Interactomics: Basics and Applications

Prof. Sanjeeva Srivastava
Prof. Deeptarup Biswas
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 59
Multi-omics Approach for Better Understanding of Cancer: Case Study

In this course we have discussed about various technologies, which provides more functional information more protein protein interaction another biomolecular interactions which gives us the functional consequence of the proteins. We talk to you about different type of technology platforms including protein microarrays label free bio sensors. And even next generation sequencing and mass spectrometry will be free. Idea is that finally, all of this data which we generate, from the wet lab experiments has to be made available to the public.

And likewise the whole research committee should start sharing the data. And as you have seen in the last few lectures, we are giving you the glimpse an idea that how you can start using the public databases and various resources from which you can download data. And now you can use open source software's to analyse the data set. In the field of cancer research and especially by many clinical research, a lot of data being generated which is very costly and very precious. The large funded program from different governments, which aims to investigate thousands of patient sample analysis.

To the new field emerging which is known as proteogenomics, where aim is to look at from the same patient sample can we analyse both proteins and gene level. And if you could do integrate that is information, then probably we are able to get more meaningful information from the same sample type. In today's lecture one of my research scholar Mister Deeptarup is going to show you how proteogenomics by integrating data from the genome and proteome, could provide us some novel insight from the literature reviews of publish datasets.

He will also talk about how proteogenomic approaches can help to resolve various issues of diagnosing different grades of cancer. Or looking at different subtype of cancer which is very difficult to understand without having a very good molecular level understanding. He will

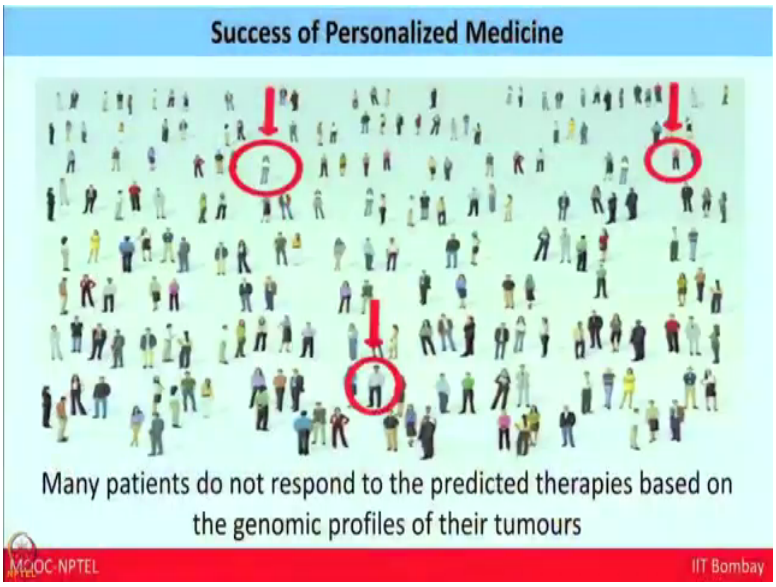
also explain how proteomics and genomic data correlation can provide a much broader and meaningful picture of progression of cancer. The different excellent program funded by national cancer institute in US from an IH have really accelerated the whole field of proteogenomic research.

Scientists are sharing data obtained from tissue microarrays, protein microarray platforms, next generation sequencer and mass spectrometry and making it publicly available. The Cancer Genome Atlas or TCGA is an excellent resource for you to get it started and do more investigation further from the raw data which you can obtain from this portal.

And likewise now CP tag and other ambitious programs are available from which one could accelerate different type of proteogenomics research. Today Deep will talk to you about the workflows of some of the cases studies, which were recently published using the workflow of proteogenomics in the area of cancer. So, let me welcome Deeptarup for his today's lecture.

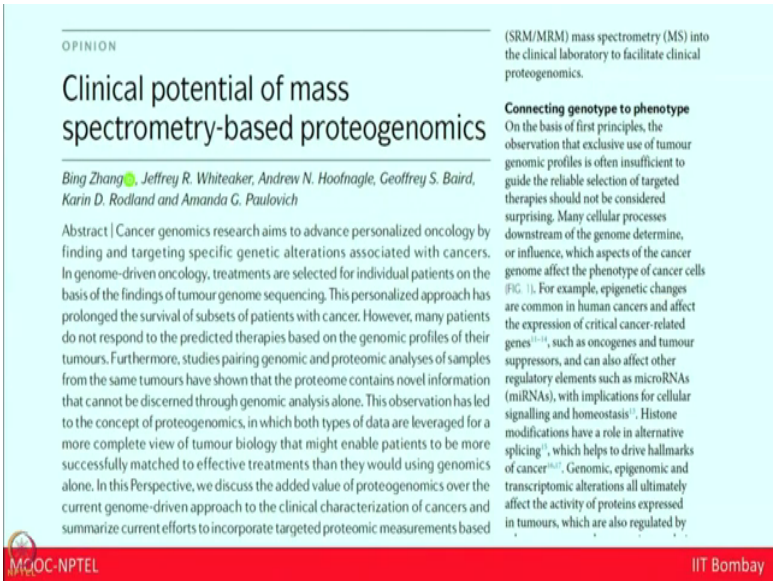
After the completion of human genome project and introduction of genomics into the disease pathobiology, there was a hope that genomics can lead to can bring revolutionary change in the cancer diagnosis.

(Refer Slide Time: 04:02)



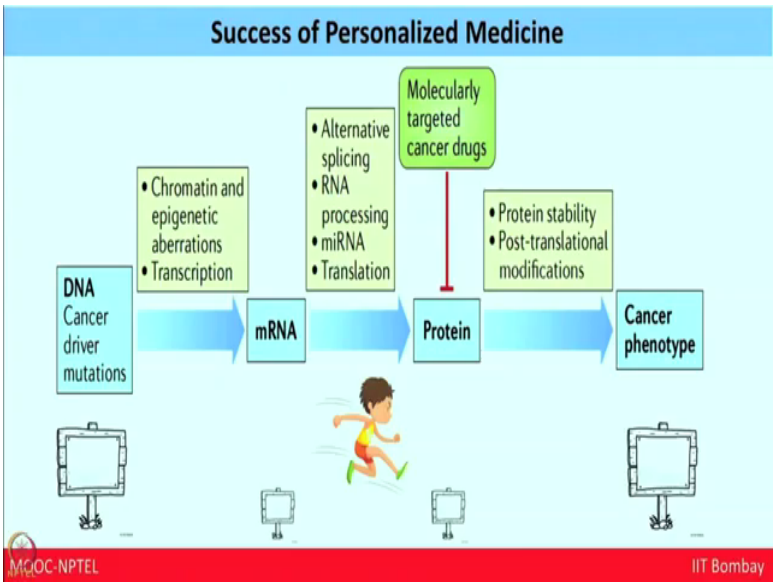
And can lead to a path to personalized medicine. But the success of personalized medicine with the help of genomics was not that much revolutionary. From overall cohort of patients only few patients were respond to the predictive therapy based on the genomic profile. There were some loopholes that were still present after the successful outcome of genomics.

(Refer Slide Time: 04:27)



So, recent paper published from Zhang group there is a clinical potential of mass spectrometry based proteogenomics. So, in this paper he has talked how the clinical potential of the mass spectrometry based proteogenomics can be introduced. The personalized medicine with the help of genomics was not that much successful total number of reasons.

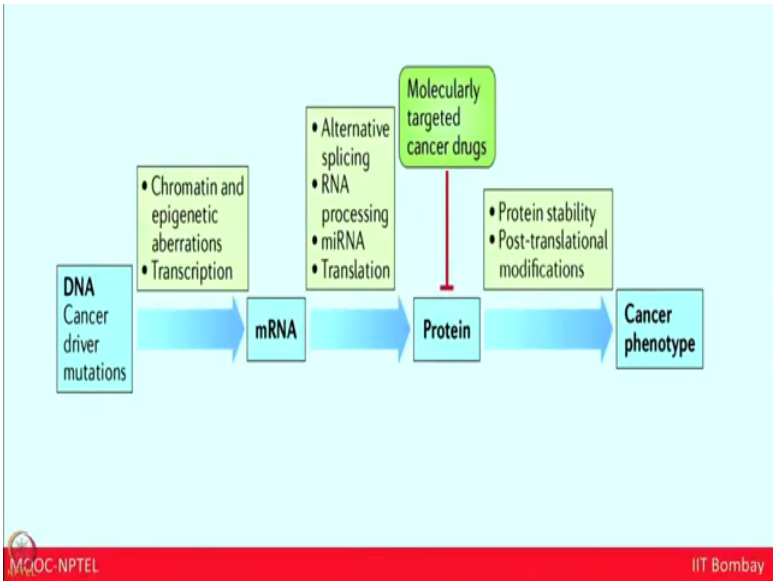
(Refer Slide Time: 04:50)



If we can see that, with the help of genomics solving the problem like cancer is like jumping from one hurdle to the last hurdle. And we are not taking into account a number of conditions and parameters that is coming in between the two hurdles. So, we are getting a complete profile of the genomics different types of mutations different aberrations; but in the same hand we are missing different epigenetic aberrations transcription and regulations alternative splicing and protein proteomics profiling.

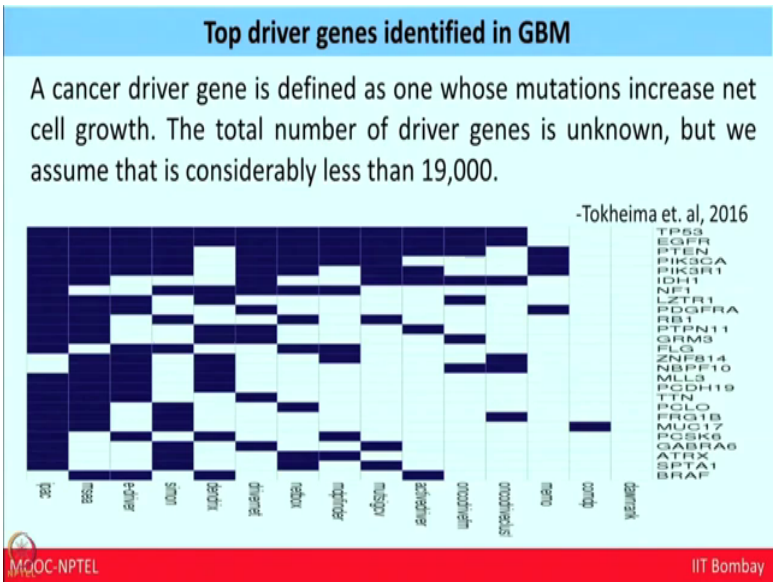
So, all this important information need to be taken into account. To understand the pathobiology of the cancer and then only this can this tool can be used for the diagnosis and treatment.

(Refer Slide Time: 05:47)



So, the message from this slide is that all this information starting from DNA to mRNA to protein need to be considered to reach to the goal and, to diagnose to bring a revolutionary change in the cancer and a cancer diagnosis and treatment. So, before I move how proteogenomic is playing a role in cancer diagnosis, I want to give a brief account of what is cancer driver genes?

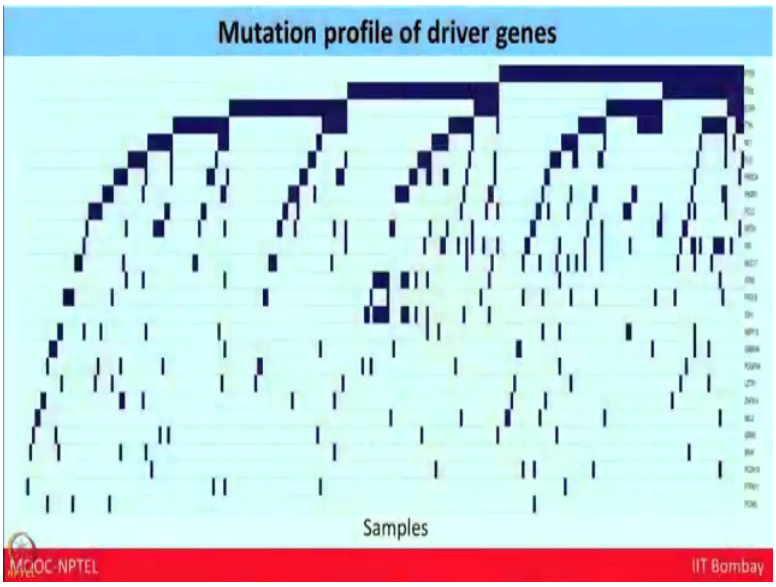
(Refer Slide Time: 06:21)



So, cancer driver gene is defined as one whose mutation increase net cell growth. The total number of driver gene is unknown, but we assume that is considerably less than 19,000 which has been given by Tokheima et al in 2016. So, from driver d v repository.

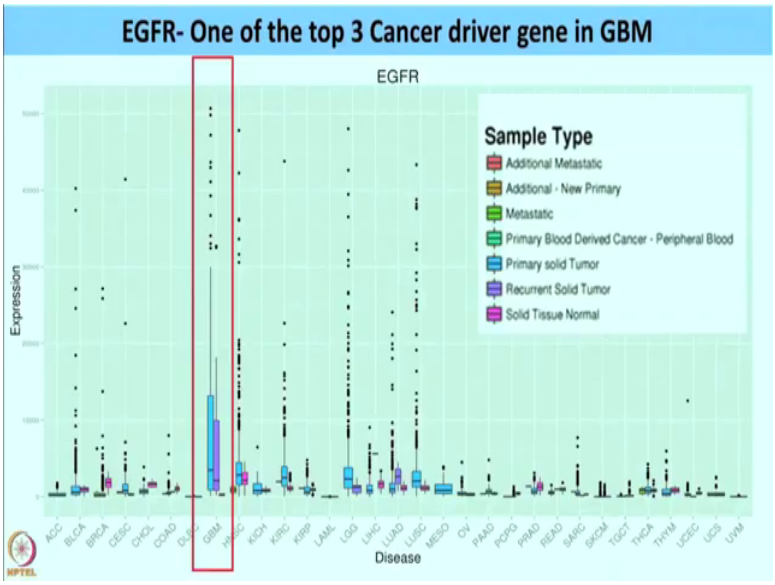
You can see like the top driver genes includes TP 53, EGFR PTEN. And how this hallmark driver genes are important in the glioblastoma in the glioblastoma tier tumorigenesis we all know.

(Refer Slide Time: 07:02)



So, here is the mutation profiles of those driver genes, where the top driver genes are PTEN TP 53 EGFR. And we can see the mutational profile in terms of samples which is in the x axis.

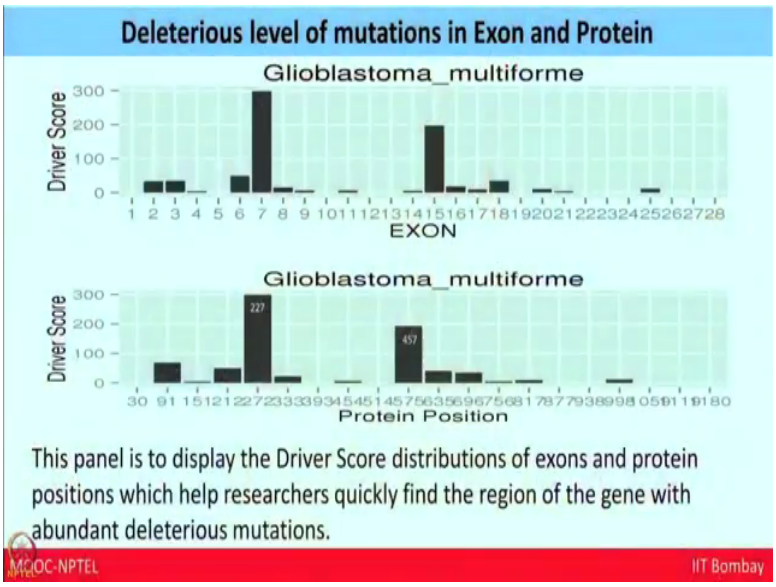
(Refer Slide Time: 07:14)



So, if I if we choose one of the top three cancer driver genes that may be EGFR and we can understand that what is the expression of this EGFR gene in glioblastoma? So, we found that the expression of the EGFR gene in glioblastoma is pretty high. So, one of the top three cancer driver gene in glioblastoma is EGFR.

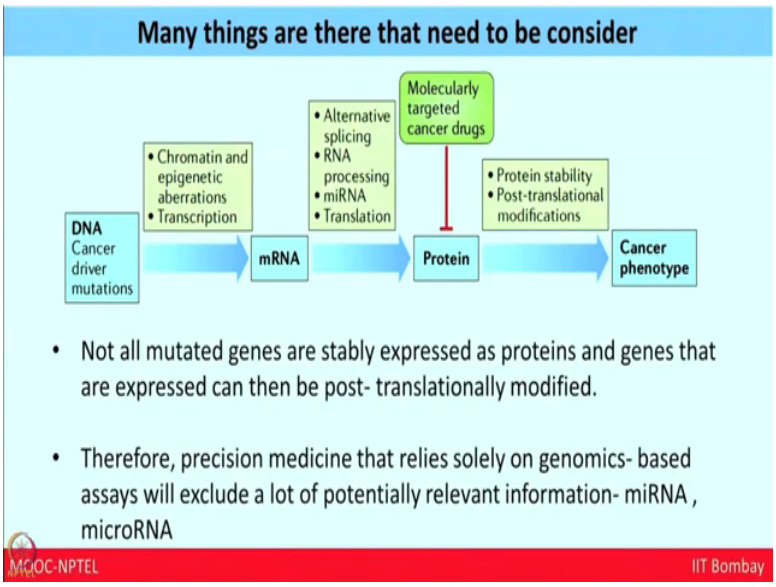
And if we want to check the expression of EGFR in terms of in taking into account the other cancer we found in case of GBM EGFR is highly lower express in both primary solid tumour and recurrent solid tumour. So, till now the genomics has given a lot of information about glioblastoma.

(Refer Slide Time: 08:06)



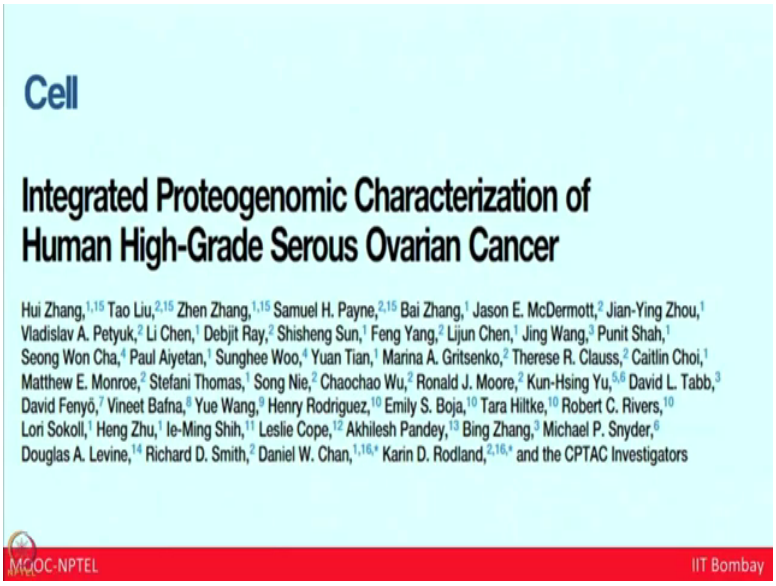
But if we taken into account the correlation between the EXON and protein, we will found that the drivers code related to protein and exon is also giving some new information. This panel is to display that drivers code distribution of exon and protein position which help research researchers quickly find the region of the gene with abundant deleterious mutations.

(Refer Slide Time: 08:24)



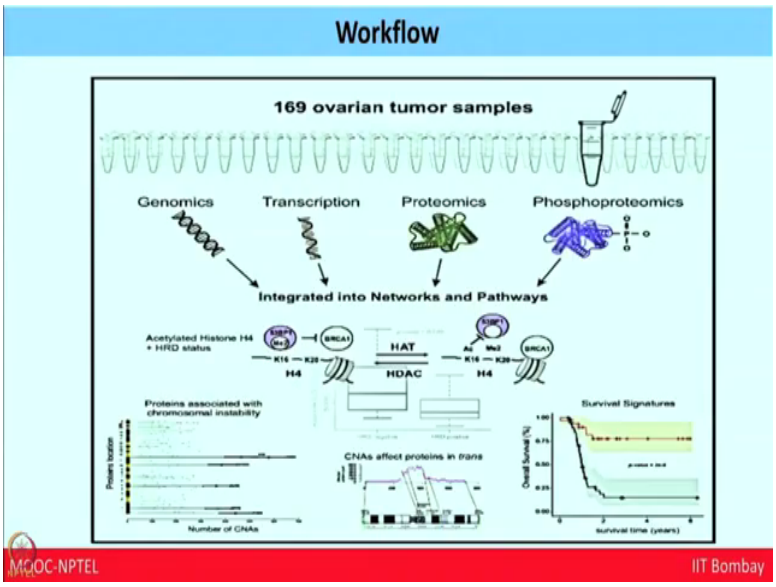
So, now we understand that we did not consider a lot of things between the genomics and the precision medicine. That not all mutated genes are stably expressed as proteins. And genes that are expressed can be post translationally modified. Therefore, precision medicine that relies solely on genomic based assay will exclude a lot of potentially relevant information like miRNA microRNA.

(Refer Slide Time: 08:53)



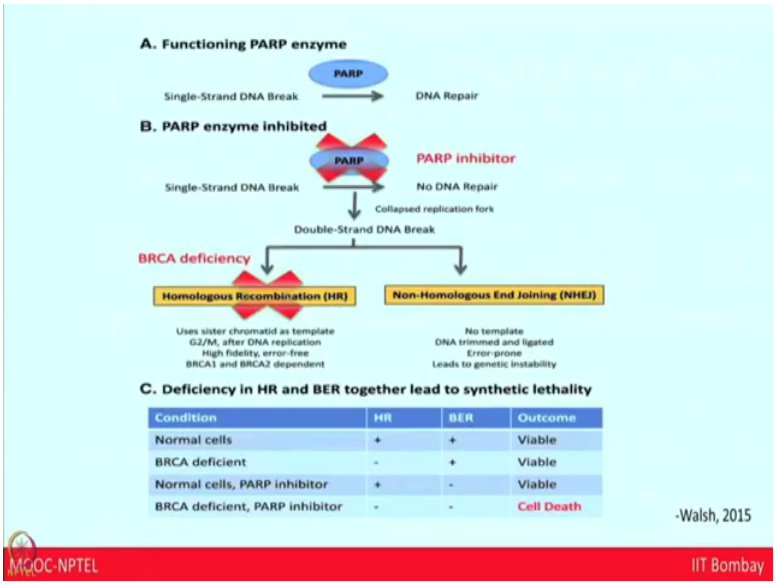
So, the support the previous statements and to give you a complete glimpse how the powerful tool of proteogenomics can be very helpful, to solve different kinds of cancer.

(Refer Slide Time: 09:08)



So, in this study they have taken 169 ovarian tumour samples from TCGA meta data and they have they tried to analyse rather correlate the genomics, transcriptomics, proteomics and phosphoproteomics.

(Refer Slide Time: 09:24)



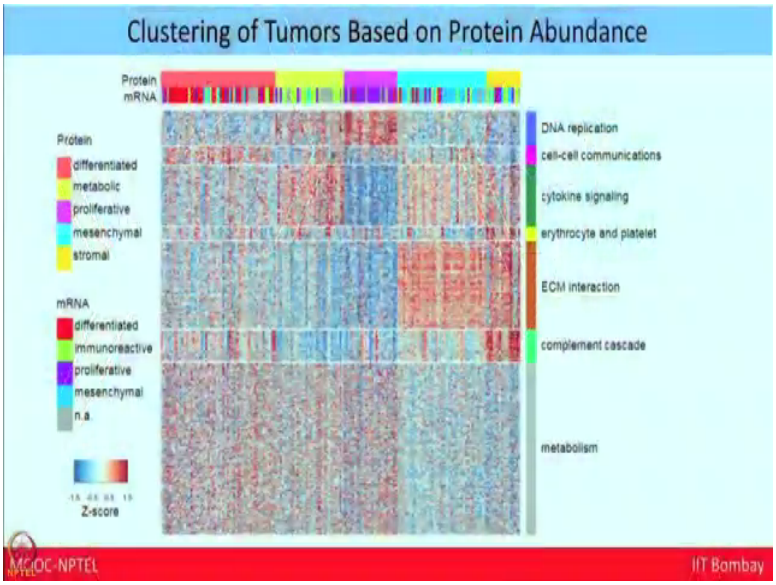
So, before going into the paper let me give a glimpse of this kind of mutation and how this mutation can lead to lethality of a cell. So, the diagram has been taken from Walsh et al 2015; where we can see the functioning of PARP enzyme and how PARP enzyme is helping in DNA repair of single strand DNA break. If PARP enzyme is inhibited so, there is no repair takes place and which helps which rather lead to collapse replication fork.

And the BRCA deficiency do not allow homologous recombination to happen. In C, the deficiency in the H r homologous recombination and base excision repair together lead to synthetic lethality, than the correlation.

So, the sample information, tumours were selected by examining the associated t c g a metadata to select tumours. On the basis of putative homologous recombination deficiency,

presence of germline or somatic BRCA 1 or BRCA 2 mutations, BRCA 1 promoter methylation or homozygous deletion of PTEN's were taken.

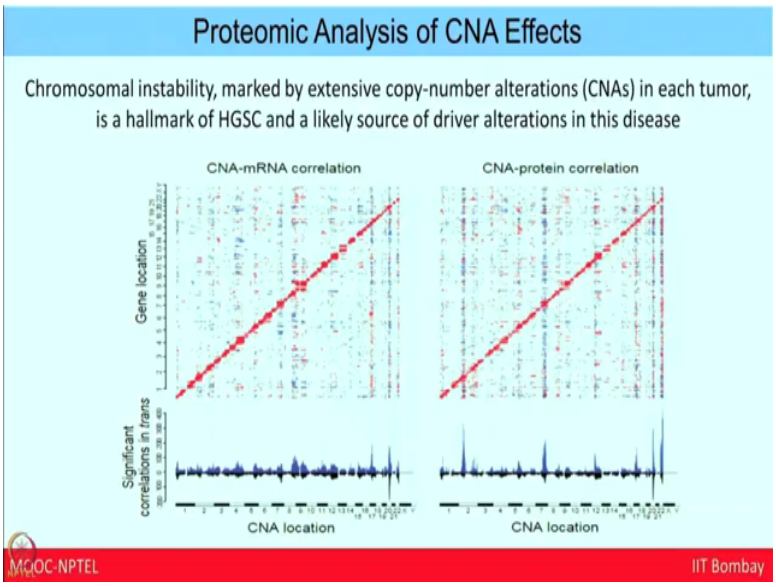
(Refer Slide Time: 10:36)



So, this clustering will is giving us the complete landscape of what are the different pathways are involved. And how protein and mRNA are playing role and what is the correlation between the protein and mRNA in this pathway.

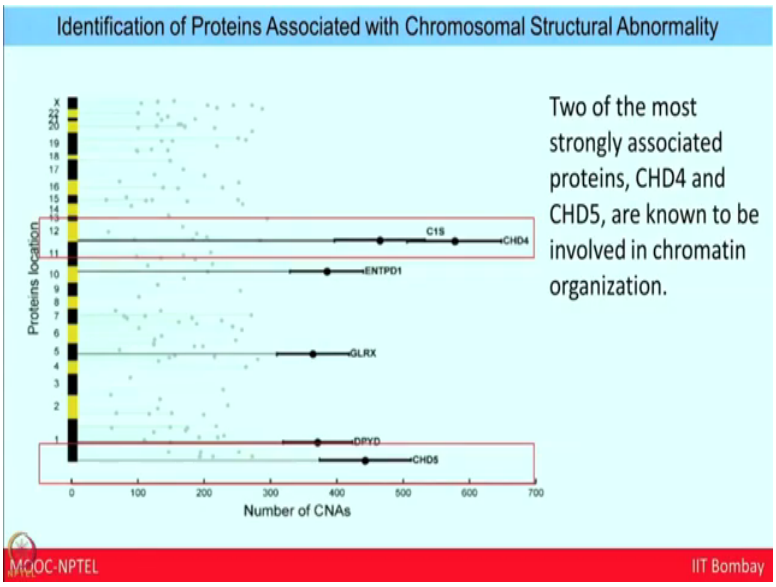
So, till now we understand that the protein and mRNA correlation is there and how this protein and mRNA co correlation is also playing a role in terms of biological pathway. But now they also tried to understand that how CNA that is copy number aberration in each tumour is playing a role with protein and mRNA correlation.

(Refer Slide Time: 11:17)



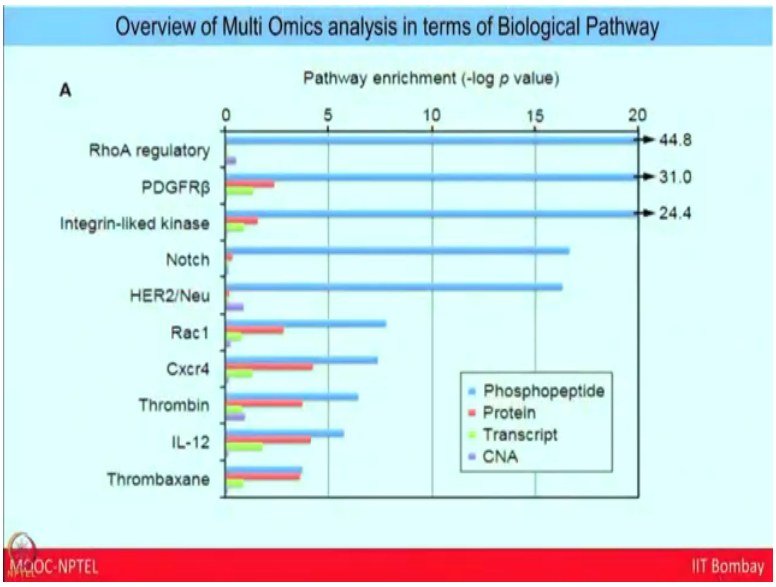
The blue one at the complete profile of the data generated whereas, the black one is the data that is present that is all that is already present in the database.

(Refer Slide Time: 11:33)



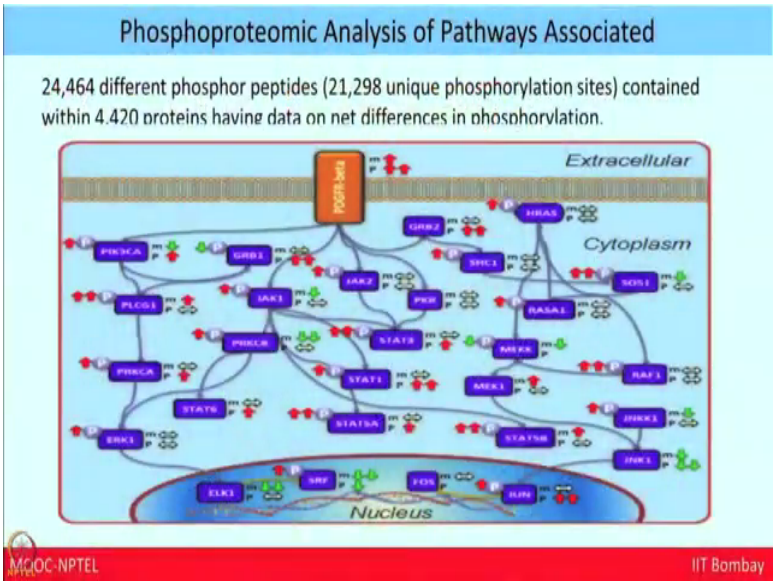
So, from this CNA mRNA correlation and CNA protein correlation they found that two important two important protein; that is CHD 4 and CHD 5 are having the maximum number of CNA CNAs. So, when the further studied they found that these two proteins are involved in chromatic organization. So, to understand the complete biological pathway they take they took phosphopeptides proteins transcripts and CNA.

(Refer Slide Time: 12:00)



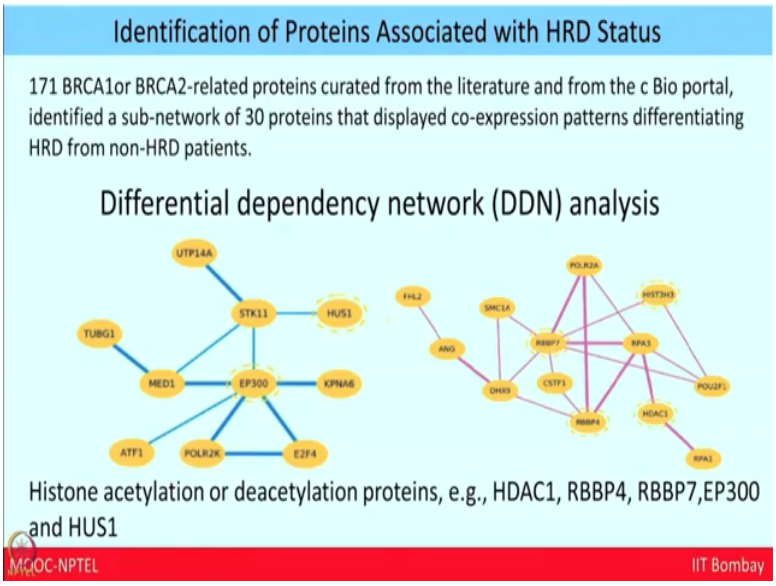
And they found that these are the top pathways that is playing a role in this cancer pathobiology. So, out of which PDGFR beta which we all know is a angiogenic, receptor is also showing an important correlation in terms of biological pathway. To understand the complete landscape of the cancer pathobiology they incorporated mRNA protein.

(Refer Slide Time: 12:26)



And phosphopeptide data into one picture and where we can we see that the PGA for beta is up regulated in both mRNA and protein. So, this up regulation of the PGA for beta is not only giving a clue to a active angiogenesis, but also showing that. How, what are the different downstream regulatory factors that are also up regulating or down regulating in terms of mRNA and protein.

(Refer Slide Time: 12:59)

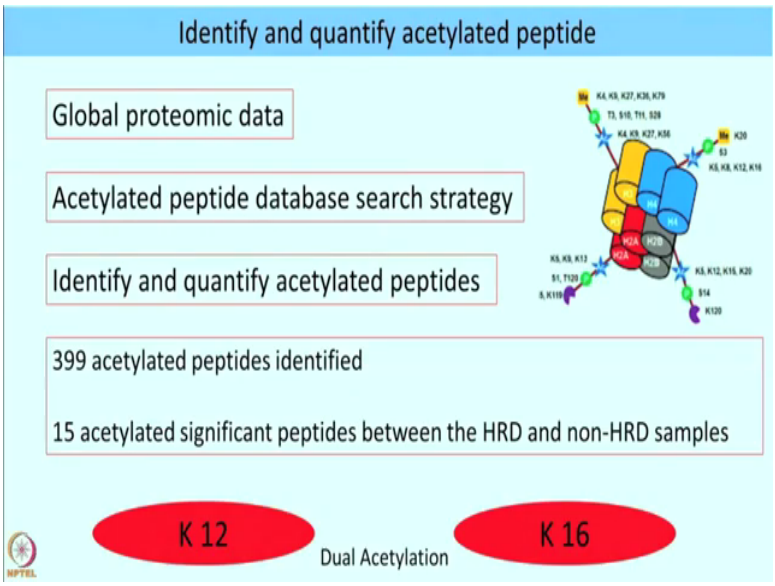


So, further they tried to do a DDN analysis. So, DDN analysis is differentially dependency network analysis where, the proteins curated from the literature and from the c bio part portal. So, c bio part portal helps you to get that data out from the TCGA.

And they identified a sub network of 30 protein that displayed co expression pattern differentiating from HRD from non HRD patient. And from this DDN analysis they found that histone acetylation or deacetylation proteins are coming are playing are coming into the clusters and which includes HDAC1, RBBP4 our RBBP7, EP300 and HUS1.

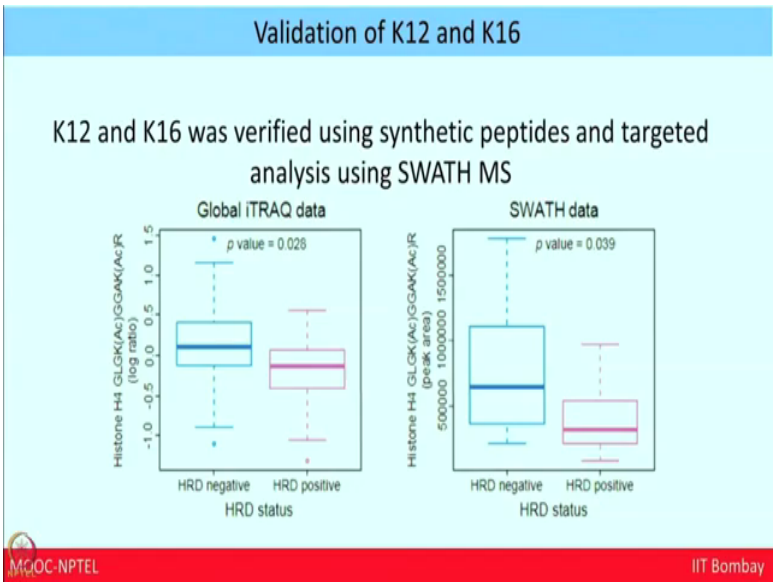
So, from the last part of the study they understand that histone acetylation and deacetylation are playing an important role. So, this clue was enough to give an idea that acetylated peptides need to be studies.

(Refer Slide Time: 13:57)



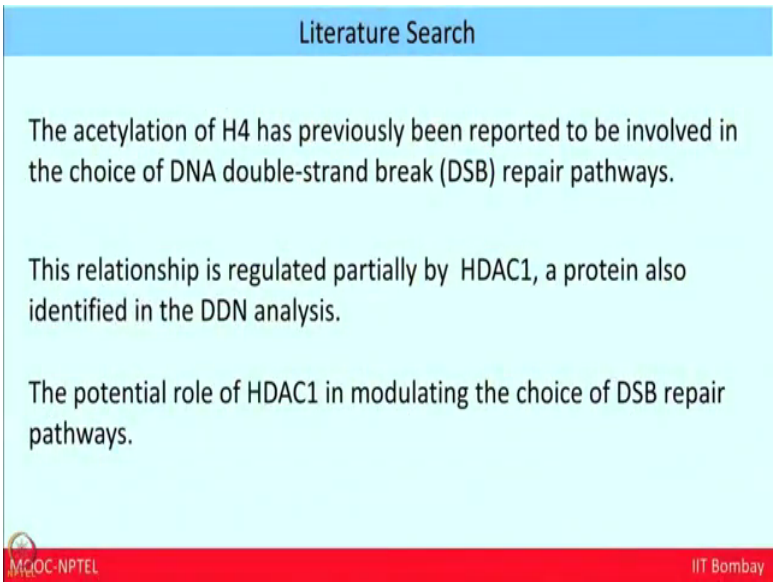
So, from the global proteome data, they prepare a acetylated peptide database search strategy and identify and quantify the acetylated peptides. From there they identified around 399 acetylated peptides and 50 acetylated significant peptide between HRD and non HRD. So, as so, from this 15 acetylated significant peptide they found that K 12 and K 16 is a that is acetylation of lysine in 12 and 16 were found.

(Refer Slide Time: 14:29)



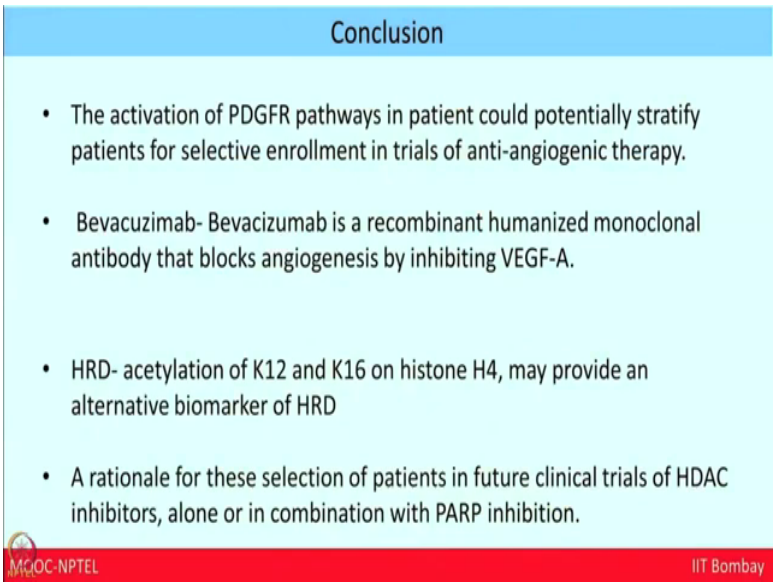
So, they validated the K 12 and K 16 using synthetic peptide and targeted analysis using SWATH MS. In the same thing, they found that the K 12 in terms of iTRAQ data were upregulated in HRD negative and same thing has been validated in sort and they found a same upregulation in HRD negative.

(Refer Slide Time: 15:05)



So, they went back and further search in the literature and they found that the acetylation of the H 4 has previously reported to be involved in the choice of d DNA double strand break DSB repair pathway. The relationship is regulated partially by HDAC one a protein also identified in DDN analysis the potential role of HDAC in modulating the choice of DSB repair pathway has been identified.

(Refer Slide Time: 15:30)

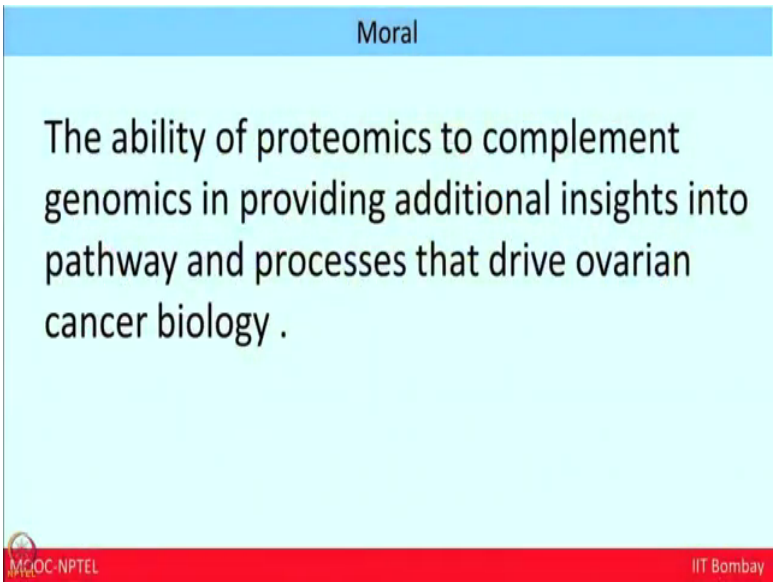


So, the conclusion from the study, we understand that the activation of PDGFR pathway in patient would portion potentially stratify; selective enrolment in trial of anti angiogenic therapy.

Recombinant human monoclonal antibody that blocks the angiogenesis by inhibiting VEGF-A has already been trailed in patients. So, the PDGA for pathway the involvement of PDGA for pathway in this cancer is also giving this recombinant humanized monoclonal antibody role in limelight. Apart from this HRD acetylation K 12 and K 16 on histone H 4 may provide an alternative biomarker of HRD.

A rationale for these selection of patient in future clinical trials of HDAC inhibitors alone or in combination with PARP inhibition can be also tried.

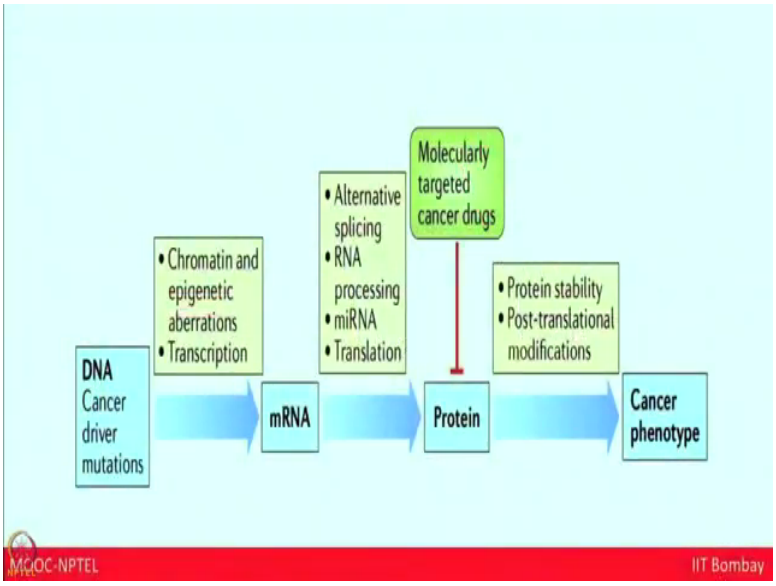
(Refer Slide Time: 16:23)



So, the moral from the study we understand, the ability of proteomics to complement genomics is, providing additional insights into the pathway and processes that drives ovarian cancer biology. Not only that complete data which we are getting from the genomics is not enough to lead to a well profile diagnosis and treatment of cancer.

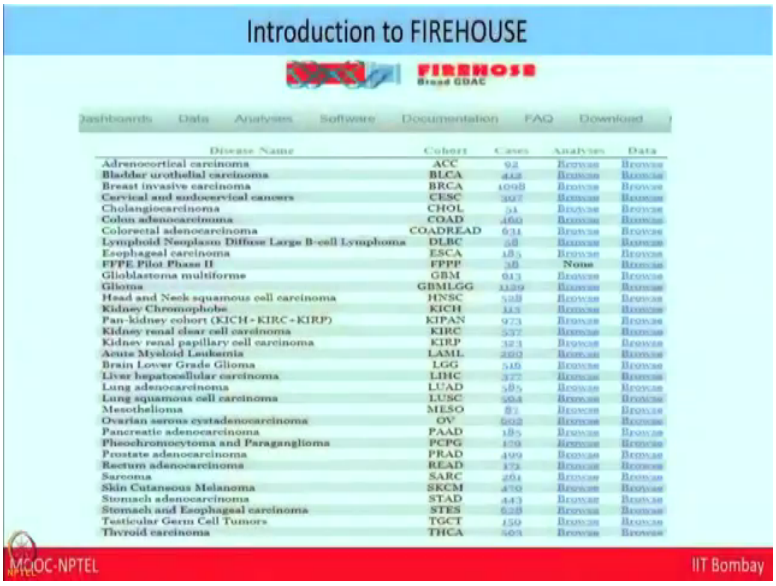
So, all the important things like mRNA information protein information and PTMs, the both translational modification information need to be gathered and further correlated among themselves. And then only we can reach to a conclusion and we can take this information and further validated in clinical trials.

(Refer Slide Time: 17:06)

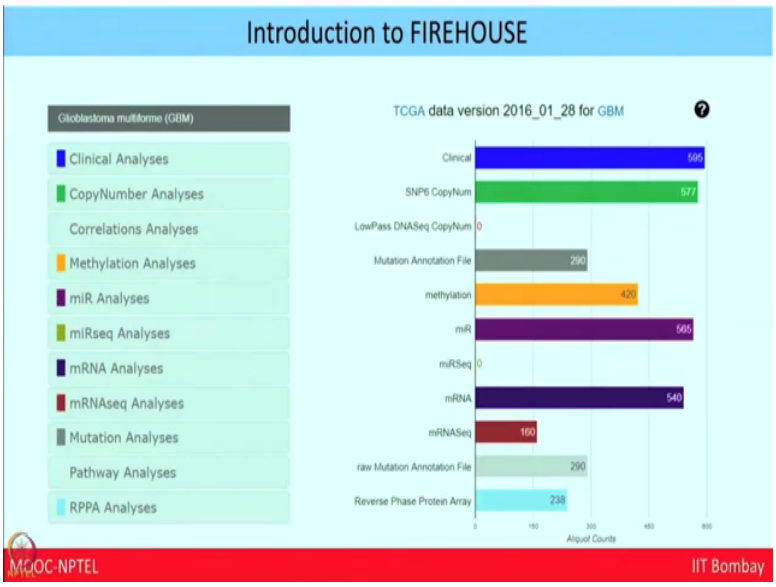


So, now we understand that how cancer driver mutation mRNA protein need to be taken into account to reach to the molecular target or cancer drug. From the last study, we understand that how the group has only generated the proteomics data. And they have tried to co correlate the their proteomics data with the already available mRNA CNA data from the databases.

(Refer Slide Time: 17:34)



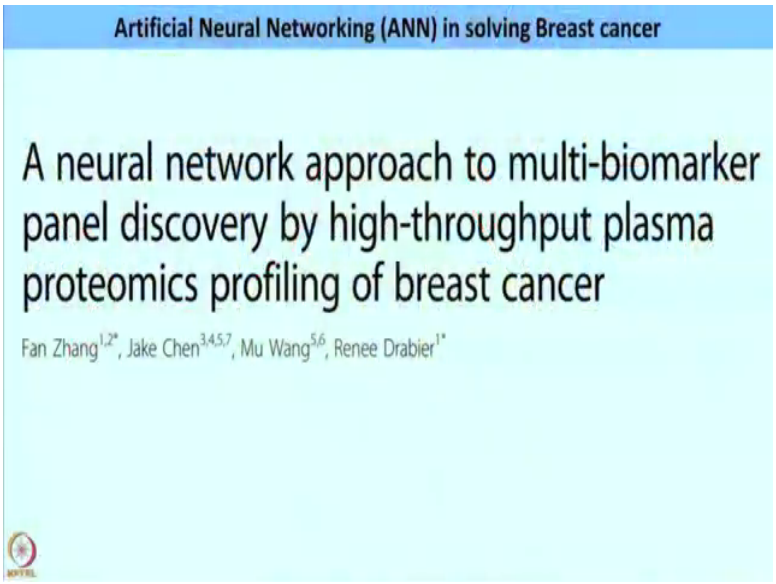
(Refer Slide Time: 17:36)



So, firehouse can be used to download this kind of data, like if we select a disease name that may be glioblastoma multiforme. And we can see like all the data which are available in the TCGA, can be downloaded from here. So, TCGA data version from 2016 from glioblastoma, clinical, SNPS, methylation and mRNA sequencing data and reverse phase protein array data's are already available. So, we can use this firehouse to download the data. So, now, we are able to understand how proteogenomics and correlation of mRNA and protein can give us better insights of a particular disease.

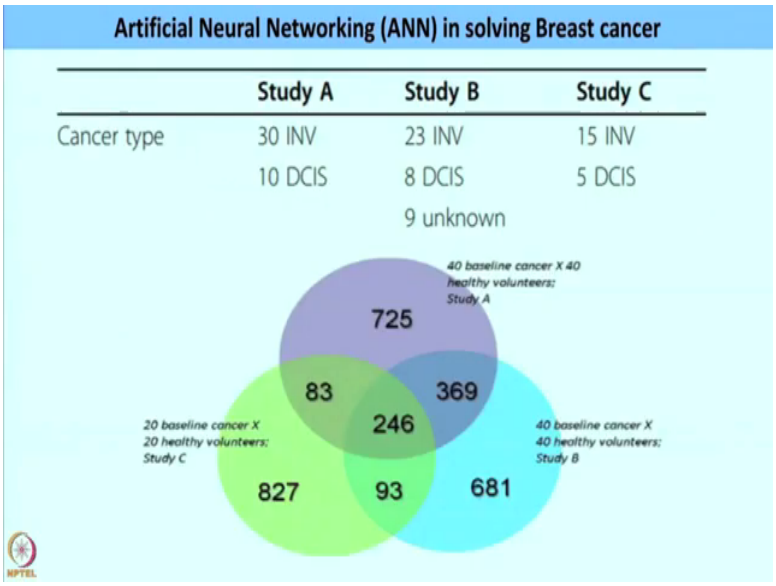
But to deal with this amount of big data, prepare a panel which can help in the treatment or diagnosis of cancer. We need to think about different predictive and machine learning based analysis. I have taken an example of a paper a neural network approach to.

(Refer Slide Time: 18:35)



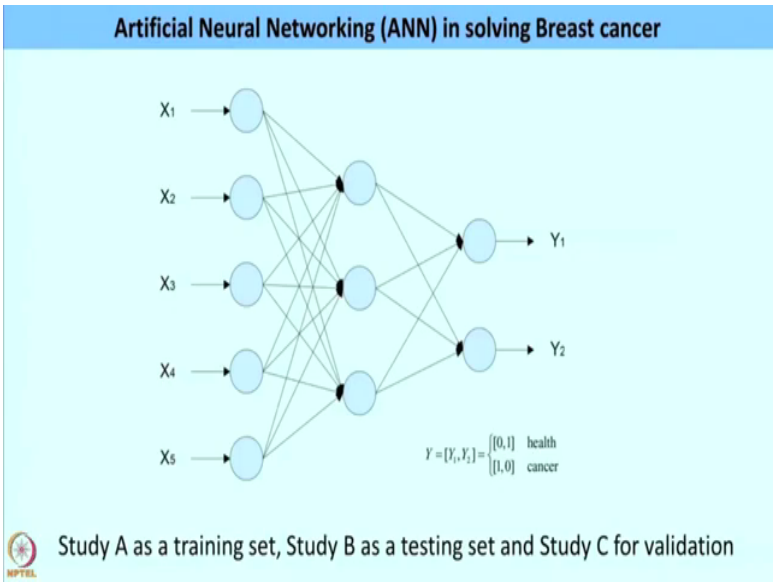
Multi biomarker panel discovery by high throughput plasma proteomics profiling of breast cancer.

(Refer Slide Time: 18:43)



Where in study A and study B 40 cancer types and 40 controls were taken; where is in study C 20 cancer types and 20 controls were taken. Further they have done the proteomic analysis and they found the 246 proteins are common between three studies.

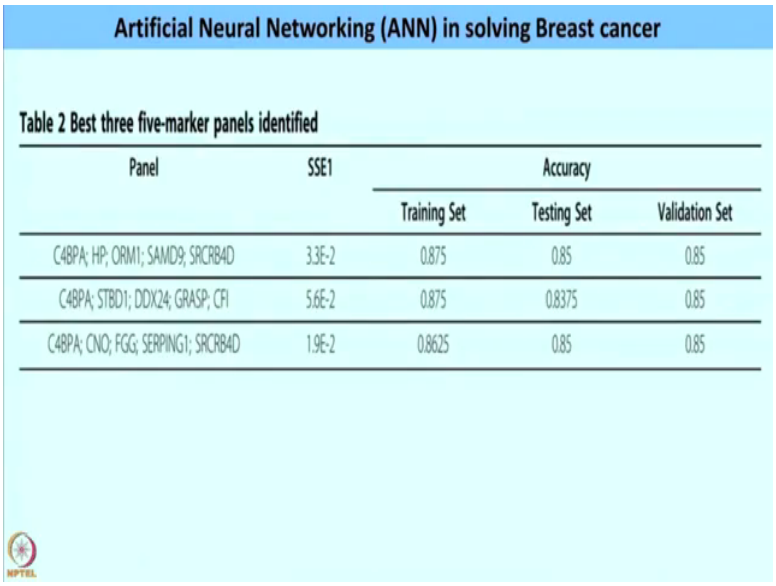
(Refer Slide Time: 19:03)



After this analysis they have taken the data and tried to prepare a artificial neural networking model taking study A as a training set, study B as a testing set and study C validation. So, in this kind of artificial neural networking in most of the cases for the training set maximum; that means, around 70 percent or more data need to be taken.

Whereas, for study B, 30 percent data need to be taken the model further validated with blind data set to check the efficiency of the model. In most of the cases, the accuracy of the model need to be more than 80 or 85 percent.

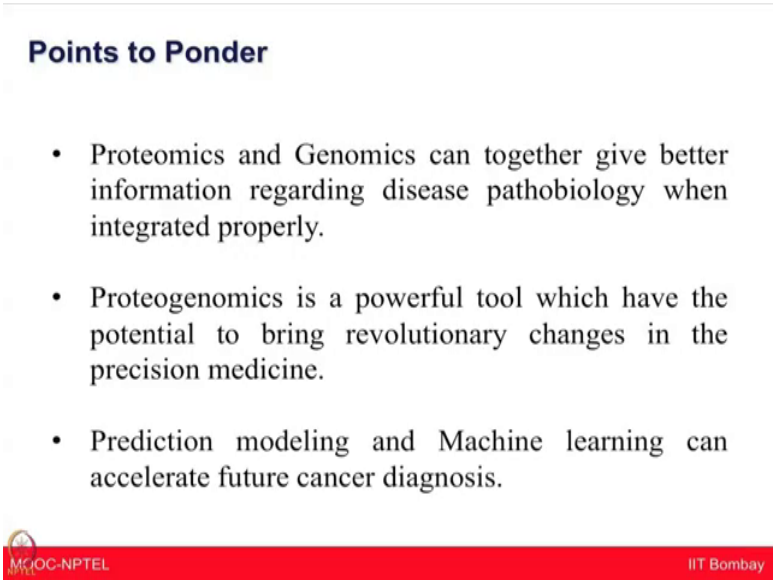
(Refer Slide Time: 19:45)



So, this artificial neural networking gives a panel base three panels with 5 markers and with the accuracy more than 85 percent. So, further this panels were taken forward and checked in large cohort of samples to validate the data. So, like this we can use artificial neural networking and different made of machine learning strategies to understand and predict top candidates that are playing key role in tumorigenesis and further development of the cancer.

So, the main concept is the different protein understand the complete pathobiology. And then only the landscape of disease can be drawn and from there we can understand and we can and that can lead to a drug target or precision medicine.

(Refer Slide Time: 20:39)



So, by now you know that there is a huge amount of data that is available in the public repositories and databases, which could be utilized and extracted for the further data analysis. The big research programs like human protein atlas the cancer genome atlas or TCGA.

As there as different laboratories working worldwide including broad institute of Harvard and MIT, have shared the data into various databases. All of these researcher and scientist are making their data publicly available. More recently, now the cancer moon short project and international cancer proteogenome consortium; also aims to investigate the proteogenomic data from the same patients and intention is to make it publicly available and shared with the entire scientific community.

All you have to do is track the data perform different type of analysis and then make meaningful insight. You can always define a unique question from the same sample and look at what is the best answer from the large number of sample data sets available to you.

You can also compare the data from different laboratories or even integrate data obtained from different population; look at the effect of the same disease in different geographical locations, different races, different age groups, as well as the effect of different treatment or certain diseases which may have you know the recurrence nature. Many of these things could be investigated from these kind of publicly available dataset.

I hope these manuscript which we are discussed today, have given you a very impressive glimpse of how genomics, transcriptomics, proteomics and sometimes your metabolomics, together could provide you the much in depth information. At the silver level which was otherwise not possible few years ago. I hope you will be able to use some of these technologies and some of these datasets in your own research.

In the next lecture, I will talk to you more about the various revolution which are happening in the field of omics; in general of course, more in the interactomics and proteomics. And try to give you much more sense about what is exactly happening in this whole field, which is really remarkable and revolutionary in nature.

Thank you.