An Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 06
Introduction to Genomics - IV Epigenome

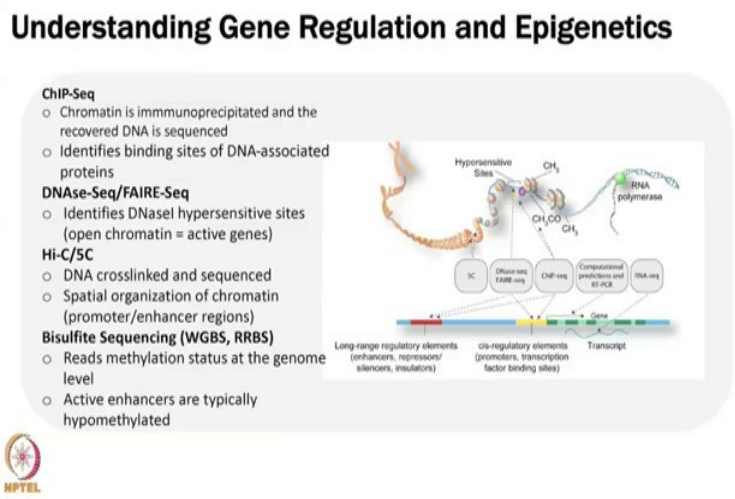
Welcome to MOOC course on Introduction to Proteogenomics. In the last few lectures, Dr. Kelly Ruggles have given you a very detailed insight of genomic revolution, studying genome and transcriptome. Today she is going to talk about epigenome. The epigenomics deals with the modifications in expression and function of genes due to the addition of different functional groups. Today will be Dr. Kelly Ruggles fourth lecture, and she will talk about epigenomics and use of ChIP-Seq technology to perform epigenomics studies, various type of publicly available databases like CPTAC, The Cancer Genome Atlas, TCGA, and ENCODE will be described.

She will also talk about the DNA seq where one can sequence gene by using DNAs, and then compared them with the reference genome. The lecture will also illuminate about DNA methylation and whole genome bisulfite sequencing known as WGBS, which can facilitate in finding sites of methylation. However, because of high cost and inefficiency of reduced representation bisulfite sequencing is preferred over WGBS.

Dr. Kelly will further talk about the role of epigenetics in histone modifications leading to expression of different genes. She will also cover the Hi-C which helps in understanding the interaction and folding of chromosomes with neighboring chromosomes or within itself. So, let us welcome Dr. Kelly Ruggles for her lecture today.

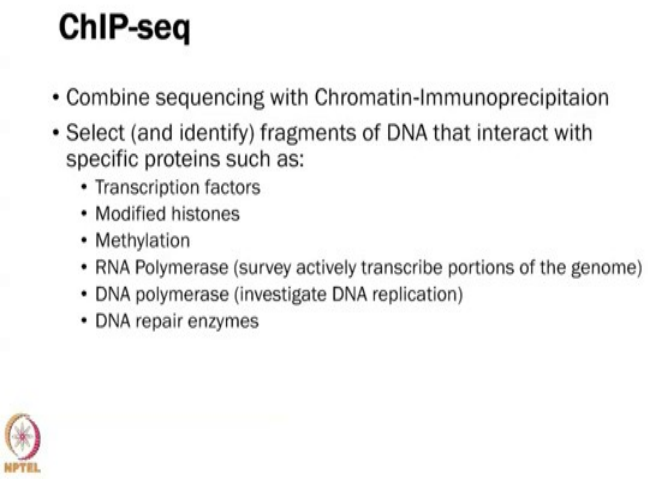
Now, I will talk about is epigenomics. So, there are lots and lots and lots of methods for epigenomics. I feel like they are constantly.. I have collaborators I come with come to me with more and more of these methods that are slightly different, and have new names. But I am, so I am just going to talk about some basic ones, but you will read if you go into depth with this, you read about more and more of these. There seem to be hundreds, but they are all slightly different, they are all doing similar things.

(Refer Slide Time: 02:35)



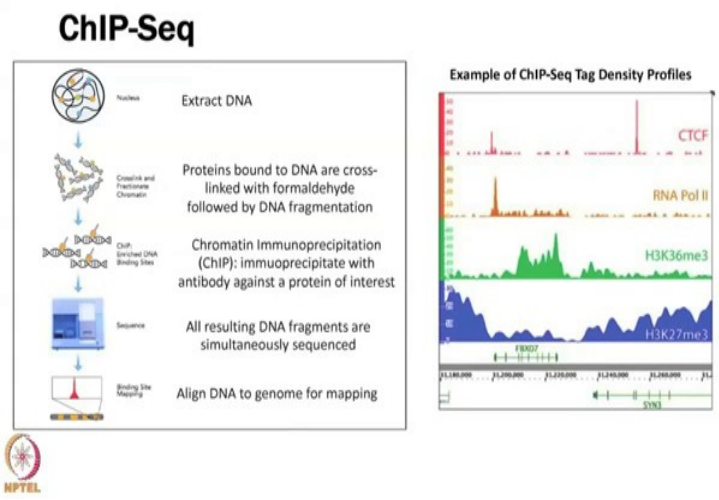
So, ChIP-Seq which is we are going to talk about in a lot of detail, but this is really trying to identify DNA associated proteins. DNA-Seq which identifies active genes based on DNAase hypersensitivity. Hi-C, which is cross linking DNA is just understand how because the DNA sort of folded on each other on itself in the nucleus. So, understanding how certain parts of the DNA interact with each other. And then bisulfite sequencing, which is looking at methylation status. So, now talk about each of those in a little bit more detail.

(Refer Slide Time: 03:16)



So, ChIP-Seq is combining sequencing with a chromatin-immunoprecipitation. So, what you do is you have to pick proteins of interest. So, let us say you are interested in a certain transcription factor or you are interested in a modified histone, there was all sorts of things that any protein that you are interested in that interacts with the DNA right, you can pull down that protein, and then look at what sequence is associated with it. It is essentially the theory behind this.

(Refer Slide Time: 03:44)

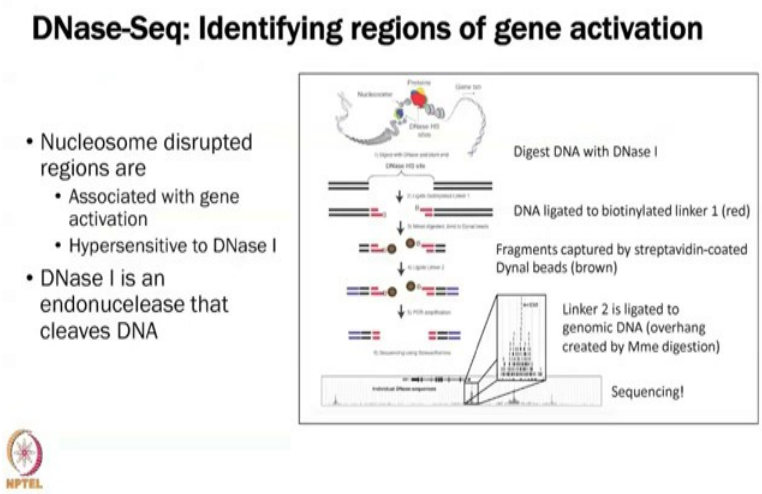


So, you extract DNA from your nucleus. And you when you extract the DNA, there are proteins that are bound to the DNA. And then you cross link them with formaldehyde. So, you know that there they continue to stay bounds to the DNA, and then you do this chrome this chip immunoprecipitation meaning you have an antibody against whatever protein you care about. You pull down all of the protein more of now all of it, but some of the proteins within your sample that is attached to DNA sequence, and then you take the DNA fragments that were attached to that protein, and then you sequence them and you align them to the DNAs.

So, you know whatever protein you are interested in these were the sequences that it is interacting with. And this is a typical. So, this would be a trans the CTCF would be the protein, and then you would pull down and you can see that there is a lot of reads in this area here. And you could look at RNA polymerase; you could look at a methylation status. So, you can look at a whole bunch of different anything you can kind of pull out

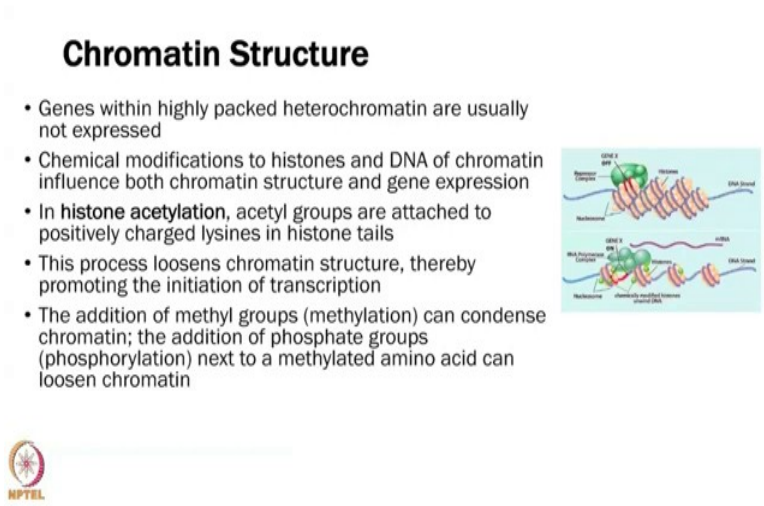
your sequence and after you do this hybridization, you can look at with this kind of a method. So, any questions so far?

(Refer Slide Time: 04:52)



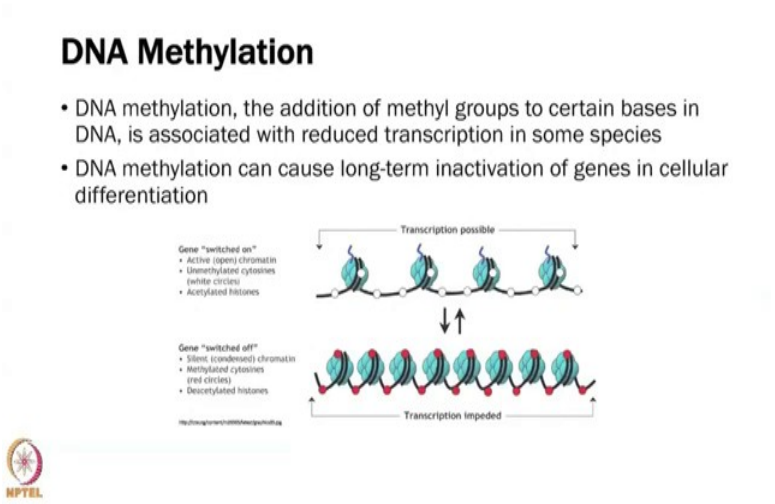
DNase-Seq is a way of looking at gene activation. So, there are it is been shown that that regions of the genome that are hypersensitive to DNase are active. So, what you can do is, you can cleave DNA with DNase. And then you can look at the you can kind of look at where the sites of its cleaved. And then you can take the chunks that are closest to that cleavage, and actually just sequence those. So, you can see where what regions of the genome have are sensitive to DNase, because they were able to be cleaved and then pulled down using these Dyna-l beads.

(Refer Slide Time: 05:43)



So, in addition to that, you again you can look at the structure of the chromatin. So, genes are packed with in this heterochromatin, and the genes that are packed within it are not expressed. So, there is lots of different modifications that occur on chromatin to sort of open it up and make it available for expression. And there was lots of ways of measuring openness and activity based on understanding how this chromatin structure occurs. So, for example, histone acetylation, the loose, if it is acetylated, it actually loosens up the structure and allows for some transcription. So, you can measure levels of this. Or you can measure the addition of methylation groups. And we will talk a little bit about how you do that and what that means.

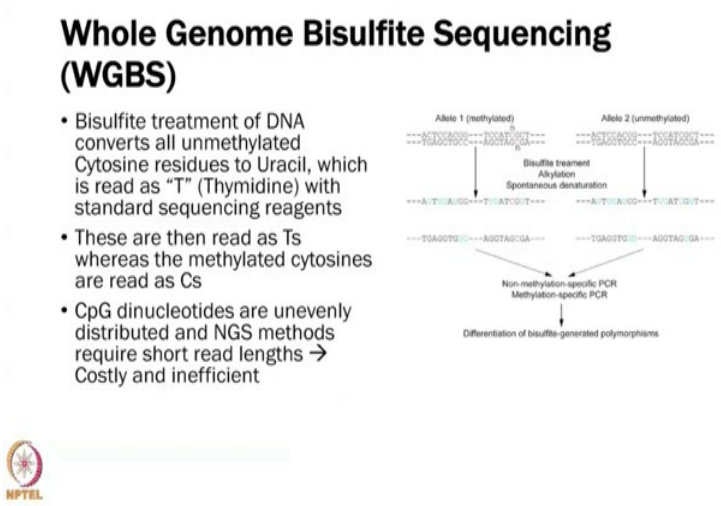
(Refer Slide Time: 06:30)



So, DNA methylation I think it is the most one of the more commonly used epigenomics methods just because it is like one of the easier ones when you are doing a whole genome analysis there. I am not gonna show a slide on this, but they are also methylation arrays that are similar to the arrays we talked about, where you have different methylation areas of where you can measure specific patterns of methylation on the genomes.

So, you have just a chip, and you put your DNA over it, and then you can measure levels of methylation in different areas. But it is been shown that adding methylation to the DNA actually reduces transcription. So, if you have a methylation and the transcription is lower, and I am that is a generalization there is lots of interest like this is a complex system, and we will talk a little bit about that, but that is sort of the idea. So, by measuring methylation, you are you can measure how active certain genes are.

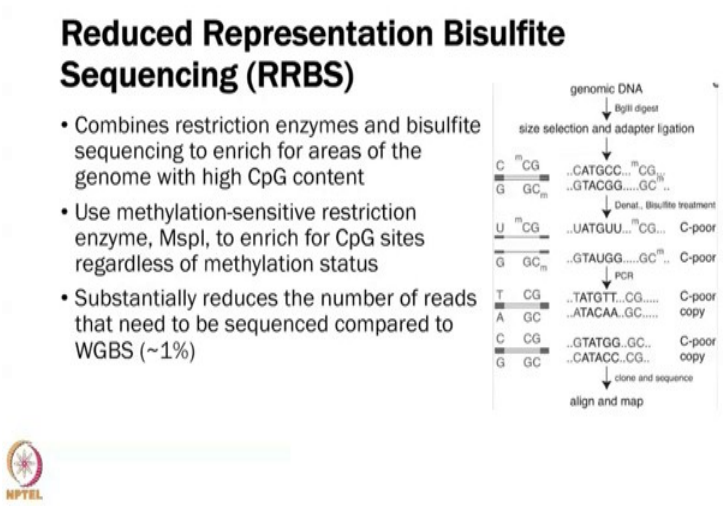
(Refer Slide Time: 07:28)



And the ways besides this methylate, the methylation chip, there is also this whole genome bisulfite sequencing or WGBS, where you bisulfite treat your DNA and it converts all of the unmethylated cytosines to uracil, which are then read is this as T. And when you sequence it, and then you can kind of deal with this in the informatics side.

So, you know that every time there is an unmethylated C, it is turn to T. And then you have a control, where you do not do this bisulfite sequencing. And so then you can look and see but you just do a whole genome sequencing of your bisulfate and your non-bisulfate treated sample. And then you can look and see which ones were methylated and which ones were not. This is just super expensive not a lot of people do it; because it is essentially whole genome sequencing times two because you have to have a bisulfite and a non-bisulfite sequence sample. So, it is one way of doing it, but again it is really expensive.

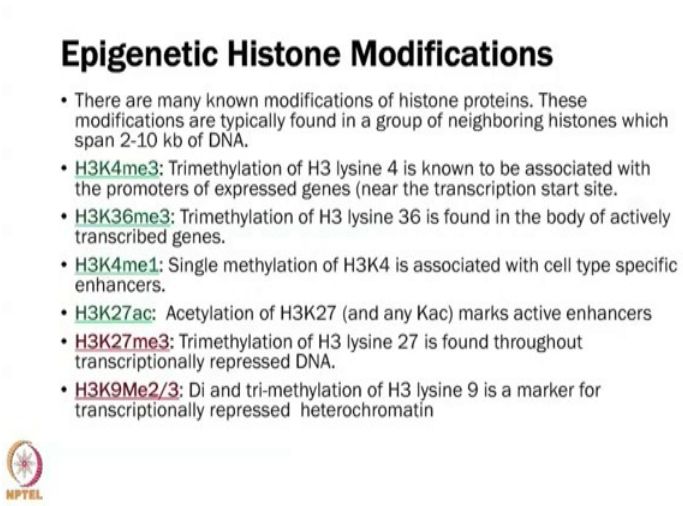
(Refer Slide Time: 08:33)



So, they have come up with a different approach which is called Reduced Representation Bisulfite Sequencing or RRBS, where it combines that method with restriction enzymes. So, you can enrich for specific sites that have more of these methylations occurring with CpG sites. I am using a specific restriction enzyme. So, it is a methylation sensitive restriction enzyme, so meaning that it will cut if there is methylation and not cut what if there is not.

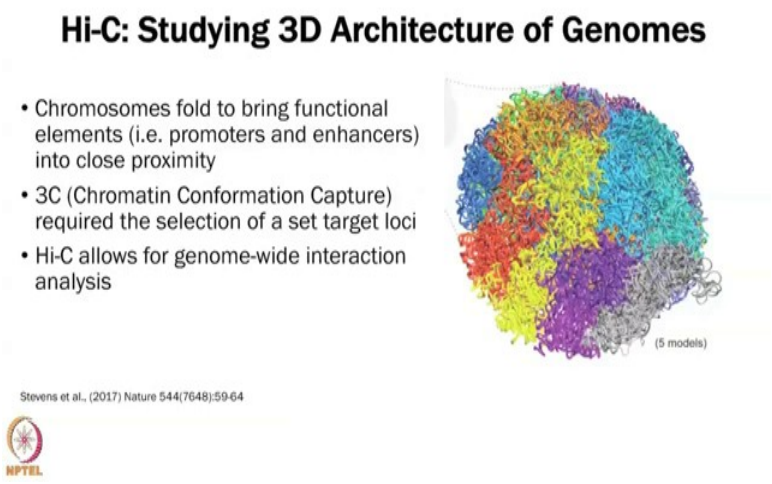
So, it allows you to just get rid off a whole bunch of stuff you do not care about. So, it takes your whole genome sequence, and it makes it only 1 percent. So, it takes things that you do care about. So, this is a more common method. I work have worked with this data and it is a bit of a mess to be honest, but it is science and that is what we do. So, it is a method that is available as well.

(Refer Slide Time: 09:29)



And I will include this here, I am not going to go through all of it, but there are these are different modifications, and sort of how they have been shown to be involved in changing expression of genes. So, there is a whole bunch of different modifications, methylation modifications and histone modifications that can occur.

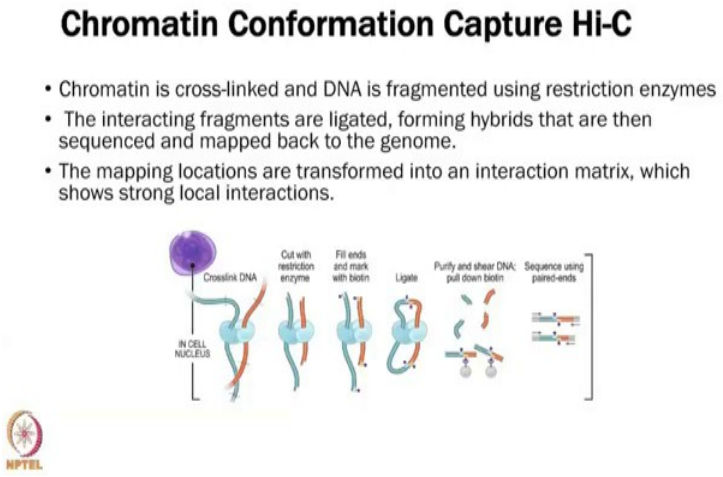
(Refer Slide Time: 09:52)



So, the last thing I am going to talk about I think for the epigenetics is this Hi-C. So, this is studying how the 3D structure of how the actual chromosomes are folded on themselves, and with other chromosomes. So, here is this is a figure I took from this

paper that just shows models of these are each colors are different chromosome, and how they are how they are the architecture of the chromosomes within the nucleus. You can see they are folded on themselves and folded with other chromosomes, and how they interact with each other is something that is become of great interest to a lot of people. So, how we measure this is I am using this chromatin conformation capture or Hi-C, which allows you to look at how these chromosomes are folded on each other.

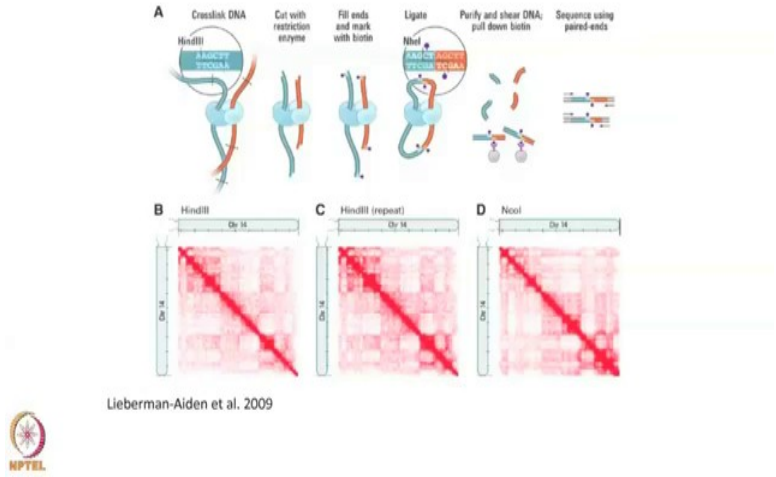
(Refer Slide Time: 10:52)



So, what happens is you cross link your DNA, and then fragment it using restriction enzymes to you. So, you have your DNA folded on itself and then you again cross link it. So, it is it stuck together, and then you cut it up. So, then you have a stuck together small pieces. And then you so here that is here, and then you cut with the restriction enzyme. Then you sort of cap the ends and you ligate them together. So, you take these two pieces that were interacting and you just ligate them. So, they are circular. And then you remove these you end up sharing it, making it into pieces and then sequence them.

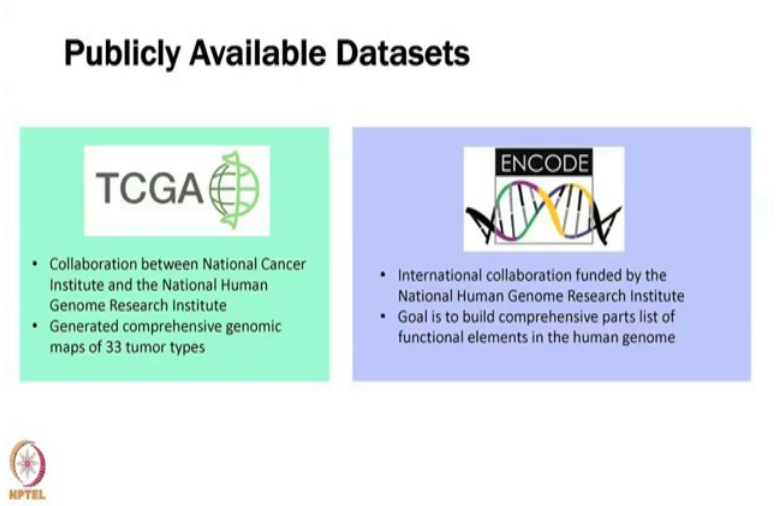
So, you end up getting an area here right this piece that you ligated where is these two different parts of DNA that could have come from totally different chromosomes that are now ligated together and so you are really just interested in this where this ligation occurred. So, areas where things that do not make sense that they are connected or connected. And then you do the sequencing, and then you can see what it looks like is connected together.

(Refer Slide Time: 12:01)



And you get things that look like this. So, this is chromosome 14 how it is folded on itself. And so you can see different patterns of where certain parts of the chromosome is folded on itself. So, this is just using different restriction enzymes of chromosome 14, and looking to see how the ligation patterns, and how the seq these areas here are connected throughout the chromosome.

(Refer Slide Time: 12:28)

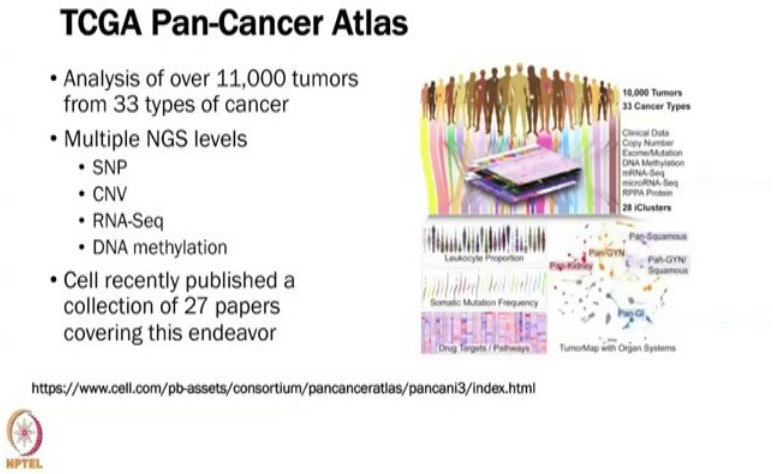


So, the last thing I wanted to just touch on with these publicly available datasets. So, the first is the TCGA. this has everyone has anyone heard of the TCGA?

Student: (Refer Time: 12:27).

Some people, you are going to hear a lot about it this week. So, the TCGA is a collaboration between the National Cancer Institute and the National Human Genome Research Institute. And what they have done is generated genomic maps of 33 tumor types, and then there is encode which is another of one of these large consortiums, where it is really looked at epigenetics. So, all of the parts list of the functional elements, so they have done a lot of the epigenetic work in mostly in cell lines.

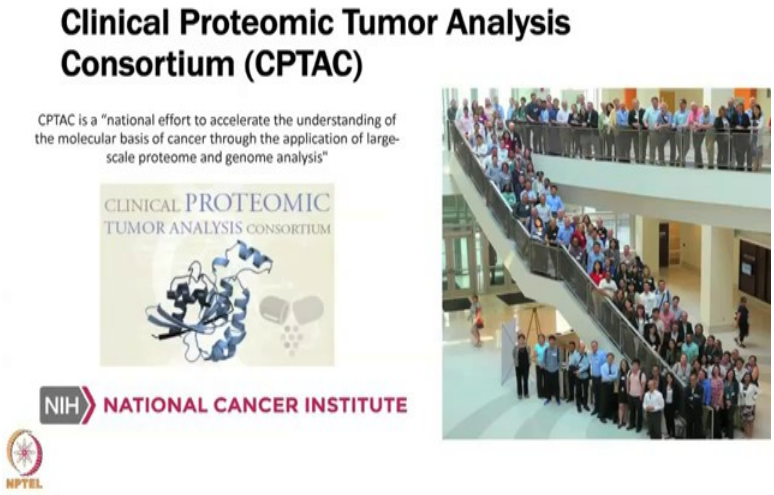
(Refer Slide Time: 13:13)



And so, the TCGA has been working for many many years, and has you know has published a lot of these papers to that are just sort of characterizing specific tumor types. So, characterizing breast tumors, characterizing ovarian tumors, looking at what mutations occur a lot and what subtypes exists in those tumors. And recently within the last few months, there was this pan-cancer atlas that was published where they took all of the tumors it did not matter what cancer it came from they analyzed them all together. So, this is 11000 tumors; they had a lot of these different sequencing levels that we have discussed.

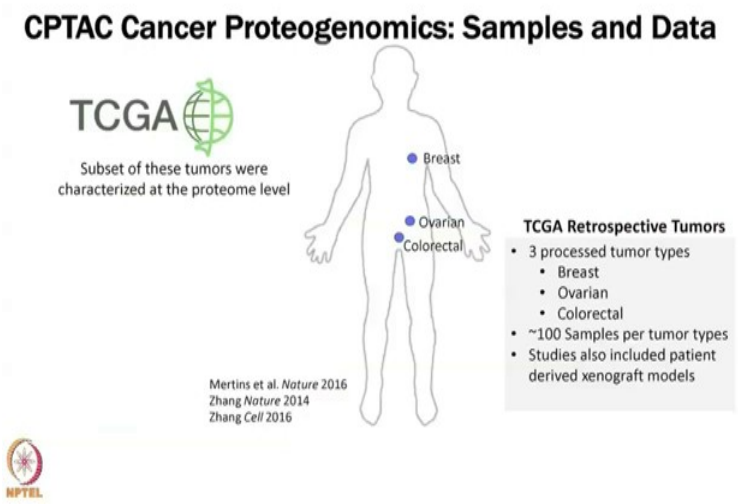
So, SNPs, copy number, RNA-Seq, DNA methylation and if you are interested there is I included it here. There is a collection of 27 papers that go through all of the many sub projects they came out of this. And so if you are interested in, you can take a look at that here, there is a lot of information there on these papers.

(Refer Slide Time: 14:22)



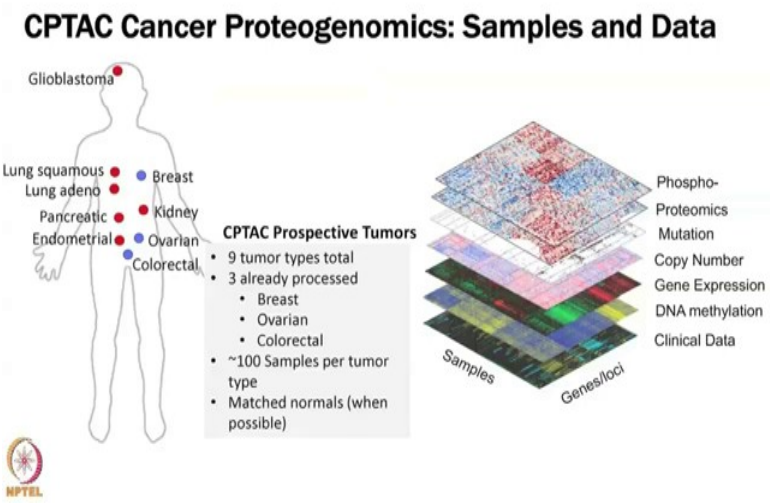
And with that I wanted to also introduce the Clinical Proteomic Tumor Analysis Consortium or CPTAC. So, many of the speakers who are here including myself have been involved in CPTAC for many years. And the goal of CPTAC is to accelerate the understanding of the molecular basis of cancer through the application of large scale proteome and genome analysis. So, we are really trying to take what the TCGA did and make it a proteogenomic analysis of cancer, and trying to figure out what more we can we can gather from using proteomics that we were not able to understand using the genomics. So, you can see here there are a lot of people involved in this. Our meetings are giant, and we always have to take a picture which is good because it is good for these talks.

(Refer Slide Time: 15:04)



So, the first the last iteration of CPTAC really just took tumors from the TCGA, and looked at them at the protein level. So, in this we looked at breast, ovarian and colorectal. These papers have been published already, everyone except for Bing who is not here yet worked on the breast analysis. You are going to hear a lot about the breasts CPTAC data. You will be very familiar with it. There were about a 100 samples per tumor type, and we took the same samples that the TCGA had done a subset of them. And then we did proteomic analysis on them, and then we did some integration of the of the data types that you will hear about for many of the people today.

(Refer Slide Time: 15:50)



And now the next iteration of CPTAC which was is ongoing. We are now looking at a perspective tumors meaning tumors that were not collected by TCGA. They are being collected specifically for CPTAC, and there are 9 tumor types total, and 3 of them are repeat. So, they are perspective samples from breast, ovarian and colorectal. And then there is about 100 samples that are collecting for these other tumor types as well. Again doing the same types of analysis and trying to better understand cancer by integrating all of these different methods.

Student: Other which has different colors and some are with the blue color.

Here the blues are the ones that we did the last time and that we are repeating with perspective, but that we have already kind of looked at and then the reds are the new ones that are yeah totally new.

Student: The homogeneity of all these type of cancers like you know the question raise many times that the cancer tissue itself is very heterogeneous.

Yeah I mean that is the problem we talk about a lot. And we know kind of the pathologists will look and see how heterogeneous, they can actually predict like how much tumor versus non-tumor we have, and we do a try and account for that. We do not do single cell RNA-Seq on these. It is just out of the scope, so yeah.

Student: Because it is also written that you tried for match normals wherever possible.

Yeah, we did try for match normals, which has been complicated.

Student: So, we could try to take the normal tissue around the tumor.

Yes, that is the goals. And for some tumors it is easier to use match normals; and for others, it is harder to define match normals for certain things like breast tumors right, because breast tissue is very fatty. So, match normals and breast is harder than in lung for example, right.

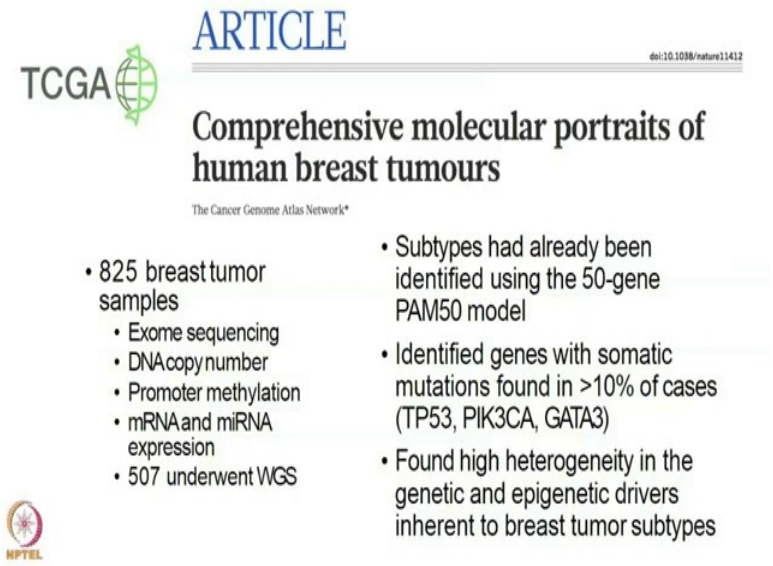
Student: The case would be with glioblastoma also.

Yes, yeah.

Student: Then how you get normal tissue around it.

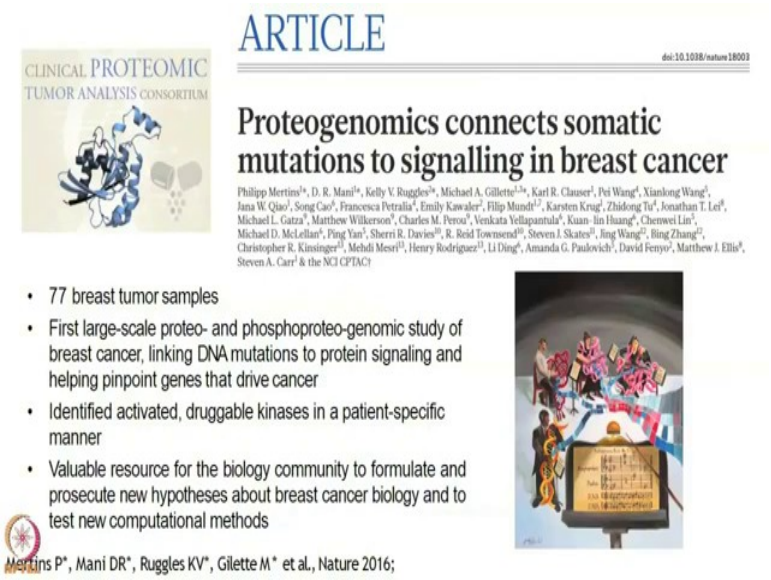
Yeah. So, it is a challenge, but it is the goal is to have match normals, because we it would be better to have a normal thing to compare our cancer too. So, we are going to look at the subset of the TCGA breast cancer study that was used for the CPTAC study.

(Refer Slide Time: 17:57)



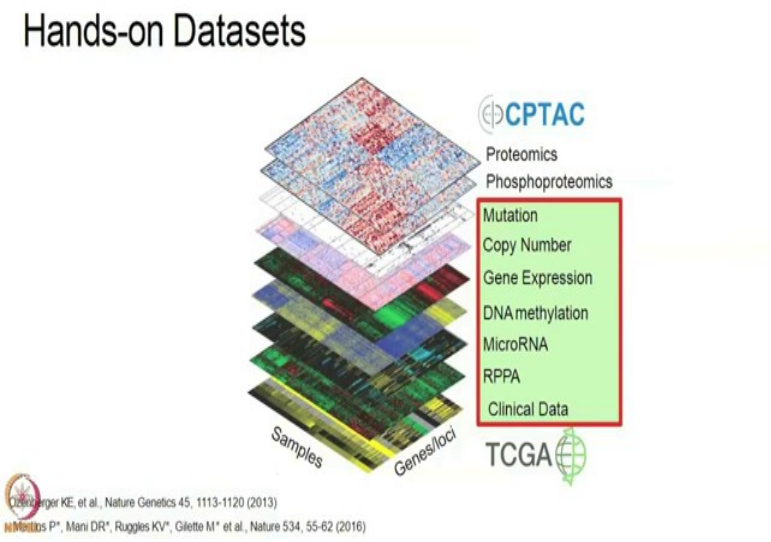
So, I just want to introduce the breast TCGA study. So, this was a study that was published, I think it was in 2012, where they looked at 825 breast tumors, they did exome sequencing, DNA copy number, methylation, mRNA and micro RNA expression. And then subtypes the these breast tumor is based on a PAM50 model which is a typical way of sub typing breast tumors. I will talk a little bit about what those subtypes are because you are probably hear about them a lot this week. And then identified genes identified genes with somatic mutations in different samples. So, they were really just characterizing breast tumors from a genomic perspective looking at the epigenetic drivers.

(Refer Slide Time: 18:44)



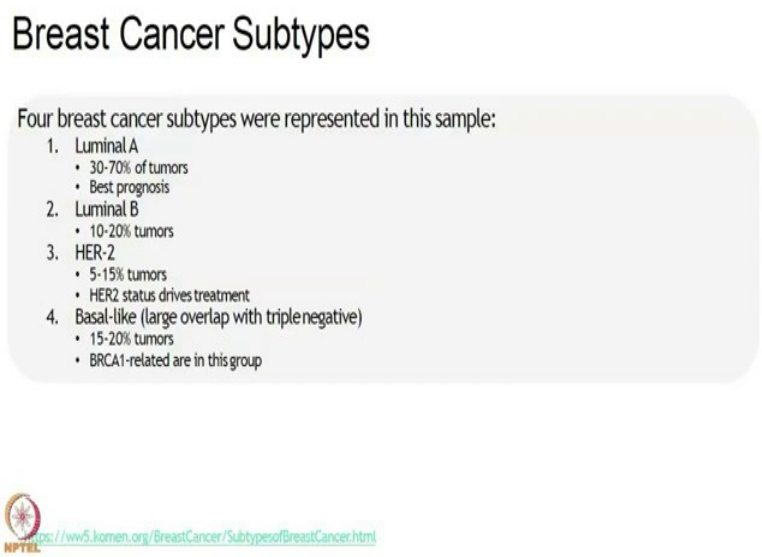
So, then last year, we published a paper that looked at the subsets of these tumor, so 77 of these tumors at the protein and phosphoprotein level. And this is the paper and this paper really looked at how the DNA mutations and protein signaling were connected. We looked at druggable kinesis, and patient specific manner. And it is really was to provide a resource for the community just like to the TCGA did, but now adding proteomics and phosphoproteomics to it.

(Refer Slide Time: 19:18)



So, for this for this hands on, we are just going to look at the genomics data since that was what I was tasked with. So, we are not going to do a proteomics analysis yet, but we are just going to look at some of the TCGA based genomics data from the 77 samples that we looked at.

(Refer Slide Time: 19:35)



And just wanted to mention that there are these PAM50 and subtyping was done for the tumors. So, there are 4 different subtypes that you will hear about luminal A, luminal B, HER-2 and basal-like, and they all differ in terms of prognosis, and in terms of how these patients are actually treated. So, there have been sub typing, and you will see that throughout many of the figures. The basal-like are typically that have the worst prognosis, so that is something that we tend to focus on quite a bit.

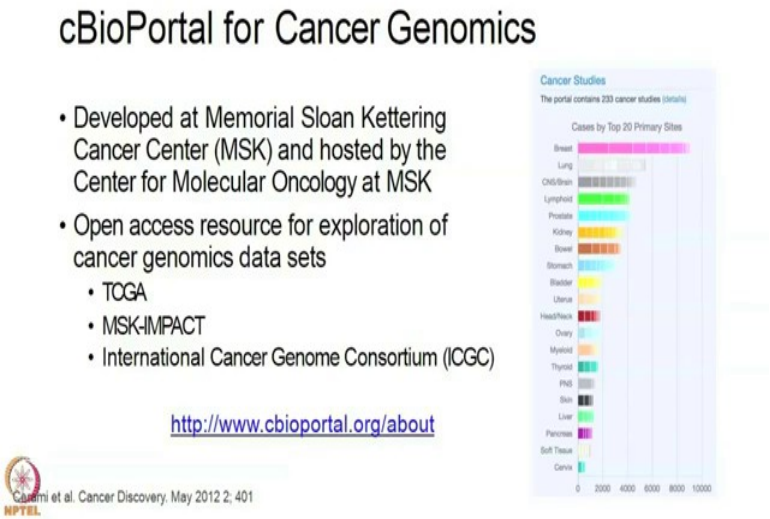
(Refer Slide Time: 20:12)

So, for this hands on the objective is really to, so this is figure 1 b from the Mertins et al., paper. And so what we wanted to do was to have you explore a lot of this genomics data that was used in this paper using a publicly available website, so cBioPortal. Has anyone used cBioPortal before?

Student: Used.

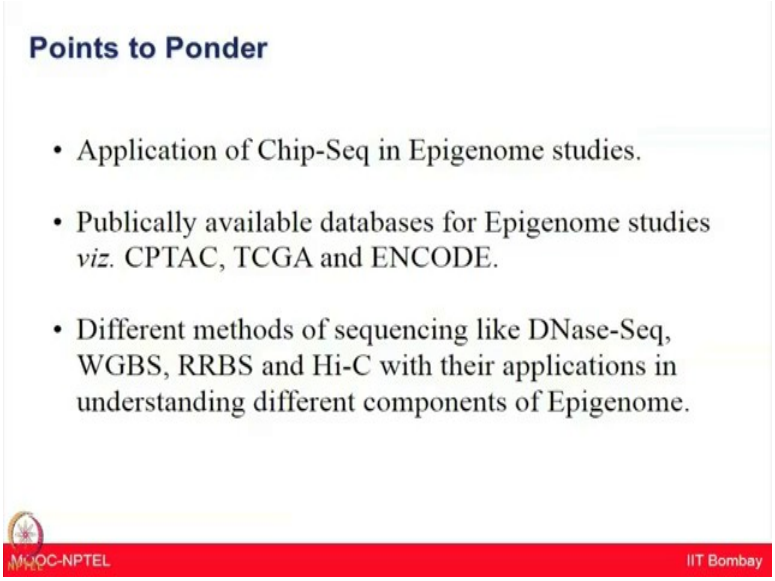
Ok, great. So, that no hopefully enjoy it ok.

(Refer Slide Time: 20:41)



So, I just wanted to introduce the cBioPortal. So, it is actually was developed by Memorial Sloan Kettering. And it is hosted that and their Molecular Oncology Center. At this point, there is a lot of people working on it, it is one of these huge, they have done a really good job of hosting of large amounts of cancer genomics data. So, they have actually taken the TCGA in addition to some other data sets and actually have it readily available on the website. So, you do not have to download anything, we did not want you to have to download lots of enormous files. So, you can just play with the data and explore the data on the websites.

(Refer Slide Time: 21:17)



So, in conclusion, I hope today you have learnt what is epigenetics and various methods which could be used for analyzing the DNA methylation in a given gene. We also learned about high throughput approaches like WGBS and RRBS which could help in searching and analyzing DNA methylation in the gene. We also learnt that chromatin conformation captured Hi-C may help in understanding the folding and interactions of chromosomes with adjacent or self genes using different set of restriction enzymes. Dr. Ruggles also briefed about TCGA and ENCODE which are publicly available databases containing very useful genomics and epigenetics information.

In the next lecture Dr. Ruggles will be giving a hands on demonstration on how to use cBioPortal for accessing gene mutations and its expression in published datasets. I hope these lectures are giving you not only understanding about you know how to analyze

genome, transcriptome and even you know epigenome, but also giving you information for various repositories databases and software tools available which could be freely available and can be utilized for your own research.

Let us continue this discussion about genomic technologies in the next lecture, and then we will have you know another transition in the concepts, and we will have another speaker to give you more fundamental concepts.

Thank you.