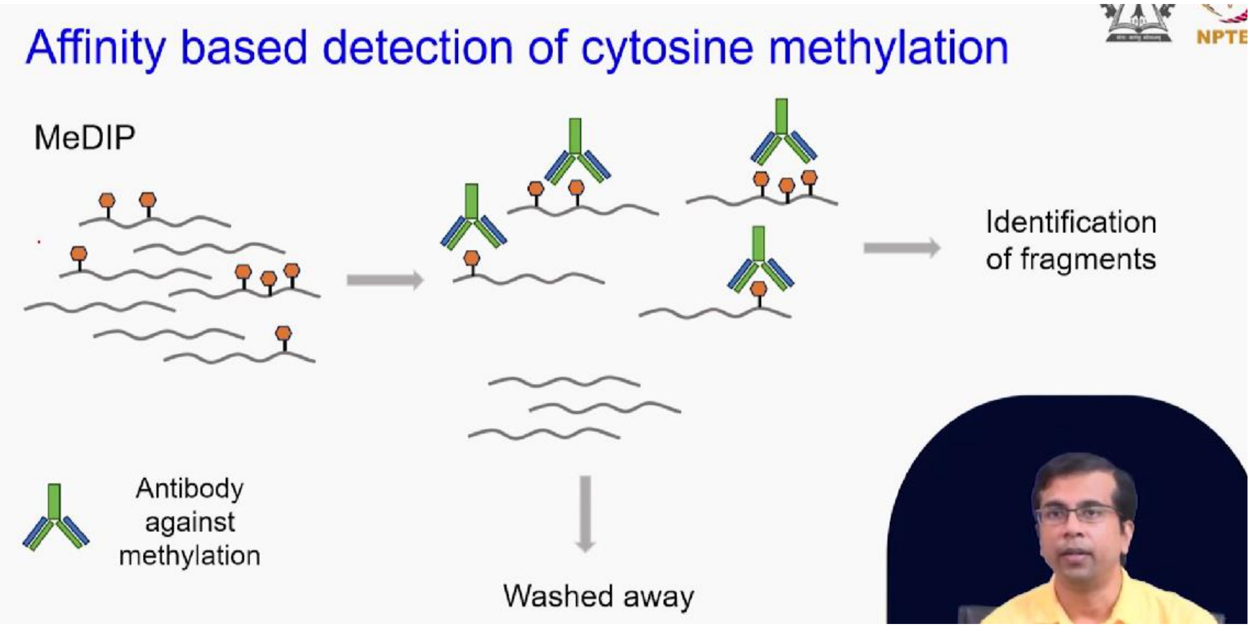
So, again, an enzyme that we just talked about, right, something like that that will recognize this site and will be sensitive to methylation. So, what will happen is that on the left side, this fragment will be digested. So, you will get two fragments from here, ok? So, this will be digested, you will get two fragments, and then once you try to amplify, you will not get any PCR product. On the other hand, on the right side, because there is methylation, the restriction enzyme will not work, the fragment will remain intact, and you will get a PCR product. So, based on this principle, you can then say, "Okay, which site is methylated? So, in any region around a methylated site, if it is still amplified after the treatment with the restriction enzyme there, that would mean there is methylation, right? So, by this method, you can then identify some of the sites where there is methylation. But again, as you can see, this will be limited in application or in throughput because, again, if you want to do this genome-wide, you need these primers, right? So, again, these primers are specific to specific regions, and you cannot have like 100,000 or millions of primers designed to do this. So, you can do this for targeted sites if you are interested in certain genomic regions or certain promoters. That is another method; I will not go into details. So, it is called amplification of intermethylated sites, or AIMs.



PCR amplification based detection of cytosine methylation

Amplification of inter-methylated sites (AIMs) (Jordà *et al.*, 2009)

So, it actually applies two sets of restriction enzyme digestions. So, one is the first; the first enzyme is sensitive to methylation, and the second one is insensitive to methylation, but they recognize the same sequence set, right? And then finally, there is some sort of ligation and amplification, and what it will do finally is sequence or fragments that reside between two methylated sites, ok. So, again, you can imagine, right, this will identify only limited regions, and even with next sequencing, you will not be able to identify sites that are seen between two unmethylated C bases, ok. Now moving on to affinity-based detection of cytosine methylation. So, the one method that we will talk about is called MeDIP, or methylated DNA immunoprecipitation, and what it does is that it can detect 5-mC and 5-hmC modifications. So, it relies on affinity binding, right? So, you have an antibody that will detect this methylation, and it will immunoprecipitate the fragments that carry this methyl group or the antibody, ok. So, here is a cartoon for the process, ok. So, here you have these DNA fragments; some of them carry this methylation, right? So, here you can see 1, 2, 3, etcetera, and some fragments do not have any methylation, ok. So, in the next step, what we do is actually apply an antibody or mix with an antibody that will recognize this methylation, ok, that recognizes this cytosine methylation, and they will specifically bind to DNA molecules that carry this methylation, ok.



Affinity based detection of cytosine methylation

MeDIP

Identification of fragments

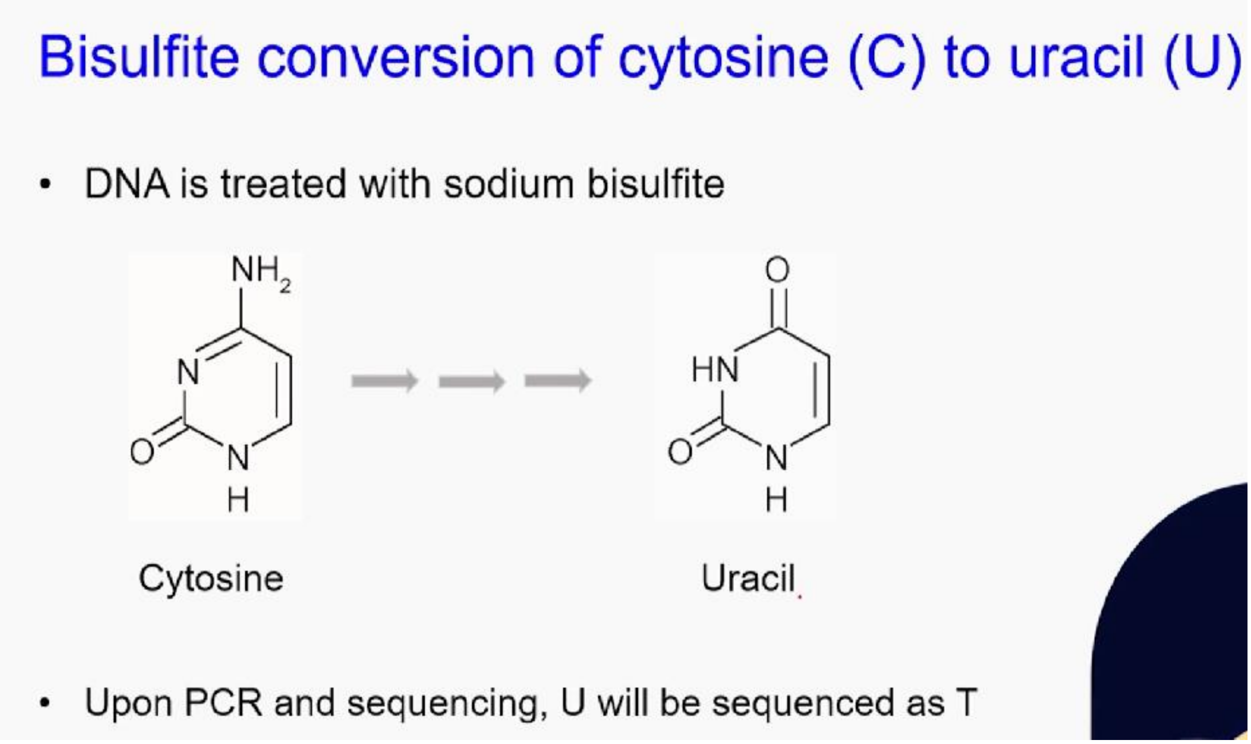
Antibody against methylation

Washed away

So, you have generated these DNA fragments; some fragments will have methylation, and some fragments will not have methylation. So, by applying this antibody, we will have this antibody binding against this affiliation. So, this antibody has affinity for this methylation, right? So, the specific antibody will then separate out DNA fragments that carry methylation, and the rest of the fragments will be washed away and will just identify the specific fragments that are bound to the antibody, ok? Now, we can identify these fragments using PCR, right? So, we can combine with PCR to identify these fragments, which is called maybe PCR, right? So, again, if you want to apply PCR methods, you need a primer. So, here you are testing against specific regions, right? So, you are kind of making a guess and designing primers to see whether you are getting PCR amplification or not, and you can do this only for a handful of regions or a handful of promoter sequences. You can also combine the MeDIP with a microarray; this increases throughput; and finally, you can apply this method with NGS, which is called MeDIP-seq. So, here the main concern is the antibody, and the drawbacks are firstly the low resolution. So, we do not know the exact base where this modification is, right? Because the antibody will bind a fragment, and in a fragment carrying methylation, you can have multiple C bases. So, you do not know which base actually was methylated, and for that reason, the antibody actually detected that fragment, right? So, that is something you cannot know using this method. So, if you have low resolution, you will know within, let us say, 50 base pairs that we have this methylation. It is also heavily dependent on antibodies, right? So, if you have ever worked with antibodies, you will see this is where the biggest challenge will be, right? Getting good quality and specific antibodies raised against methylation, ok? So, this process is heavily dependent on antibodies, and the results will be influenced by that, which means you also need to do control experiments. So, at the end of immunoprecipitation, you will see enrichment for certain regions, and this enrichment will be done with respect to the control experiment, right? So, we will have to compare against control experiments, and then you can identify this specific region.

So, as you can see, the whole design is quite complex, right? So, you need to have this antibody, you need to have this immunoprecipitation, then you have to also have control experiments, then you have to sequence both and then compare them to identify regions. Even after all this, you will not get very high resolution. There is another method very similar; it is again dependent on affinity, but instead of antibodies, it utilizes a protein domain called the MBD domain of methyl CpG-binding protein 2. This is a protein that exists in cells, and this actually recognizes this methyl CpG group, right? So, on the C base, this methyl group is bound by the MBD domain of this protein. So, you can utilize this MBD domain to actually specifically pull down, right, this 5 MC, 5 HMC bases, ok. And you can do an affinity purification very similar to what we discussed just now, and then we can take those fragments and purify and then subject them to sequencing and you can identify those. Again, the drawbacks are very similar, right? So, you have limited resolution, and again, we have to run through a control experiment.

Now moving on to the technique that actually revolutionized this field of study of methylation on C bases is called bisulfite sequencing, ok? So, this is a method that was described by Fromer et al. in 1992, which is a long time before NGS actually came into the picture. So, what happened is that this method could detect methylated cytosine and differentiate from the unmethylated ones, ok? And what happens here is that you have bisulfite treatment of DNA followed by Sanger sequencing. So, we will elaborate on this a bit more, and you will understand the principle in detail, ok? So, the first step in this process is the bisulfite treatment, ok? So, the DNA is treated with sodium bisulfite. So, what actually happens is that the cytosine, ok, that is shown in the picture here, right? So, on the left, this is cytosine.



Bisulfite conversion of cytosine (C) to uracil (U)

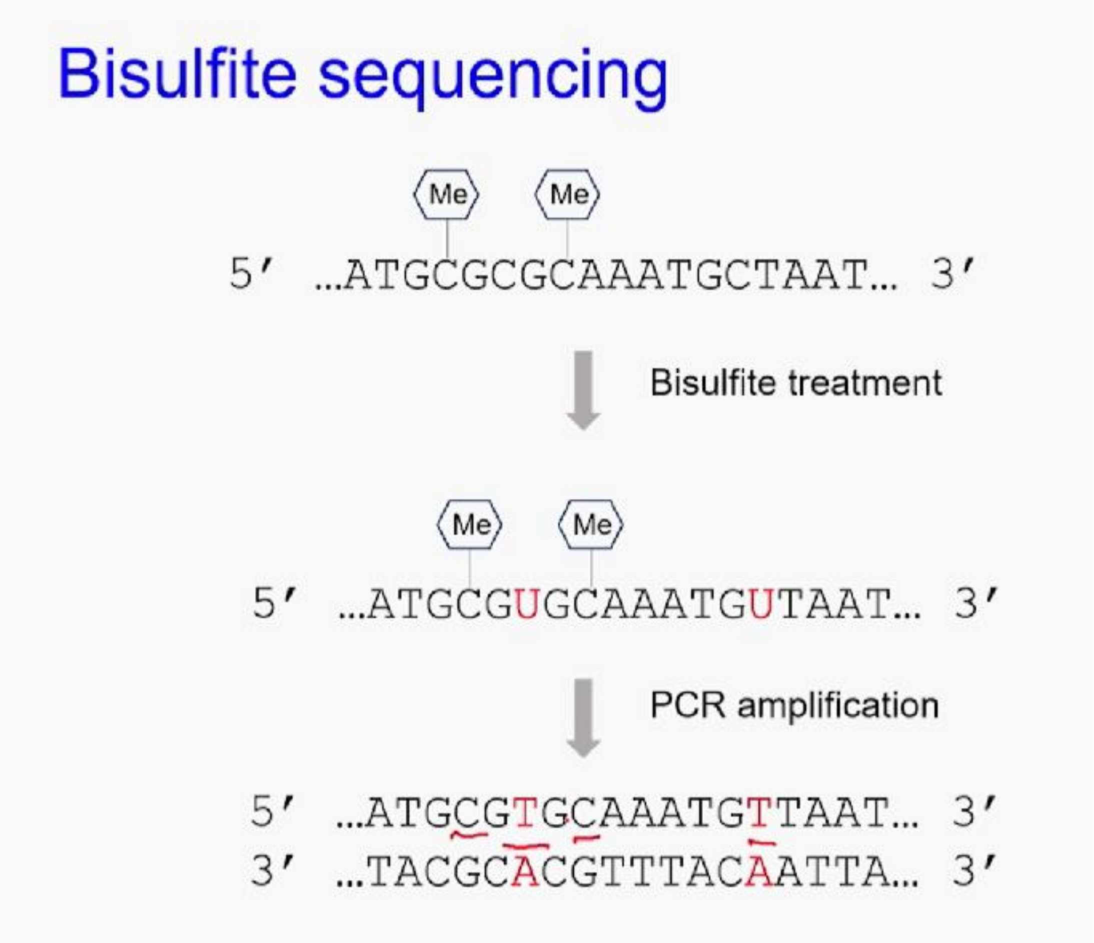
- DNA is treated with sodium bisulfite

- Upon PCR and sequencing, U will be sequenced as T

- Methylated cytosine remains the same and sequenced as C

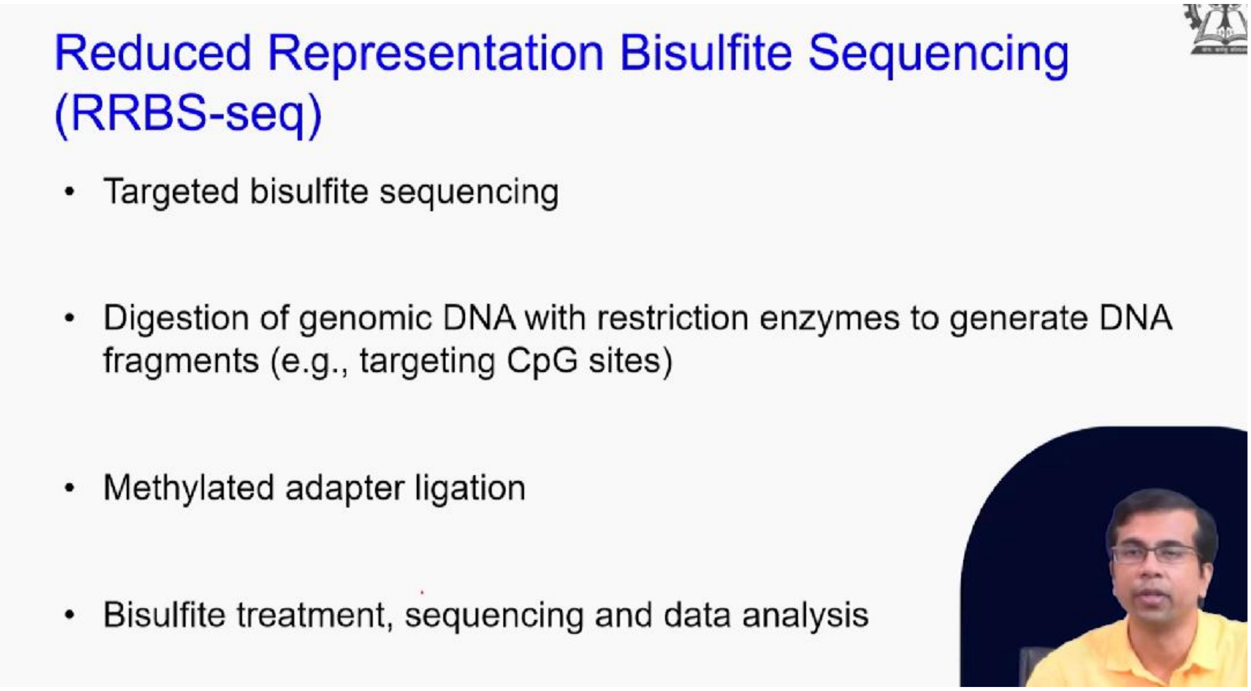
There are multiple steps. Of course, I am not going to go down those steps. So, after these steps, what you get is the uracil, ok. So, cytosine is converted to uracil by the bisulfite treatment, ok. But if you see this, right, once you have this uracil, if you do PCR on this DNA fragment and then sequence this U, this will be sequenced as T, ok. So, you will see you will observe T in the sequence data instead of the cytosine, ok.

So, keeping this in principle in mind, So, one of the peculiarities that is associated is that this will only happen if the C is un-methylated, ok. So, this conversion happens only if the C is un-methylated, ok. So, if the C is methylated or hydroxymethylated, this conversion will not happen, ok. So, methylated cytosine remains the same and is sequenced as C. So, in the sequence data, you will observe that some bases remain as C even after bisulfite treatment and some bases are converted to T, ok.


Bisulfite sequencing
Me
Me
5′ ...ATGCGCGCAAATGCTAAT... 3′
Bisulfite treatment
Me
Me
5′ ...ATGCGUGCAAATGUTAAT... 3′
PCR amplification
5′ ...ATGCGTGCAAATGTTAAT... 3′
3′ ...TACGCACGTTTACAATTA... 3′

So, based on this, you can now say which base is cytosine, right, C or normal C, and which bases are 5 mC, ok. So, you can identify those bases cytosine that actually carry this methylation. So, let us take this example of DNA sequence so that you understand the concept better. So, here you have this fragment where you have two methylated cytosines, which are highlighted here, right? So, you can see these methylations, and then you have two Cs that are unmethylated, ok? So, here are the two Cs that are unmethylated, and then you have the two Cs that are methylated. So, in the first step, when you do the bisulfite treatment, what will happen is that the unmethylated Cs will be converted to U, right? So, I have highlighted them in red. So, you can see that these Cs are converted to U, whereas the Cs that contain methylation will remain as C. Now if you then do PCR amplification, what you will get is that these Cs that converted to U will be sequenced as Cs, ok? So, again, they are highlighted in red here. So, they will be sequenced as Cs and the Cs that were methylated they will be sequenced as Cs, ok. So, based on this difference you can now identify the bases the cytosine bases that are methylated, ok. So, what we do is then we have to sequence before bisulfite treatment and sequence after bisulfite treatment and then we can simply compare, right and then identify which are the methylated cytosines and which are un-methylated, ok. So, there are other methods that are based on bisulfite treatment something called methylation sensitive single strain confirmation analysis I will not go into that or any of these methods I will just brief simply mention them you can look up if you are interested. There is something called high-resolution melting analysis. These are methods that do not use exchange sequencing; you have methylation-specific PCRs or MSPs; you have array-based methods in place. So, now we can extend this bisulfite sequencing with NGS, ok? So, remember, bisulfite sequencing was developed a long time before NGS was in the picture. So, researchers used mostly Sanger sequencing, but now with the appearance of NGS, we can now apply this method to the whole genome, ok? So, the process is very simple, right? So, now we treat the whole genome with bisulfite and then we do the whole genome bisulfite sequencing, ok. Here is the paper that actually applied this method and we can get this cytosine methylation at single base resolution, ok.

So, this was done for the human genome and we could get this data for the whole genome, ok. So, here is a very simple method, right? Again, without any complications of immunoprecipitation or any other method, here you have simple treatment, and then you go and do the sequencing, like simple genome sequencing, and you get the data, ok. So, one of the limitations or one of the drawbacks of whole genome bisulfite sequencing is that when you are looking at the cytosine methylations, these are probably limited to only very small regions across the genome, right? So, there are only certain places where this methylation can happen, right? So, especially the CpG islands, for example, in the human genome or in the mammalian genome. Now, what do we do in human genome bisulfite sequencing? We are sequencing the full genome to search for specific regions or look at specific regions. What if we could reduce the amount of sequencing and actually only sequence the regions where the bisulfite sequencing could happen, ok? And precisely this method suggested is called reduced representation bisulfite sequencing or RRBS-seq, ok? So, in this method, what happens is that this is the paper that actually suggested this method, and this is actually targeted bisulfite sequencing, ok? So, what happens is that in this method, we first digest the genomic DNA with restriction enzymes, ok?, and then generate these fragments. Now we can use certain strategies to actually make your life easier. For example, you can target the CpG sites with enzymes that are insensitive to methylation patterns. So, what will happen is that now if you are targeting in and around CpG sites, you can then take those fragments, right, and then sequence those fragments, and immediately at the end of these fragments, you will have these methylation patterns, ok.



Reduced Representation Bisulfite Sequencing (RRBS-seq)

- Targeted bisulfite sequencing
- Digestion of genomic DNA with restriction enzymes to generate DNA fragments (e.g., targeting CpG sites)
- Methylated adapter ligation
- Bisulfite treatment, sequencing and data analysis

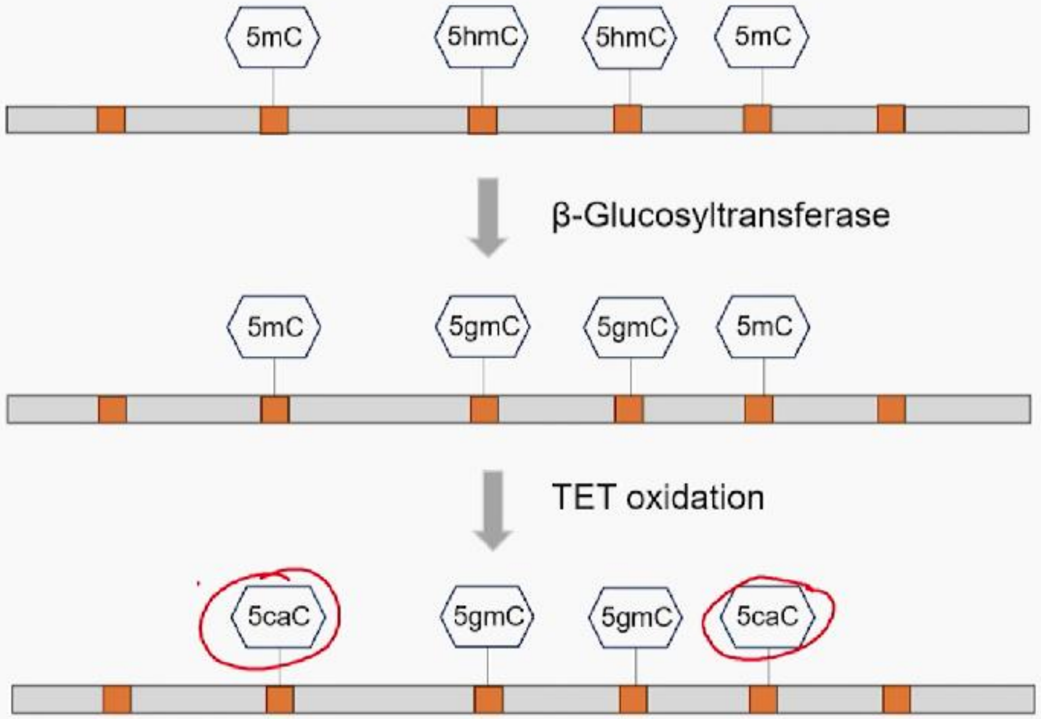
So, you can imagine right from the whole genome you create these fragments whose ends actually contain mostly CPG islands, ok. And these are the regions we are interested in looking for methylation patterns, ok. So, once you create this fragment and you sequence from the ends, you will get the sequence data. So, you do not need to actually sequence the full genome; you are not actually interested in studying or looking at this cytosine methylation.

And once you have done this fragmentation, you can then add this adapter. One thing you need to remember is that this adapter has to contain methylated Cs. So, if you have a C base in your adapter, this would have to be methylated because of the next bisulfite treatment, ok? Unless your adapter does not contain methylated C, they will be converted to T, and then, of course, this will cause problems for sequencing reactions because we will use the adapter sequence for primer binding right in Illumina. So, we do the bisulfite treatment, then we do the sequencing, and finally, we get the data from which we can analyze the methylation patterns. Now, one of the questions that we have not answered is: how can we distinguish 5 hmC from 5 mC? How can we distinguish between this 5 hmC and 5 mC? Okay, so it has been seen that there is some importance to 5 hmC. So, in many cases, 5 hmC could be a good marker for certain diseases, etcetera. So, we want to identify these 5 hmC modifications, ok? So, as I mentioned, 5 hmC and 5 mC are not converted to uracil by sodium bisulfite.

So, they will appear as C in the sequencing data. Now, there has been a method that actually was proposed in 2012 called TET-assisted bisulfite sequencing, or TAB-seq, ok. So, this method can distinguish between 5 hydroxymethyl cytosine and the 5 mC, 5 methyl cytosine, ok. So, this is how this method works, ok. So, again, imagine this cartoon where you have different sites or different Cs that are highlighted in orange and they can be unmethylated, they can be methylated, and they can be hydroxymethylated, ok. So, I have denoted them by no methylation, you can see there is no mark on them, then you have 5 MC.

So, this is methylation, and then you have 5 hmC, ok. So, these are hydroxymethylated, ok. So, what we want to do is distinguish between this hmC and mC, ok. And of course, the unmethylated Cs are also there. So, in the first step, what we do is do an enzyme treatment called beta-glucosyl transferase. So, what it does is it actually converts these 5 hmC modifications to 5 gmC modifications, ok. We will see this is some sort of protection for the next step, ok. So, in the next step there is a TET oxidation using an enzyme, ok. So, it actually converts these 5 mC groups to 5 caC. You can see this, right? So, here are these 5 CAC groups. So, they have been converted from 5 mC. So, this GM-C protection is actually important. Otherwise, these hmC groups will also be converted to caC, ok? So, gmC actually protects from TET oxidation. So, once you have this 5 caC, now you can actually apply bisulphate treatment, and they will actually be converted to uracil and they will be sequenced as T's, ok. So, what happens at the end is that these 5 hmC residues or bases are sequenced as Cs, and the rest of them are sequenced as T's, ok.

TAB-seq



5mC 5hmC 5hmC 5mC

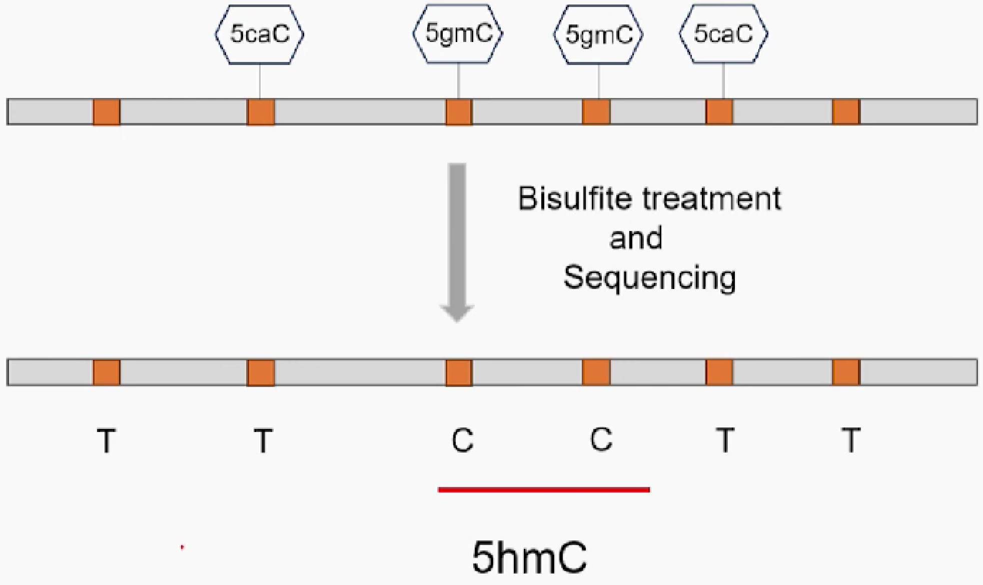
β-Glucosyltransferase

5mC 5gmC 5gmC 5mC

TET oxidation

5caC 5gmC 5gmC 5caC

TAB-seq



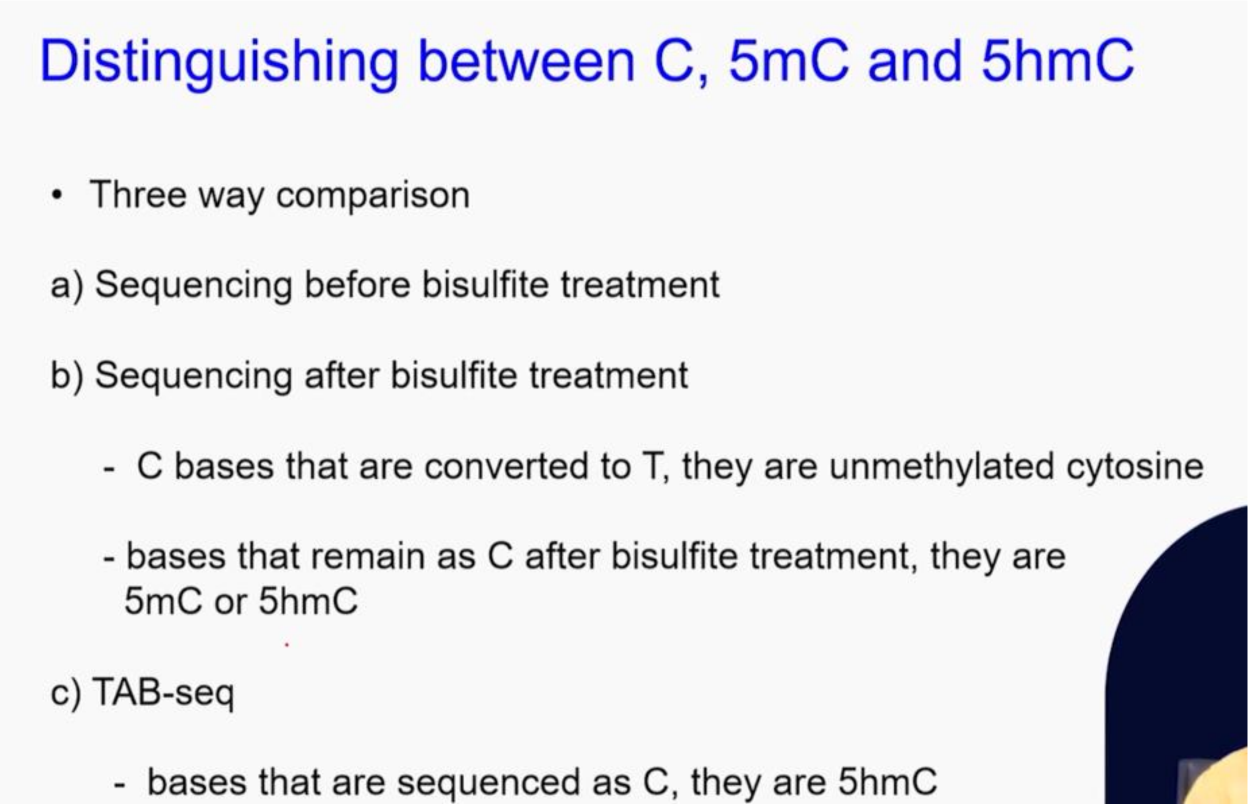
5caC 5gmC 5gmC 5caC

Bisulfite treatment
and
Sequencing

T T C C T T

5hmC

Now with this data, what you want to do now is have this three-way comparison, right? So, why do you need to do this three-way comparison? Because you want to distinguish between unmethylated cytosine at 5 mC and 5 hmC, ok. So, for that, what do we do? We do a three-way comparison, right? So, we have to sequence before bisulfate treatment.



Distinguishing between C, 5mC and 5hmC

- Three way comparison

a) Sequencing before bisulfite treatment

b) Sequencing after bisulfite treatment

- C bases that are converted to T, they are unmethylated cytosine
- bases that remain as C after bisulfite treatment, they are 5mC or 5hmC

c) TAB-seq

- bases that are sequenced as C, they are 5hmC

So, we get the original sequence, and then we get sequencing after bisulfate treatment. So, C bases that are converted to T are unmethylated cytosine. So, we identify them. Bases that remain as C after bisulphate treatment are 5 hmC or 5 hmC, ok? So, we cannot distinguish just after bisulphate treatment between 5 mC and 5 hmC. And now we do the tap set and bases that are sequenced as C and which remained as C after bisulphate treatment; they are 5 hmC, right? So, now, using these three sequencing methods, right? So, you can actually determine this unmethylated C at 5 mC and 5 hmC, ok? So, there is another method again looking at slightly different combinations called oxidative bisulfate sequencing, or oxBS-Seq. It can also distinguish between 5 mC and 5 hmC, ok. So, this actually relies on selective oxidation of 5 hmC to 5 fC followed by bisulphate treatment, ok.

So, C and 5 hmC, these are now sequenced as T, ok. And whereas, the 5 mC, those bases are sequenced as C, ok. Again looking at these comparisons, you can derive which are unmethylated, which bases are 5 mC and which bases are 5 hmC, ok. So, again, this we compare with conventional bisulphate sequencing as I just mentioned, ok. So, here are the references for this class.

And to summarize, we talked about traditional methods for detecting cytosine methylation. We talked about PCR-based detection methods, affinity-based detection methods, and restriction digestion-based detection methods, but they were all of either low resolution or had low throughput or limited throughput. So, we could not apply that across the whole genome. Microarray methods actually allowed some sort of high-throughput measurement, but then again, resolution was concerned for many of these methods. So, bisulfate sequencing actually allows genome-wide detection of cytosine methylation at single-base resolution.

And we also talked about TAB-seq and oxBS-seq. This allows us to distinguish between 5 mC and 5 hmC bases, as we have just seen. And finally, we need to do a 3-way comparison for resolving C, 5 mC, and 5 hmC bases. Thank you very much.