

Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – S15

Topics in Preteogenomics: Cancer case study

Welcome to MOOC course on Introduction to Proteogenomics. In today's lecture we are going to hear about a case study relevant for cancer proteogenomics. We are going to hear from Mr. Deeptarup Biswas about how cancer proteogenomics research could be helpful to provide some novel insights from the literature reviews some published datasets. His research is called in proteomics lab at IIT Bombay and he will talk about how proteogenomics approaches can help in resolving various issues of diagnosing various grades of cancer or looking at different subtypes of cancer, which is very difficult to understand without having a very good molecular base understanding.

He will also explain how proteomics and genomics data correlation can provide a much broader and meaningful picture of progression of cancer. He will try to also provide you the workflows of some of the case studies published in the areas of cancer proteogenomics. So, let me welcome Deeptarup for his today's lecture.

Welcome participants, till now you have learned a lot about proteomics and genomics, how to design an experiment, how to what are the condition that to be taken into account, but whether to consider proteomics or whether to consider genomics. Already a number of debates are going on and you have also heard that whether proteomics is powerful or genomics. To support this hypothesis of proteogenomics I want to give you a glimpse of how the powerful tool of proteogenomic can be, you can be used in cancer diagnosis and treatment.

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 and can lead to a path to personalized medicine.

(Refer Slide Time: 02:35)

Success of Personalized Medicine

Many patients do not respond to the predicted therapies based on the genomic profiles of their tumours

MGOC-NPTEL IIT Bombay

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: 02:59)

OPINION

Clinical potential of mass spectrometry-based proteogenomics

Bing Zhang, Jeffrey R. Whiteaker, Andrew N. Hoofnagle, Geoffrey S. Baird, Karin D. Rodland and Amanda G. Paulovich

Abstract | Cancer genomics research aims to advance personalized oncology by finding and targeting specific genetic alterations associated with cancers. In genome-driven oncology, treatments are selected for individual patients on the basis of the findings of tumour genome sequencing. This personalized approach has prolonged the survival of subsets of patients with cancer. However, many patients do not respond to the predicted therapies based on the genomic profiles of their tumours. Furthermore, studies pairing genomic and proteomic analyses of samples from the same tumours have shown that the proteome contains novel information that cannot be discerned through genomic analysis alone. This observation has led to the concept of proteogenomics, in which both types of data are leveraged for a more complete view of tumour biology that might enable patients to be more successfully matched to effective treatments than they would using genomics alone. In this Perspective, we discuss the added value of proteogenomics over the current genome-driven approach to the clinical characterization of cancers and summarize current efforts to incorporate targeted proteomic measurements based

(SRM/MRM) mass spectrometry (MS) into the clinical laboratory to facilitate clinical proteogenomics.

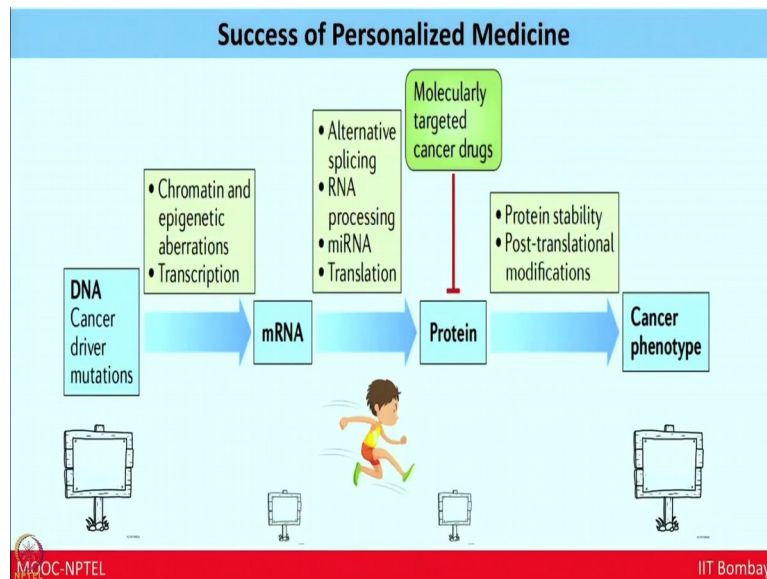
Connecting genotype to phenotype
On the basis of first principles, the observation that exclusive use of tumour genomic profiles is often insufficient to guide the reliable selection of targeted therapies should not be considered surprising. Many cellular processes downstream of the genome determine, or influence, which aspects of the cancer genome affect the phenotype of cancer cells (FIG. 1). For example, epigenetic changes are common in human cancers and affect the expression of critical cancer-related genes^{11–15}, such as oncogenes and tumour suppressors, and can also affect other regulatory elements such as microRNAs (miRNAs), with implications for cellular signalling and homeostasis¹⁶. Histone modifications have a role in alternative splicing¹⁷, which helps to drive hallmarks of cancer^{18,17}. Genomic, epigenomic and transcriptomic alterations all ultimately affect the activity of proteins expressed in tumours, which are also regulated by

MGOC-NPTEL IIT Bombay

So, recent paper published from Zhang Group there is a clinical potential of mass spectrometry based proteomics. 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 due to a number of reasons.

(Refer Slide Time: 03:23)

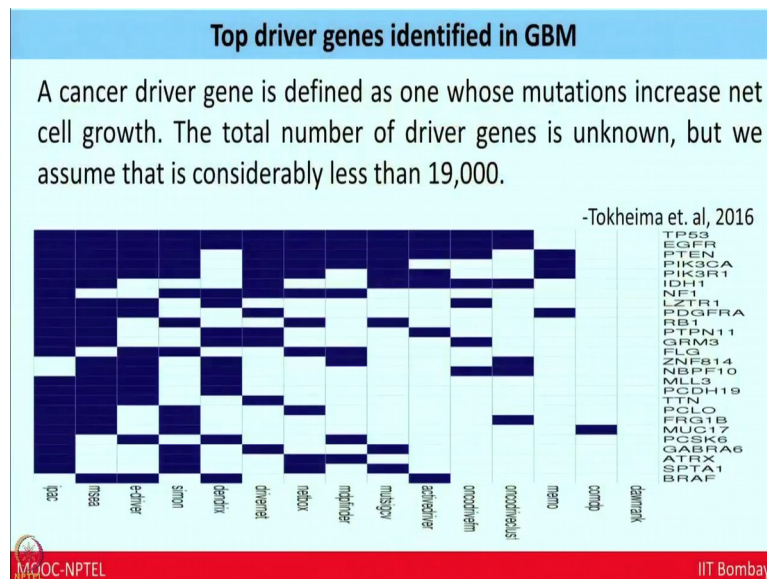


If we can see that with the help of genomics solving the problem like cancer is like jumping from one hurdle to the last hurdle and we are not taking into account a number of conditions and parameters that is coming in between the two hurdles.

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, transcriptional, regulations, alternative splicings and protein proteomics profiling. So, all this important information need to be taken 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. 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 and to bring a revolutionary change in the cancer and cancer diagnosis and treatment.

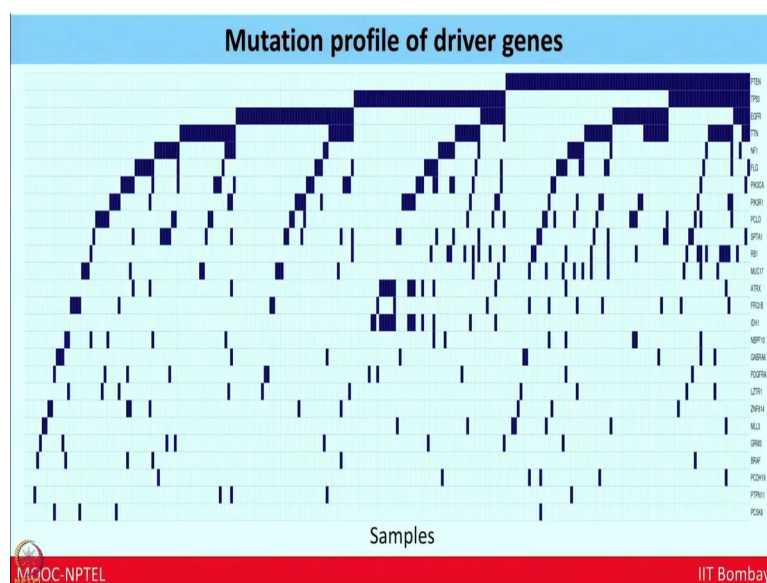
So, before I move how proteogenomics is playing a role in cancer diagnosis, I want to give a brief account of what is cancer driver genes.

(Refer Slide Time: 05:03)



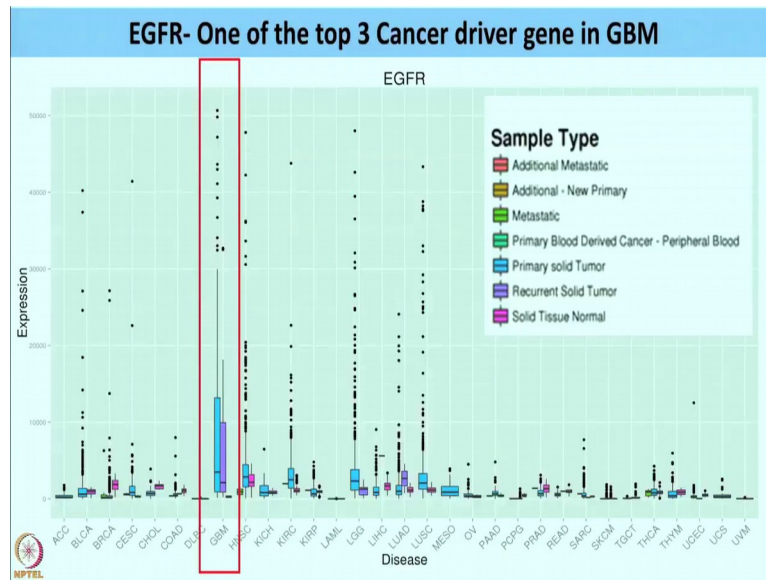
So, cancer driver gene is defined as one whose mutation increase net cell growth, the total number of driver gene is unknown, but we assume that is considerably less than 19000 which has been given by Tokheima et. al in 2016. So, from driver DV repository you can see like the top driver genes includes TP53, EGFR, PTEN, and how this hallmark driver genes are important in the glioblastoma in the glioblastoma tumorigenesis we all know.

(Refer Slide Time: 05:47)



So, here is the mutation profiles of those driver genes where the top driver genes are PTEN, TP53, EGFR and we can see the mutational profile in terms of samples which is in the x axis.

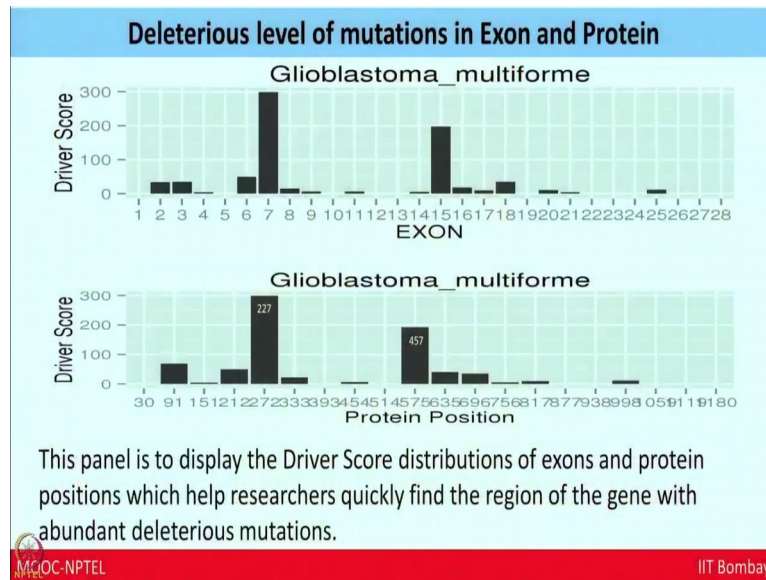
(Refer Slide Time: 05:59)



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 that GBM is having the most in case of GBM EGFR is highly is overexpressed in both primary solid tumor and recurrent solid tumor.

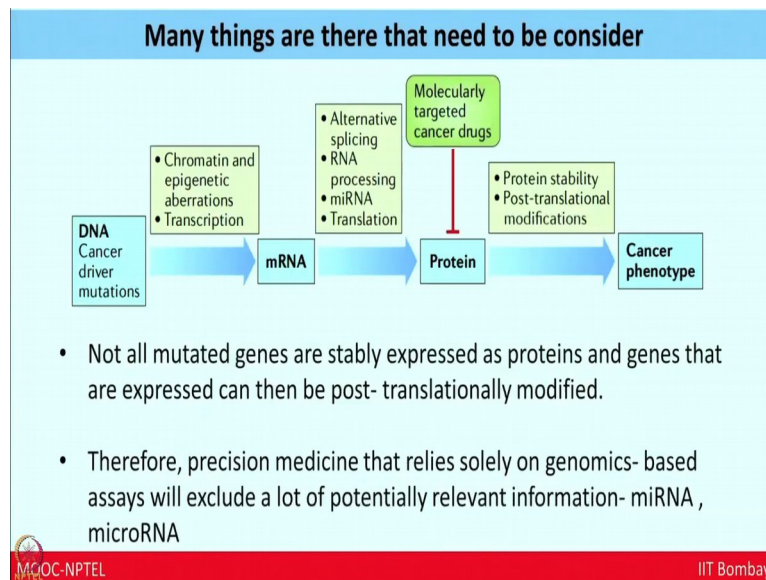
So, till now the genomics has given a lot of information about glioblastoma, but if we taken into account the correlation between the exon and protein we will found that the driver score related to protein and exon is also giving some new information.

(Refer Slide Time: 06:57)



This panel is to display that driver score distribution of exon and protein position which help researchers quickly find the region of the gene with abundant deleterious mutations.

(Refer Slide Time: 07:17)



So, now we understand that we are look, 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: 07:49)

Cell

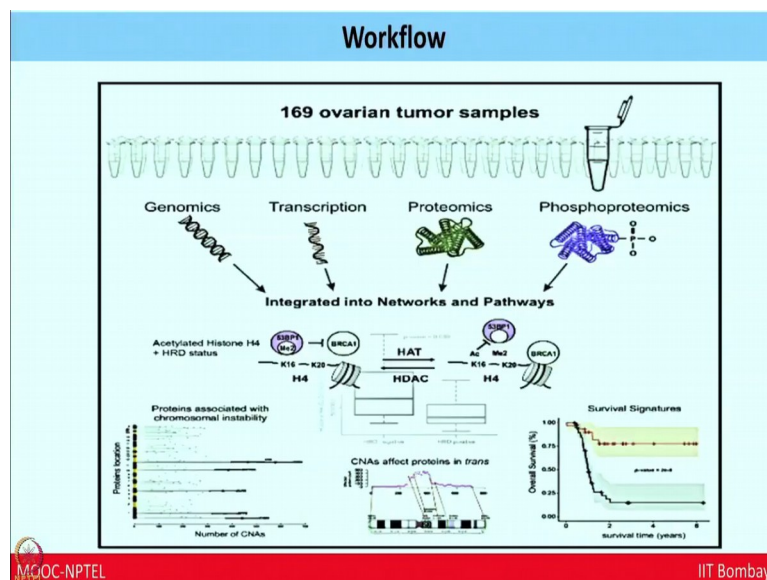
Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer

Hui Zhang,^{1,15} Tao Liu,^{2,15} Zhen Zhang,^{1,15} Samuel H. Payne,^{2,15} Bai Zhang,¹ Jason E. McDermott,² Jian-Ying Zhou,¹ Vladislav A. Petyuk,² Li Chen,¹ Debjit Ray,² Shisheng Sun,¹ Feng Yang,² Lijun Chen,¹ Jing Wang,³ Punit Shah,¹ Seong Won Cha,⁴ Paul Aiyetan,¹ Sunghee Woo,⁴ Yuan Tian,¹ Marina A. Gritsenko,² Therese R. Clauss,² Caitlin Choi,¹ Matthew E. Monroe,² Stefani Thomas,¹ Song Nie,² Chaochao Wu,² Ronald J. Moore,² Kun-Hsing Yu,^{5,6} David L. Tabb,³ David Fenyö,⁷ Vineet Bafna,⁸ Yue Wang,⁹ Henry Rodriguez,¹⁰ Emily S. Boja,¹⁰ Tara Hiltke,¹⁰ Robert C. Rivers,¹⁰ Lori Sokoll,¹ Heng Zhu,¹ Ie-Ming Shih,¹¹ Leslie Cope,¹² Akhilesh Pandey,¹³ Bing Zhang,³ Michael P. Snyder,⁵ Douglas A. Levine,¹⁴ Richard D. Smith,² Daniel W. Chan,^{1,15,*} Karin D. Rodland,^{2,15,*} and the CPTAC Investigators

MCOG-NPTEL IIT Bombay

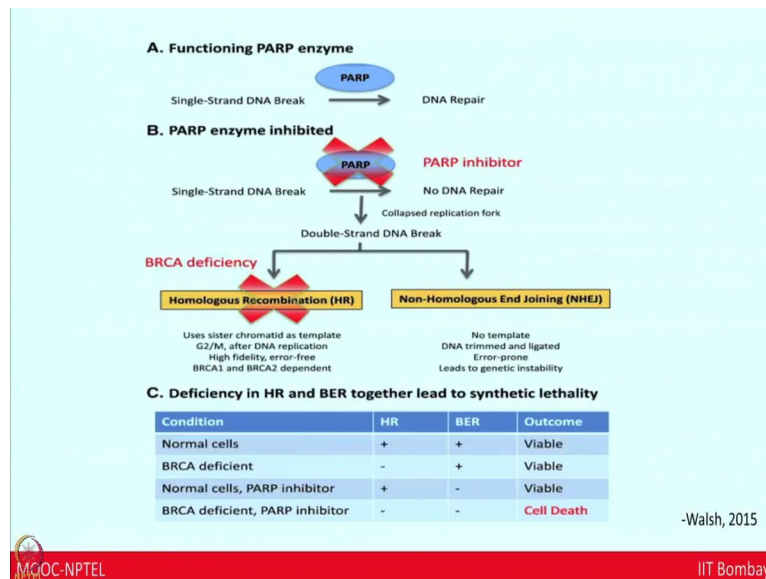
So, to support the previous statements and to give you a complete glimpse how the powerful tool of proteogenomics can be can be very helpful to solve different kinds of cancer.

(Refer Slide Time: 08:05)



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

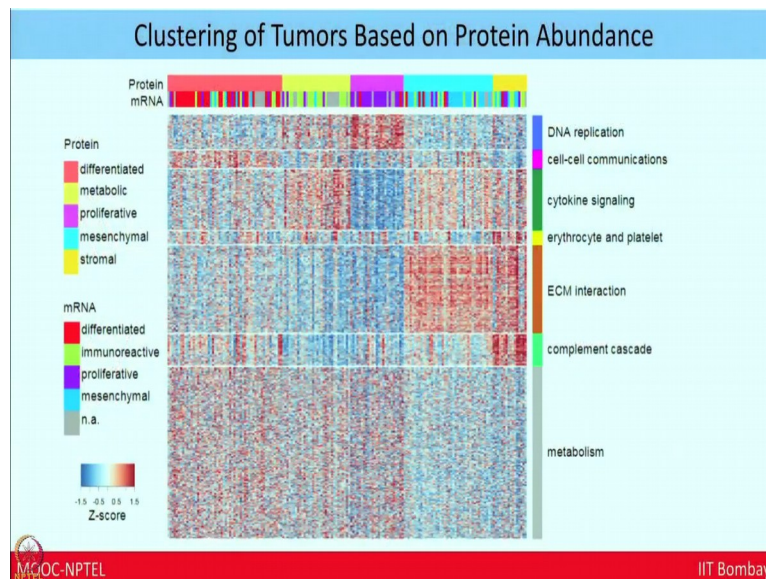
(Refer Slide Time: 08:21)



So, before going into the paper let me give a glimpse of this kind of mutation and how this mutation can be very can lead to lethality of a cell. So, the schematic I have to 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 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 HR, homologous recombination and base excision repair together lead to synthetic lethality than the correlation. So, the sample information tumors were selected by examining the associated TCGA Meta data to select tumors. On the basis of putative homologous recombination deficiency presence of germline or somatic BRCA1 or BRCA2 mutations, BRCA1 promoter methylation or homozygous deletion of PTEN were taken.

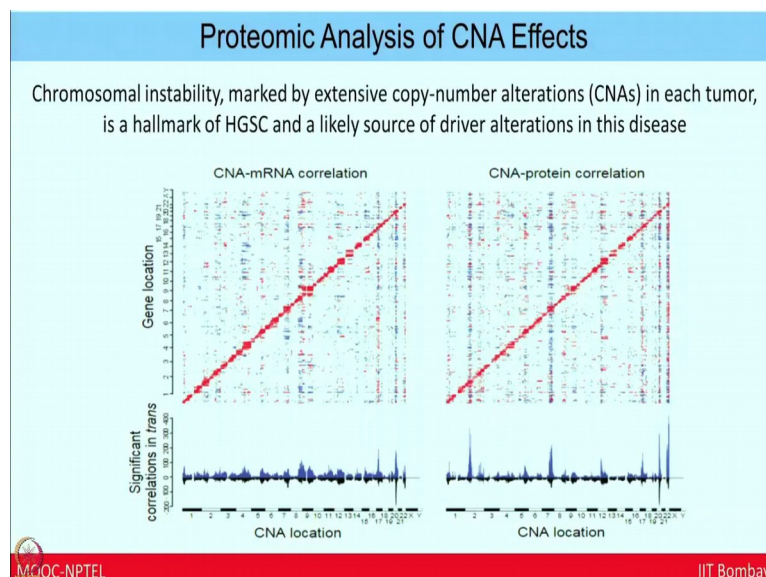
(Refer Slide Time: 09:41)



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

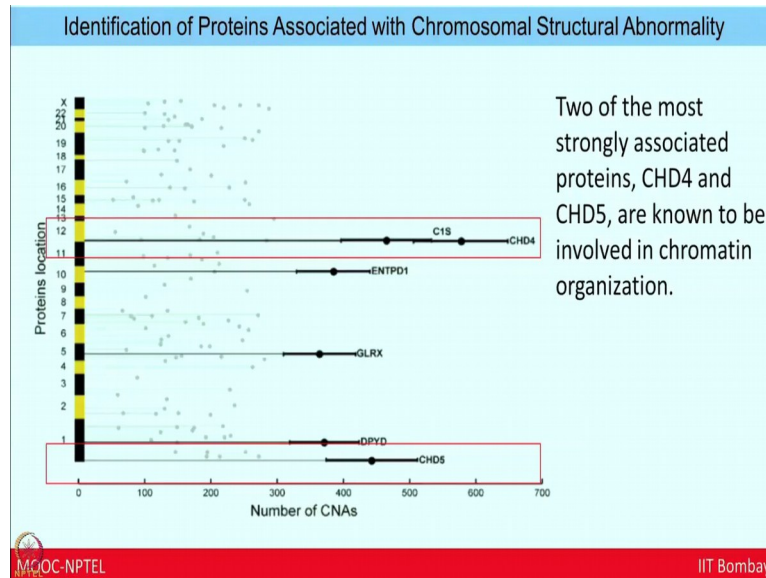
So, till now we understand that the protein and mRNA correlation is there and how this protein and mRNA correlation is also playing a role in terms of biological pathway, but now they also tried to understand that how CNA that is copy number aberration in each tumor is playing a role with protein and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

(Refer Slide Time: 10:23)



The blue one are the complete profile of the data generated where is the black one is the data that is present that is or that is already present in the database. So, from this CNA mRNA correlation and CNA protein correlation.

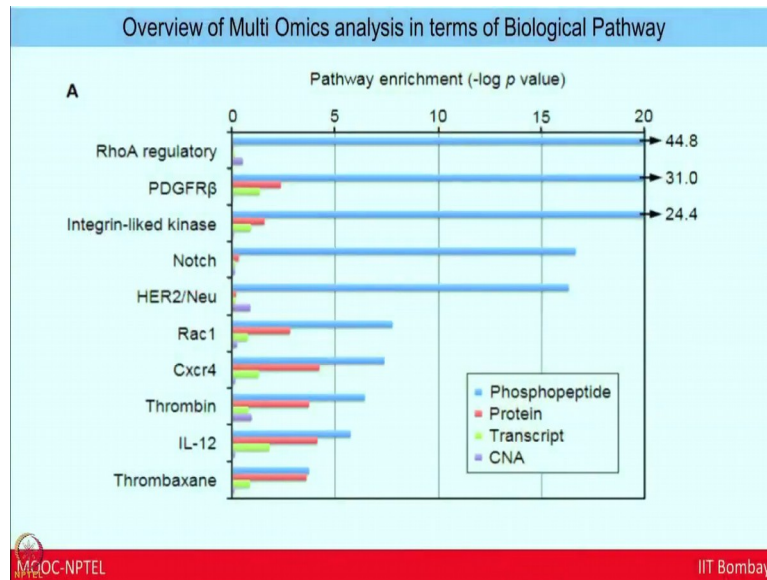
(Refer Slide Time: 10:37)



They found that two important two important protein that is CHD4 and CHD5 are having the maximum number of CNA CNAs. So, when the further studied they found that these two proteins are involved in chromatic organization.

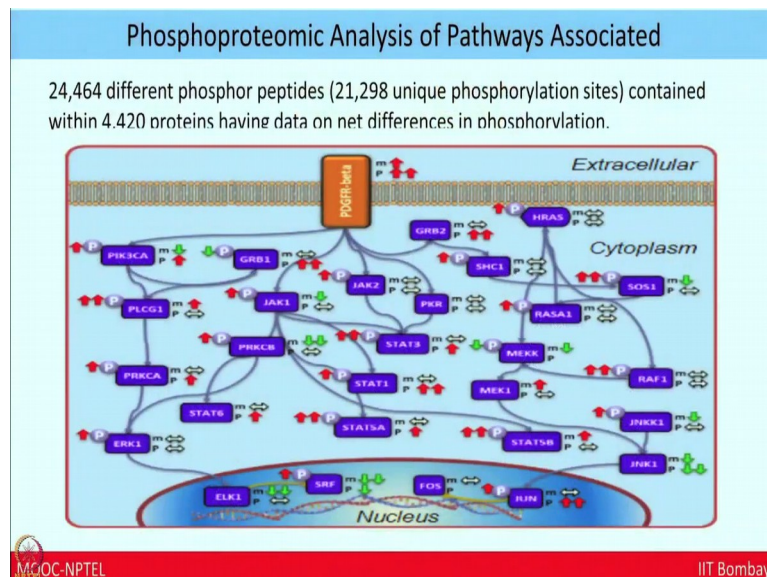
So, to understand the complete biological pathway they take they took phosphopeptides, proteins, transcripts, and CNA.

(Refer Slide Time: 11:05)



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, and phosphopeptide data into one picture and where we can see that the PDGFR beta is up regulated in both mRNA and protein.

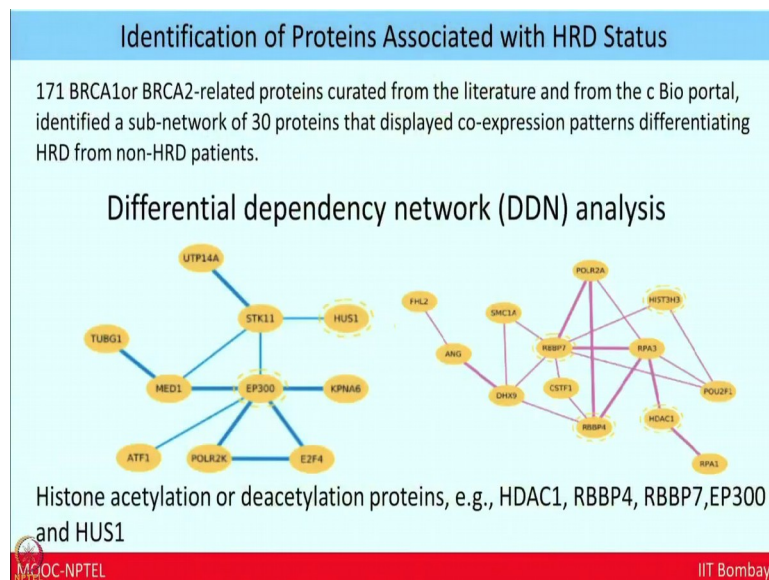
(Refer Slide Time: 11:31)



So, this up regulation of the PDGFR beta is not only giving a clue to our 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.

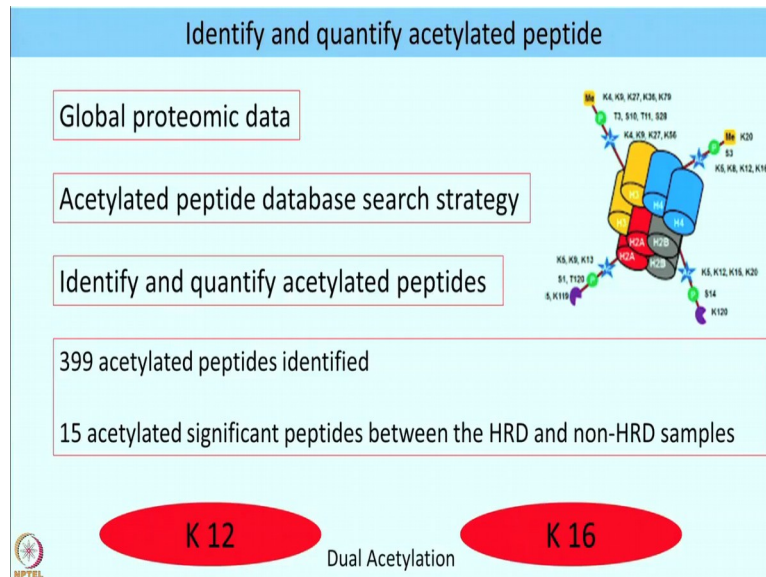
So, further they tried to do a DDN analysis. So, DDN analysis is differentially dependency network analysis where the proteins curated from the literature and from the c Bio portal.

(Refer Slide Time: 12:05)



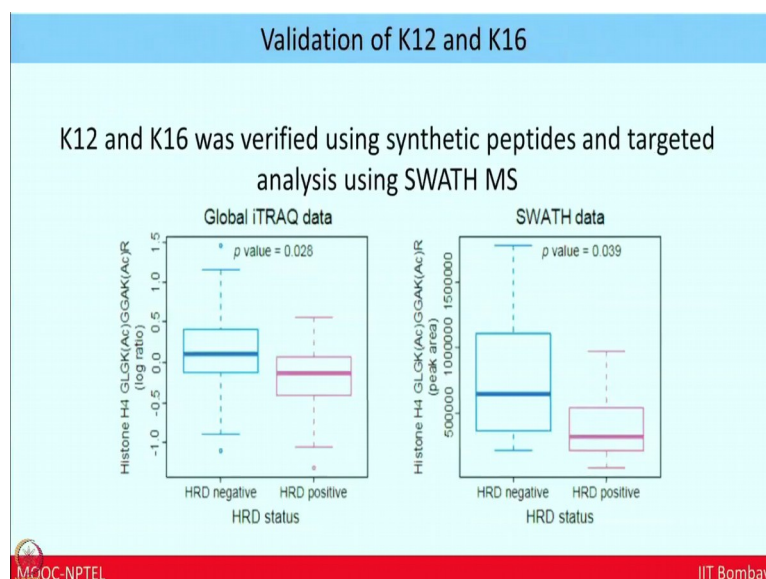
So, c Bio portal helps you to get the data out from the TCGA and they identified a sub network of 30 protein that displayed co-expression pattern differentiating from HRD from non-HRD patient and from these DDN analysis they found that histone acetylation or deacetylation proteins are coming are playing are coming into the clusters and which includes HDAC 1, RBBP4, 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 them give an idea that acetylated peptides need to be studies.

(Refer Slide Time: 13:03)



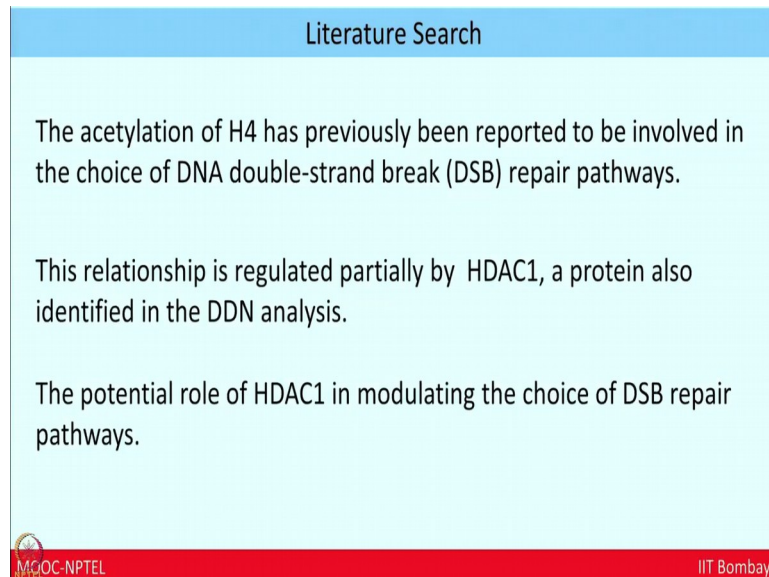
So, from the global proteome data they prepare acetylated peptide database, search strategy and identify and quantify the acetylated peptides. From there they identified around 399 acetylated peptides and 15 acetylated significant peptide between HRD and non-HRD. So, as so from this 15 acetylated significant peptide they found that K12 and K16 whether that is acetylation of lysine in 12 and 16 were found. So, they validated the K12 and K16 using synthetic peptide and targeted analysis using SWATH MS.

(Refer Slide Time: 13:35)



In the same thing they found that the K12 in terms of iTRAQ data were upregulated in HRD negative and same thing has been validated in SWATH and they found the same up regulation of same up regulation in HRD negative. So, they went back and further such in the literature and they found that the acetylation of the H4 has previously reported to be involved in the choice of DNA double brake double strand break DSB repair pathway.

(Refer Slide Time: 14:15)



Literature Search

The acetylation of H4 has previously been reported to be involved in the choice of DNA double-strand break (DSB) repair pathways.

This relationship is regulated partially by HDAC1, a protein also identified in the DDN analysis.


The potential role of HDAC1 in modulating the choice of DSB repair pathways.

MGC-NPTEL IIT Bombay

The relationship is regulated partially by HDAC1 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: 14:43)

Conclusion
<ul style="list-style-type: none">• The activation of PDGFR pathways in patient could potentially stratify patients for selective enrollment in trials of anti-angiogenic therapy.• Bevacuzimab- Bevacizumab is a recombinant humanized monoclonal antibody that blocks angiogenesis by inhibiting VEGF-A.• HRD- acetylation of K12 and K16 on histone H4, may provide an alternative biomarker of HRD• A rationale for these selection of patients in future clinical trials of HDAC inhibitors, alone or in combination with PARP inhibition.

 MOC-NPTEL IIT Bombay

So, the conclusion from the study we understand that the activation of PDGFR pathway in patient good portion potentially stratify selective enrollment in trial of anti-angiogenic therapy; recombinant human humanized monoclonal antibody Bevacizumab that blocks the angiogenesis by inhibiting VEGF-A has already been trialed in patients.

So, the PDGFR pathway the involvement of PDGFR pathway in this cancer is also giving this recombinant humanized monoclonal antibody role in limelight. Apart from this HRD acetylation K12 and K16 on histone H4 may provide an alternative biomarker of HRD. A rationale for this selection of patient in future clinical trials of HDAC inhibitors alone or in combination with PARP inhibition can be also tried.

(Refer Slide Time: 15:41)

Moral

The ability of proteomics to complement genomics in providing additional insights into pathway and processes that drive ovarian cancer biology .


MOC-NPTEL IIT Bombay

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. So, we understand that how 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 most 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.

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 already has only taken the has only generated the proteomics data and they have tried to correlate the their proteomics data with the already available genomics already available mRNA, CNA data from the databases. 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.



(Refer Slide Time: 17:03)

Introduction to FIREHOUSE

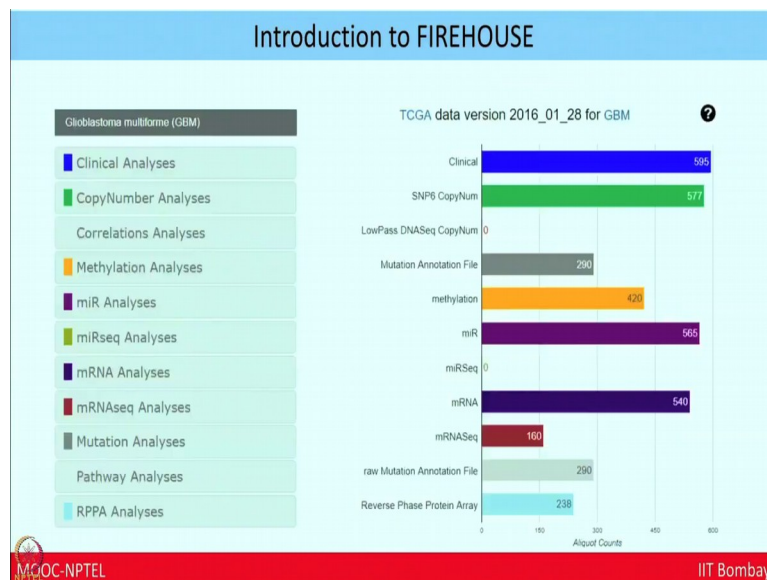


Dashboards Data Analyses Software Documentation FAQ Download

Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	Browser	Browser
Bladder urothelial carcinoma	BLCA	412	Browser	Browser
Breast invasive carcinoma	BRCA	1098	Browser	Browser
Cervical and endocervical cancers	CESC	307	Browser	Browser
Cholangiocarcinoma	CHOL	51	Browser	Browser
Colorectal adenocarcinoma	COAD	460	Browser	Browser
Colorectal adenocarcinoma	COADREAD	631	Browser	Browser
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	Browser	Browser
Esophageal carcinoma	ESCA	155	Browser	Browser
FPPE Pilot Phase II	FPPE	38	None	Browser
Glioblastoma multiforme	GBM	613	Browser	Browser
Glioma	GBMLGG	1129	Browser	Browser
Head and Neck squamous cell carcinoma	HNSC	528	Browser	Browser
Kidney Chromophobe	KICH	113	Browser	Browser
Pan-kidney cohort (KICH - KIRC - KIRP)	KIPAN	973	Browser	Browser
Kidney renal clear cell carcinoma	KIRC	537	Browser	Browser
Kidney renal papillary cell carcinoma	KIRP	323	Browser	Browser
Acute Myeloid Leukemia	LAML	200	Browser	Browser
Brain Lower Grade Glioma	LGG	315	Browser	Browser
Liver hepatocellular carcinoma	LIHC	377	Browser	Browser
Lung adenocarcinoma	LUAD	585	Browser	Browser
Lung squamous cell carcinoma	LUSC	504	Browser	Browser
Mesothelioma	MESO	87	Browser	Browser
Ovarian serous cystadenocarcinoma	OV	602	Browser	Browser
Pancreatic adenocarcinoma	PAAD	185	Browser	Browser
Phaeochromocytoma and Paraganglioma	PCPG	179	Browser	Browser
Prostate adenocarcinoma	PRAD	499	Browser	Browser
Rectum adenocarcinoma	READ	171	Browser	Browser
Sarcoma	SARC	251	Browser	Browser
Skin Cutaneous Melanoma	SKCM	470	Browser	Browser
Stomach adenocarcinoma	STAD	443	Browser	Browser
Stomach and Esophageal carcinoma	STES	628	Browser	Browser
Testicular Germ Cell Tumors	TGCT	150	Browser	Browser
Thyroid carcinoma	THCA	503	Browser	Browser

(Refer Slide Time: 17:07)




So, TCGA data version from 2016 from glioblastoma clinical, SNPs, methylation, miR mRNA and mRNA sequencing data and reverse phase protein array datas 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.

(Refer Slide Time: 18:05)

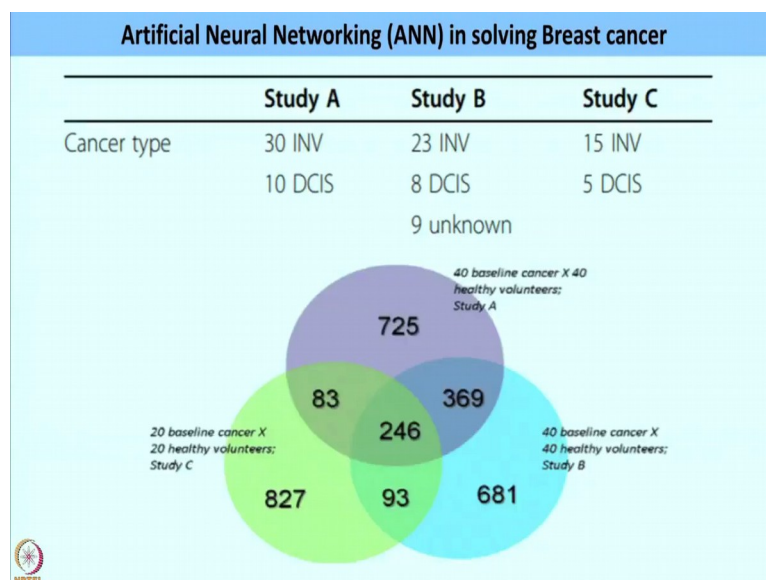
Artificial Neural Networking (ANN) in solving Breast cancer

A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer

Fan Zhang^{1,2*}, Jake Chen^{3,4,5,7}, Mu Wang^{5,6}, Renee Drabier^{1*}

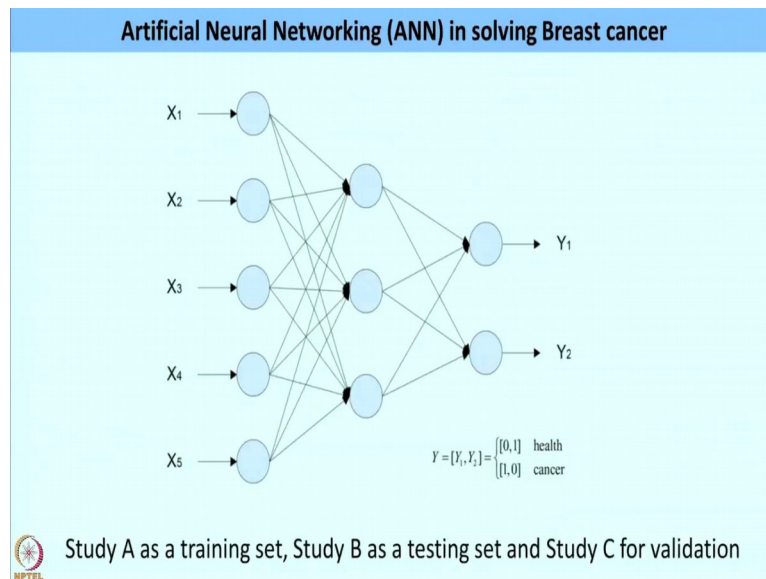


(Refer Slide Time: 18:15)



I have taken an example of a paper a neural network approach to multi-biomarker panel discovery by high throughput plasma proteomics profiling of breast cancer 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 they have done the proteomic analysis and they found that 246 proteins are common between 3 studies.

(Refer Slide Time: 18:35)



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 has a testing set and study C for 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 given.

The model further validated with blind data set to check the accuracy 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. So, this artificial neural networking gives a panel base three panels with five-markers and with the accuracy more than 85 percent.

(Refer Slide Time: 19:21)

Artificial Neural Networking (ANN) in solving Breast cancer				
Table 2 Best three five-marker panels identified				
Panel	SSE1	Accuracy		
		Training Set	Testing Set	Validation Set
C4BPA; HP; ORM1; SAMD9; SRCRB4D	3.3E-2	0.875	0.85	0.85
C4BPA; STBD1; DDX24; GRASP; CFI	5.6E-2	0.875	0.8375	0.85
C4BPA; CNO; FGG; SERPING1; SRCRB4D	1.9E-2	0.8625	0.85	0.85

So, further these panels were taken forward and checked in more checked in large cohort of samples to understand 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 up disease can be drawn and from there we can understand and we can and that can lead to a drug target or precision medicine 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: 20:27)

Points to Ponder

- Proteomics and Genomics can together give better information regarding disease pathobiology when integrated properly.
- Proteogenomics is a powerful tool which have the potential to bring revolutionary changes in the precision medicine.
- Prediction modeling and Machine learning can accelerate future cancer diagnosis.



MOOC-NPTEL

IIT Bombay

I hope from today's lecture you got a glimpse of what is happening in literature the most recent and very promising cancer proteogenomic studies, especially the CPTAC National Cancer Institute based work and those publications have made huge impact and brought the confidence about using genomic proteomic tools together and trying to provide the novel insights in looking at different type of cancers. We also got a glimpse of the workflows involved in doing these experiments which I think can provide you in a good way of thinking how you can try to utilize these tools of genomics and proteomics in your own research and then try to correlate the data analyze the data and bring something very novel which may not be possible otherwise. So, let me thank Deeptarup again for today's lecture and we will continue more interesting sessions in the next lecture.

Thank you.