INDIAN INSTITUTE OF TECHNOLOGY

NATIONAL PROGRAMME ON TECHNOLOGY
ENHANCED LEARNING

(NPTEL)

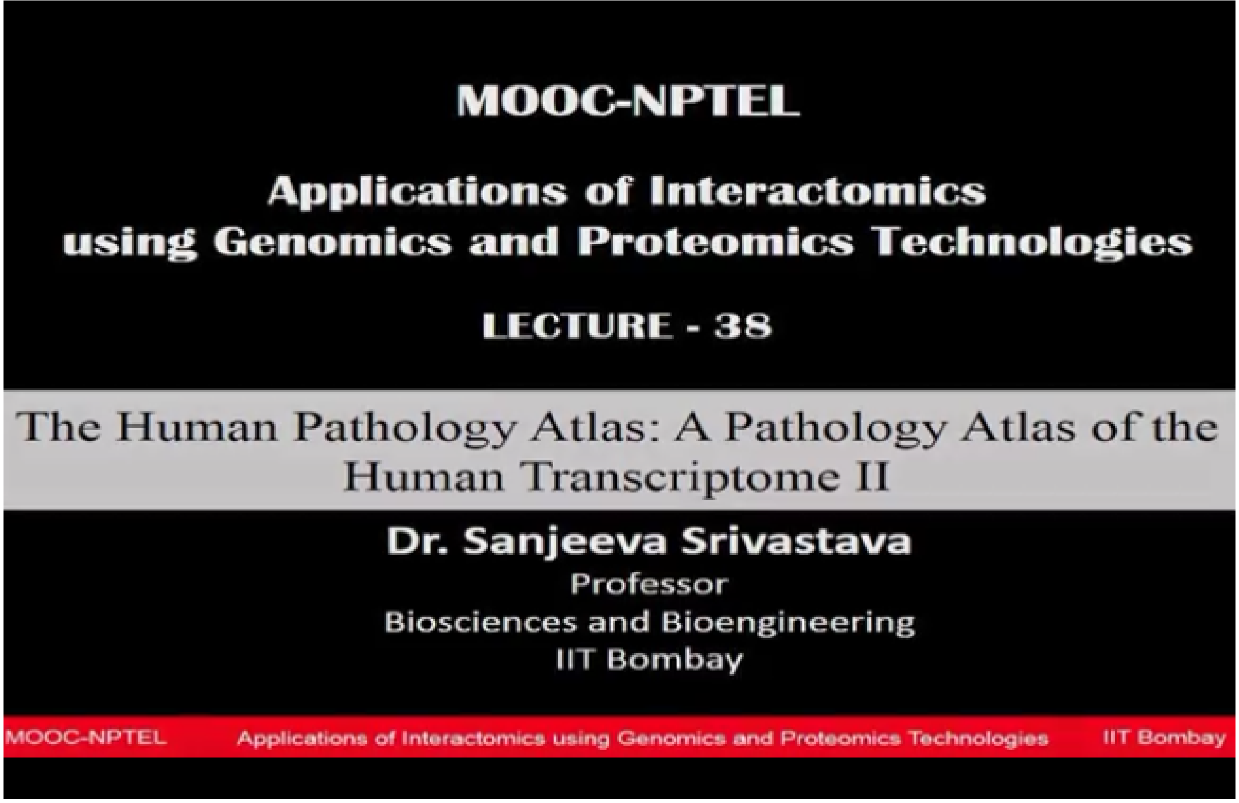
C-DEEP
IIT BOMBAY

Applications of Interactomics using
Genomics and Proteomics technologies

Course Introduction by
Prof. Sanjeeva Srivastava
Biosciences and Bioengineering
IIT Bombay

Lecture – 38
The Human Pathology Atlas: A Pathology Atlas of the Human Transcriptome-II

(Refer Slide Time: 00:27)



MOOC-NPTEL

Applications of Interactomics using Genomics and Proteomics Technologies

LECTURE - 38

The Human Pathology Atlas: A Pathology Atlas of the Human Transcriptome II

Dr. Sanjeeva Srivastava
Professor
Biosciences and Bioengineering
IIT Bombay

MOOC-NPTEL Applications of Interactomics using Genomics and Proteomics Technologies IIT Bombay

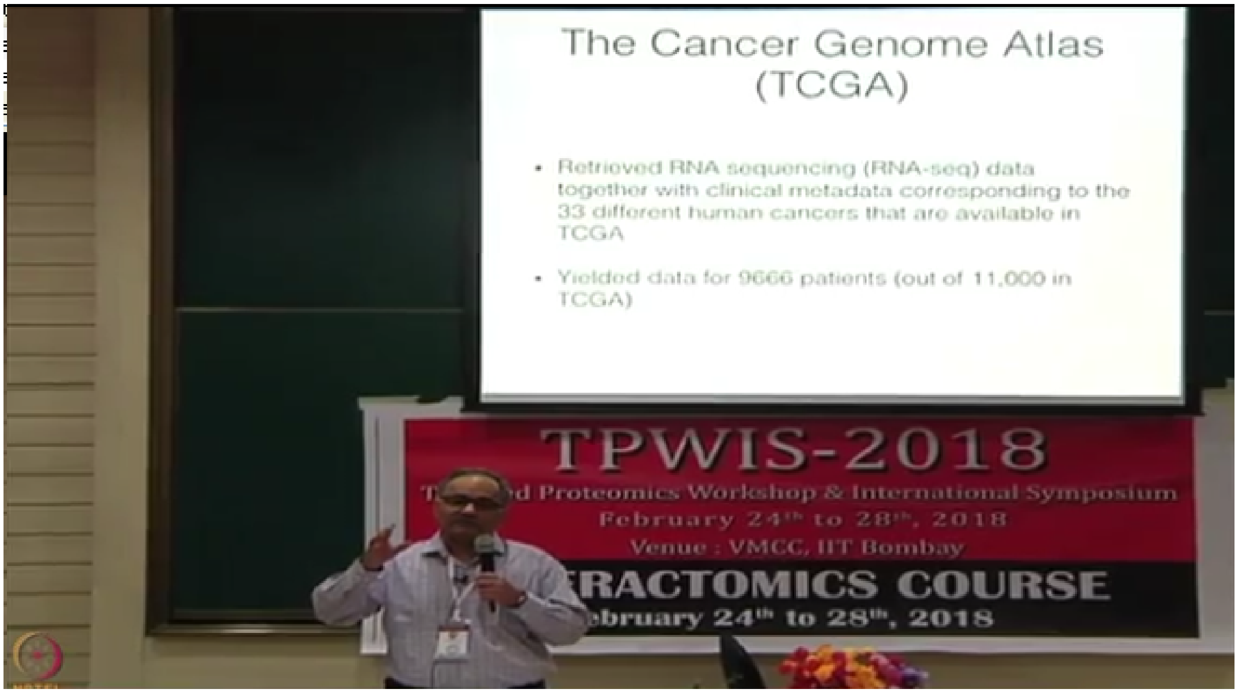
Welcome to MOOC course on Applications of Interactomics using Genomics and Proteomics Technologies. In last lecture you have to broad understanding of the mega project on human protein atlas, and also an overview of human pathology atlas project.

Today Dr. Sanjay Navani is going to continue his talk on human pathology atlas, you will learn more about the human pathology atlas which is one of the three aspects of the mega project on human protein atlas which also incorporates the tissue atlas and cell atlas, so let me welcome again Dr. Sanjay Navani for his lecture and continue discussion about human pathology atlas project.

(Refer Slide Time: 01:19)

Dr. Sanjay Navani

MBBS, MD, Laboratory Surgpath



The Cancer Genome Atlas (TCGA)

- Retrieved RNA sequencing (RNA-seq) data together with clinical metadata corresponding to the 33 different human cancers that are available in TCGA
- Yielded data for 9666 patients (out of 11,000 in TCGA)

TPWIS-2018
T... d Proteomics Workshop & International Symposium
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
...RACTOMICS COURSE
...ebruary 24th to 28th, 2018

Dr. Sanjay Navani: At this time of big data, there are lot of people protein atlas is one of them, there is a lot of big data, and the cancer genome atlas is another place with high amounts of data, can you combine, can you combine resources and produce a product that is beneficial or gives us a better answer, so it was with that thought that the RNA sequencing data with the clinical metadata which means the survival and how the patient did, was derived from the cancer genome atlas.

And we got data out of our total of 11,000 patients for about 9,600 patients which was the study pool, we looked at the global gene expression patterns for all the protein encoding genes, but here you have to be a bit careful, I'm using this words protein encoding genes, what I'm basically telling you is only one protein per gene, I'm not looking at post translational modifications, so you must remember that, that still another variable that may need to be crossed in the future.

The gene expression in 37 normal human tissues were obtained from a 162 patients from the HPA project, so the cancer RNA-Seq data with the clinical metadata from the cancer genome atlas and the normal tissues from the HPA.

Unidentified Speaker: So, how are you getting all these normal tissues? Is it of healthy individuals?

Dr. Sanjay Navani: From healthy people.

Unidentified Speaker: Yeah.

Dr. Sanjay Navani: Yeah, so I tried to define healthy, I mean one source was autopsies.

Unidentified Speaker: Okay.

Dr. Sanjay Navani: And the other source was people who had been biopsied, but who were non-cancer, that was the control that was used.

Unidentified Speaker: Okay.

Dr. Sanjay Navani: So every biopsy did not expected diagnosis of cancer they were some of those patients and some of them were done to rule out a cancer.

Unidentified Speaker: Right, right.

Dr. Sanjay Navani: So yeah, so therefore they were not entirely normal, but.

Unidentified Speaker: But there will be limited number of patients you might do a biopsy without a cancer, right.

Dr. Sanjay Navani: Yeah, well the sweets had it, they had a biobank, very well maintained biobank, I have to say, so it was a struggle and that's why…

Unidentified Speaker: They are introducing potential biases.

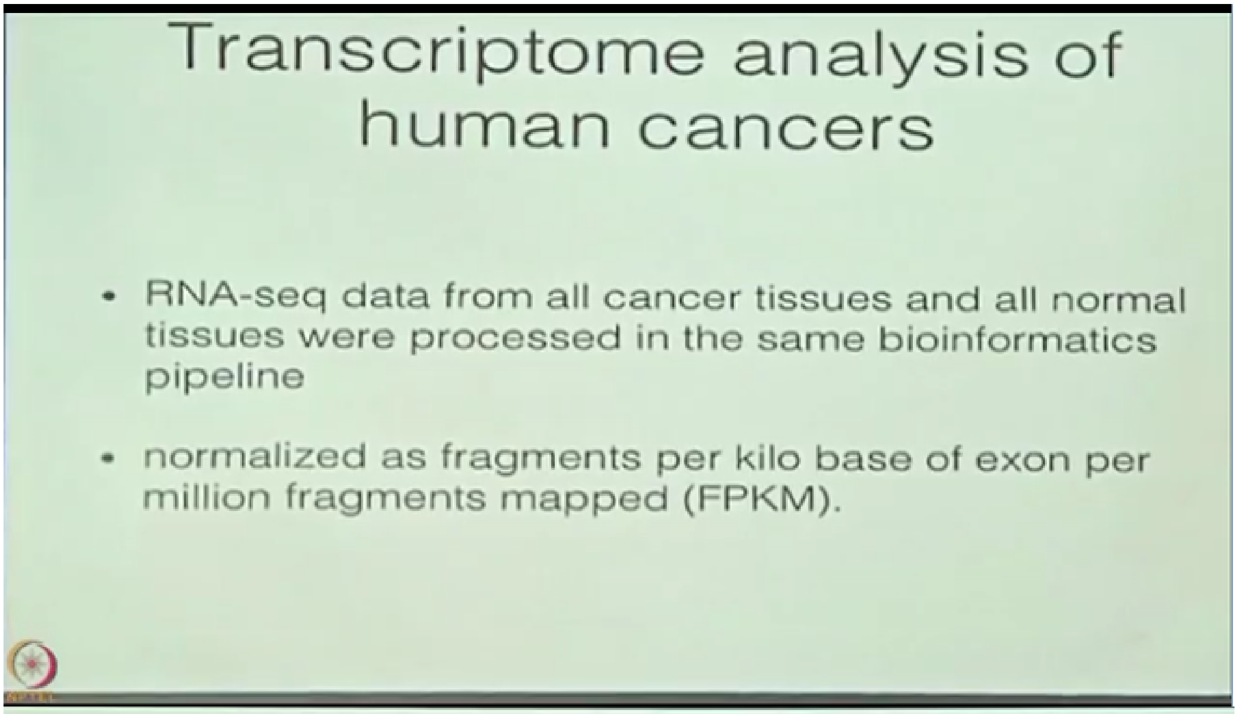
Dr. Sanjay Navani: Yeah, yeah, of course.

Unidentified Speaker: As compared to autopsy of a healthy person.

Dr. Sanjay Navani: Yeah and for sure, but it was when there was no pathologic issue seen in that tissue under the microscope, or correlated, I'm sure the patients, I'm sure must have had many other problems, so normal, I mean all of us have many problems.

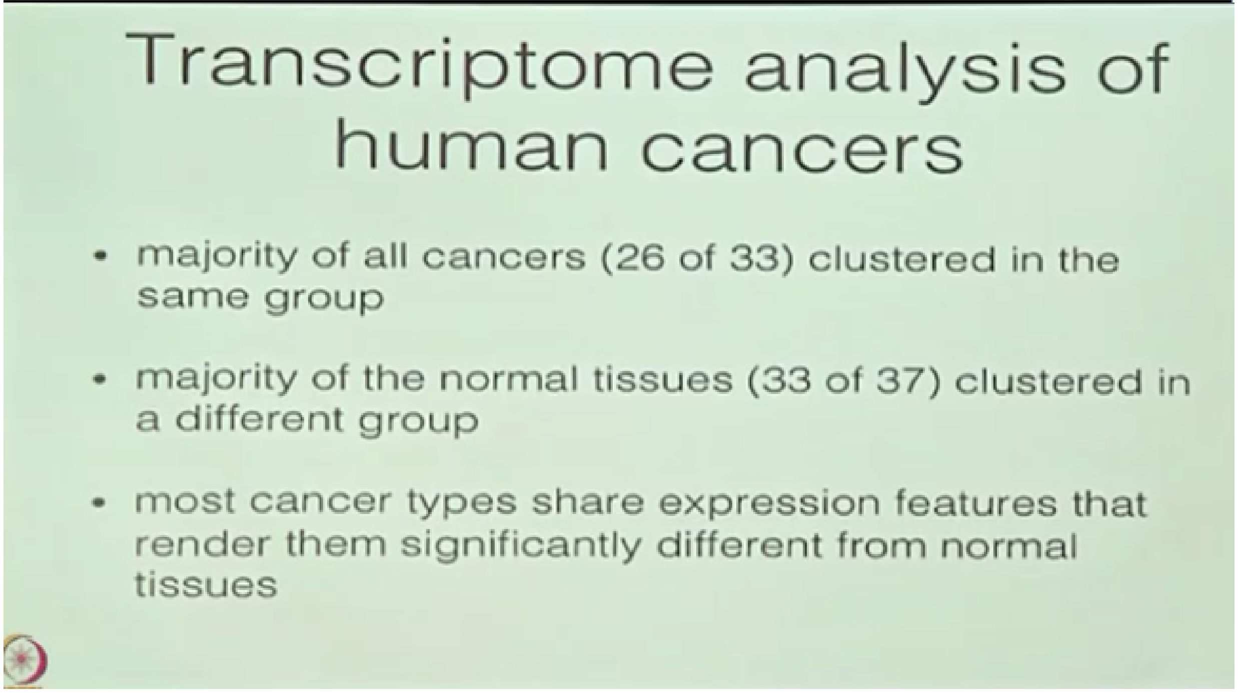
Unidentified Speaker: Yes, yes.

Dr. Sanjay Navani: So we are not normal too, so that was the best that was possible under that circumstance and all the RNA-Seq data both from the cancer as well as the normal tissues, they were processed in the same pipeline, and they were given normalized according to the FPKM, so that was how it was expressed.
(Refer Slide Time: 04:54)



Transcriptome analysis of human cancers

- RNA-seq data from all cancer tissues and all normal tissues were processed in the same bioinformatics pipeline
- normalized as fragments per kilo base of exon per million fragments mapped (FPKM).

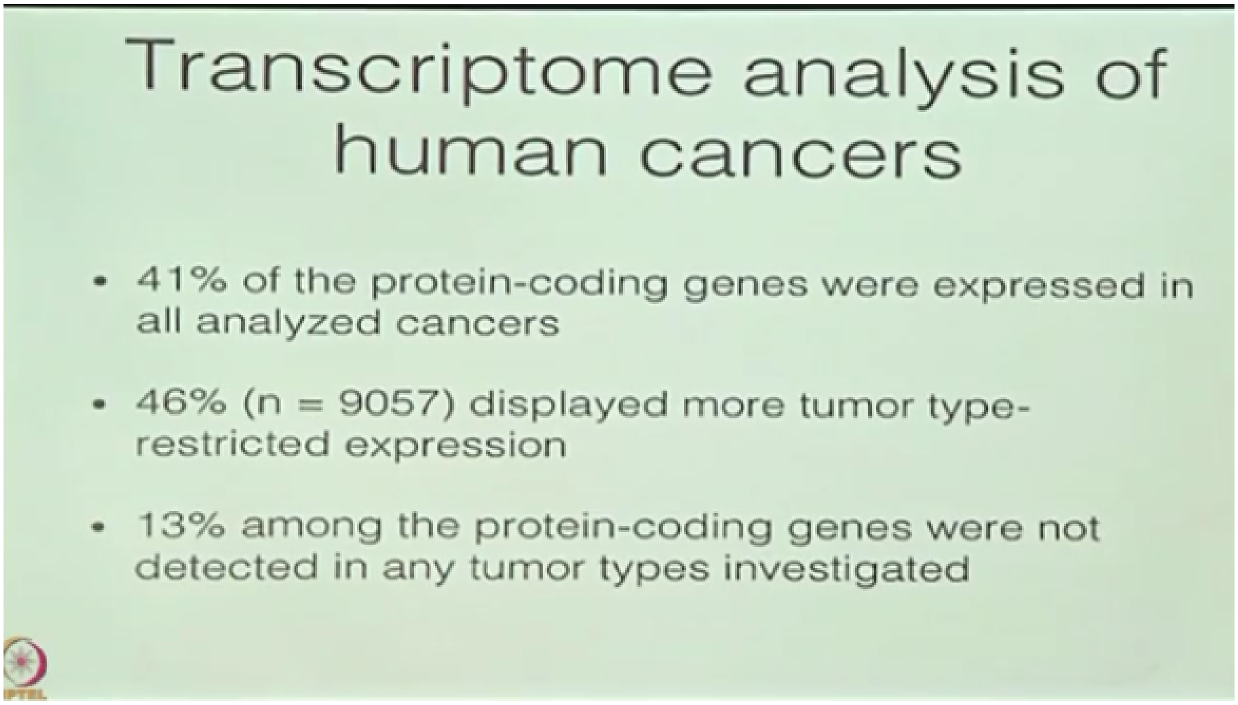


Transcriptome analysis of human cancers

- majority of all cancers (26 of 33) clustered in the same group
- majority of the normal tissues (33 of 37) clustered in a different group
- most cancer types share expression features that render them significantly different from normal tissues

When we looked at the data initially majority of all the cancers 26 out of 33 cancers clustered in the same group, majority of all normal tissues 33 out of 37 clustered in the different group, and the conclusion was of course the first basic conclusion was that most cancer types share expression features that make them quite different from normal tissues and that was what we expected.

(Refer Slide Time: 05:32)

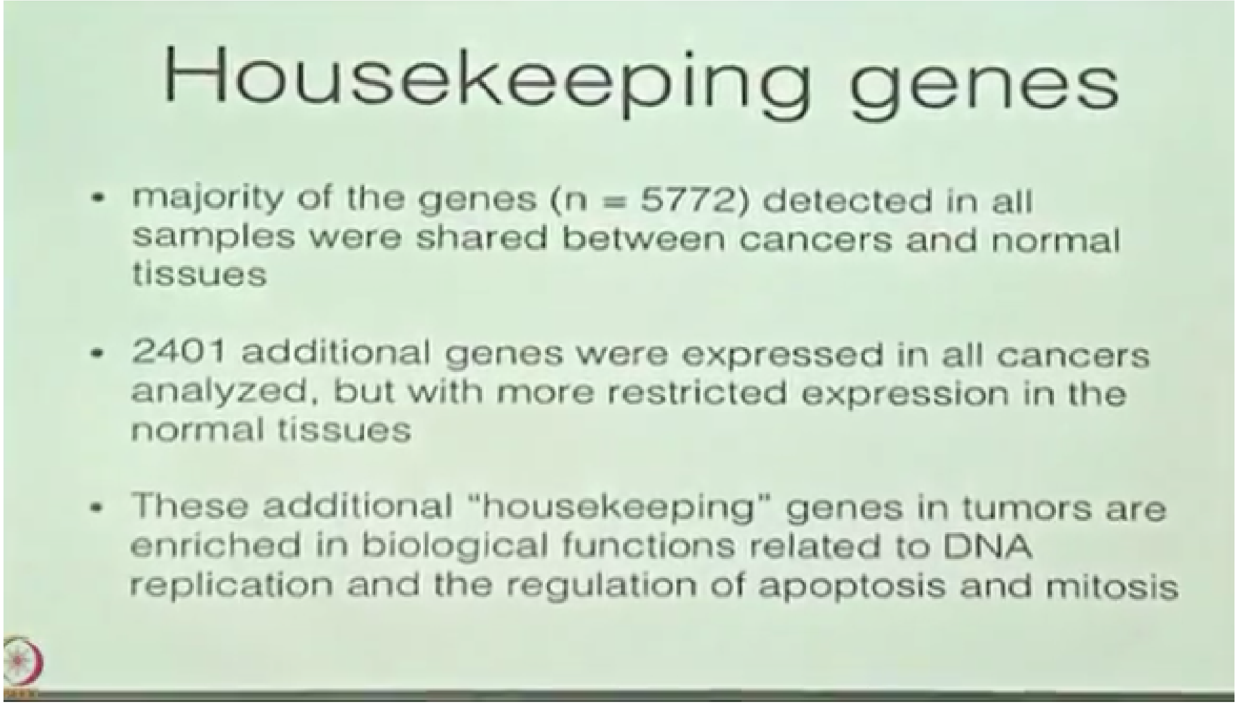


Transcriptome analysis of human cancers

- 41% of the protein-coding genes were expressed in all analyzed cancers
- 46% (n = 9057) displayed more tumor type-restricted expression
- 13% among the protein-coding genes were not detected in any tumor types investigated

Out of all the protein encoding genes 41% were present in all cancers, so a breast cancer was not necessarily that different from gallbladder cancer, there was a large overlap, secondly 46% and this I considered to be very important at this stage of our research activity, they were, they showed a restricted type tumor expression, so in the tumors they were different, but when we compared it to normal tissues those showed same genes were different, and 13% of the protein encoding genes were not seen at all, now that's a big question mark, where are those genes? We've probably read a bit about the missing genes and the missing proteins and stuff like that, some of them, there are only theories for that group, what are the proteins that those genes are coding for, are they important only before birth, but I won't get into that now.

The housekeeping genes were detected in all samples, both cancers as well as normal tissues,
(Refer Slide Time: 07:10)

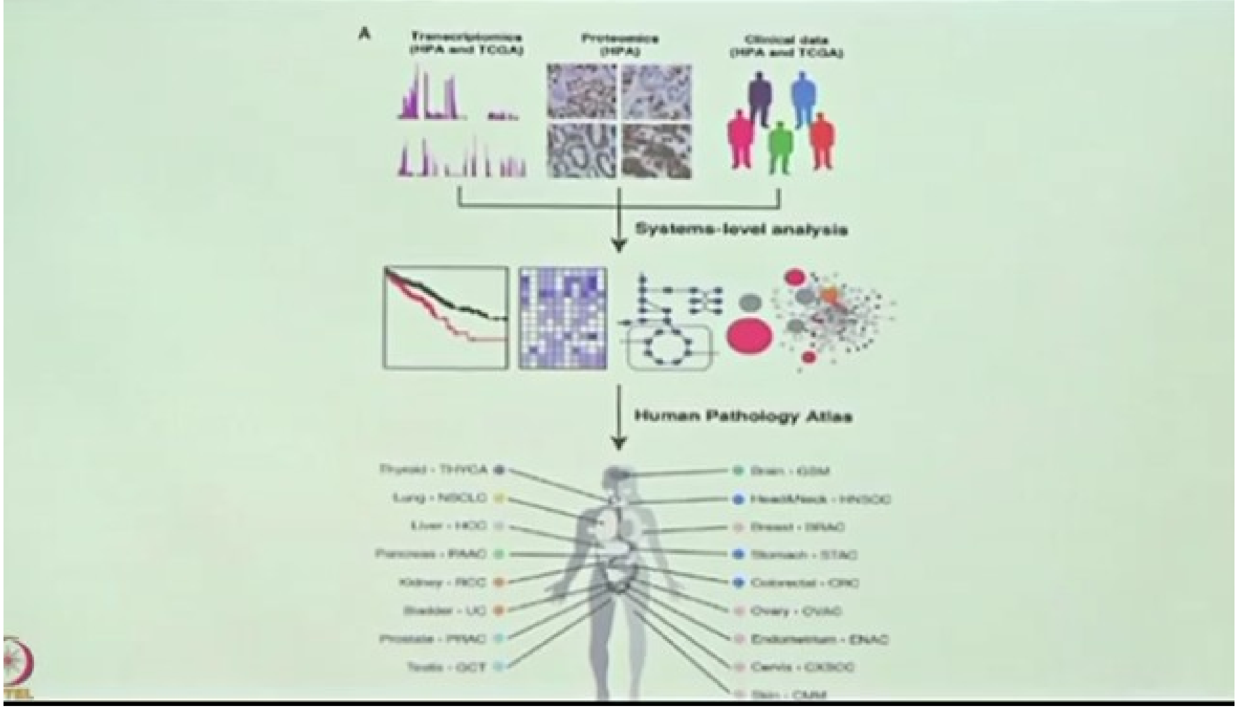


Housekeeping genes

- majority of the genes (n = 5772) detected in all samples were shared between cancers and normal tissues
- 2401 additional genes were expressed in all cancers analyzed, but with more restricted expression in the normal tissues
- These additional "housekeeping" genes in tumors are enriched in biological functions related to DNA replication and the regulation of apoptosis and mitosis

the housekeeping genes are very important because they are increased in cancers, because they do all the activities looking after every cell, they are the same in every cell, only because the cancer cell is multiplying very fast they are more in the cancer cell, but what we can't forget is they also present in normal tissues,

(Refer Slide Time: 07:32)

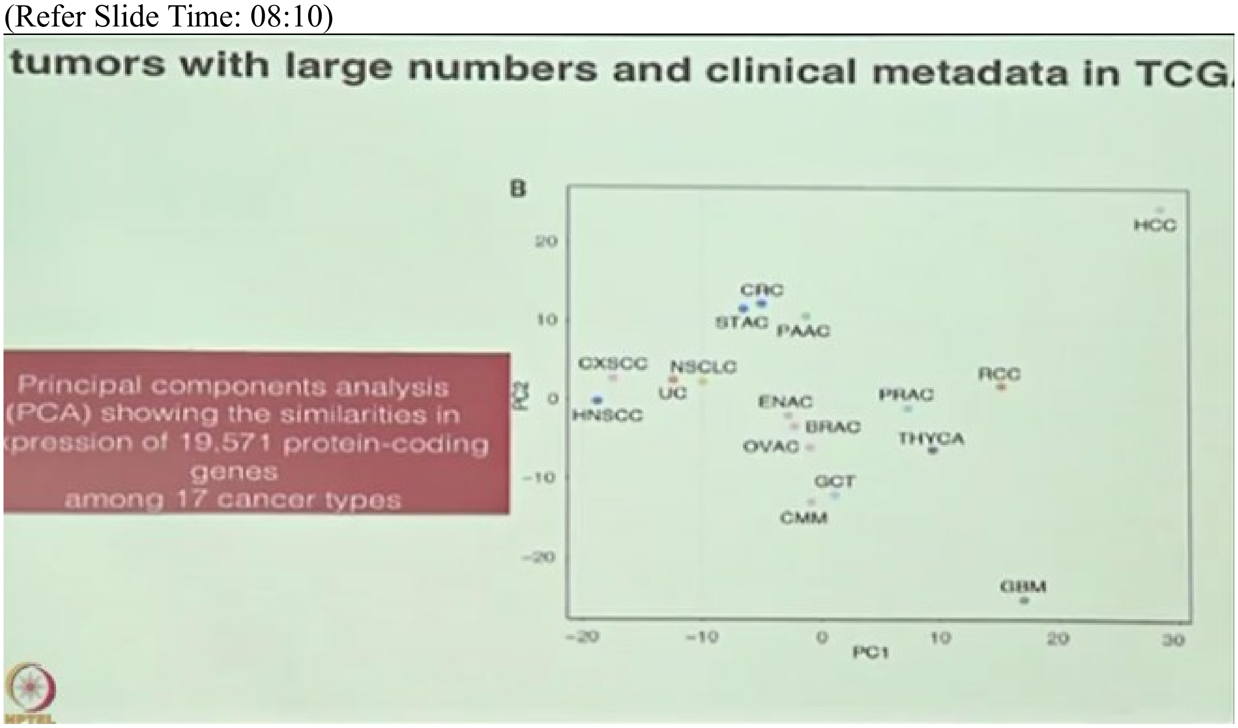


that was how the program worked out.

Unidentified Speaker: _7:37_

Dr. Sanjay Navani: So the transcriptomics was taken from the cancer genome atlas and the HPA, the HPA contains the immunohistochemistries stained images, and the clinical data was got both from HPA and the cancer genome atlas, a systems level analysis was done that gave rise to the human pathology atlas, so when look at this diagram on the website you can click any of these cancers and you will be taken to that data.

(Refer Slide Time: 08:10)



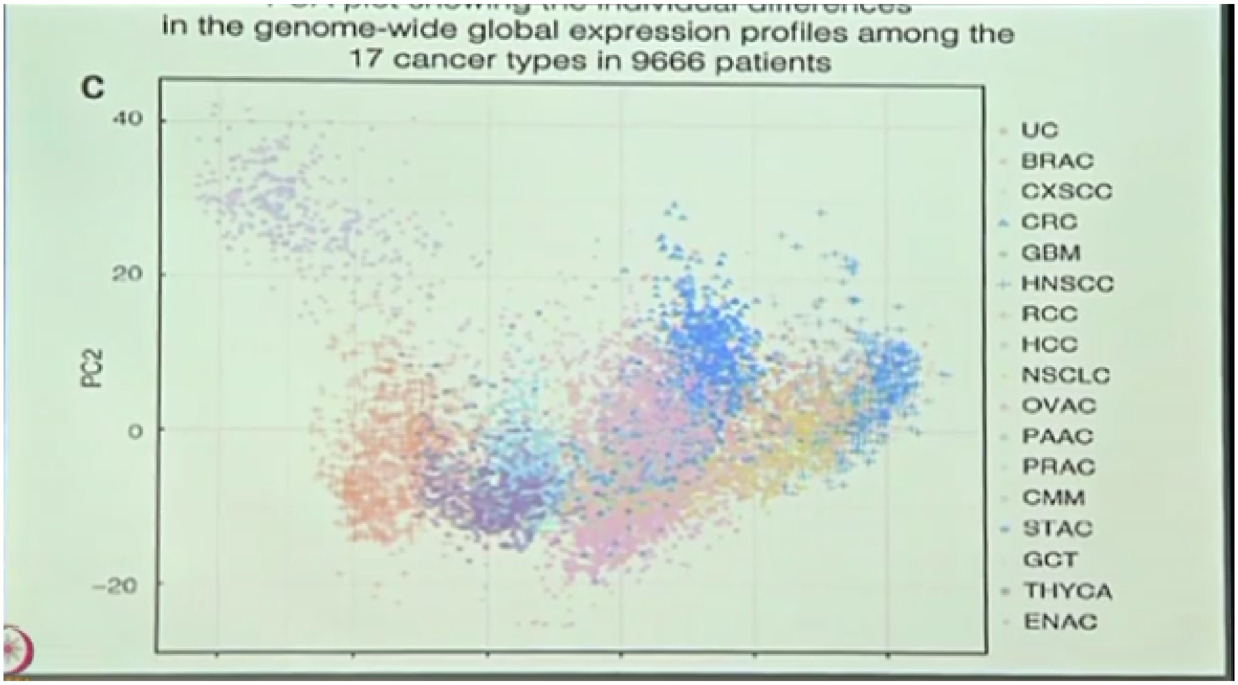
tumors with large numbers and clinical metadata in TCG

Principal components analysis (PCA) showing the similarities in xpression of 19,571 protein-coding genes among 17 cancer types

We then narrowed the search to 17 tumors with large numbers and clinical metadata, from the original 33 we came down, 37 we came down to 17 because these cancers had adequate numbers, and as you can see these cancers which are grouping here are gastrointestinal cancers, this is colorectal carcinoma that is stomach adenocarcinoma, that is pancreatic adenocarcinoma.

This group which is coming here are the squamous cell carcinomas, so it's head and neck squamous cell carcinoma, cervical squamous carcinoma, if you see this group here these are the endometrial adenocarcinomas, ovarian adenocarcinomas and the breast adenocarcinomas, so they group together from the same system, they were only two very far outliers, one was hepatocellular carcinoma very different from everything else and glioblastoma multiforme which is a very brain tumor, which look quite different from everything else, the surprises were yet to come.

(Refer Slide Time: 09:26)

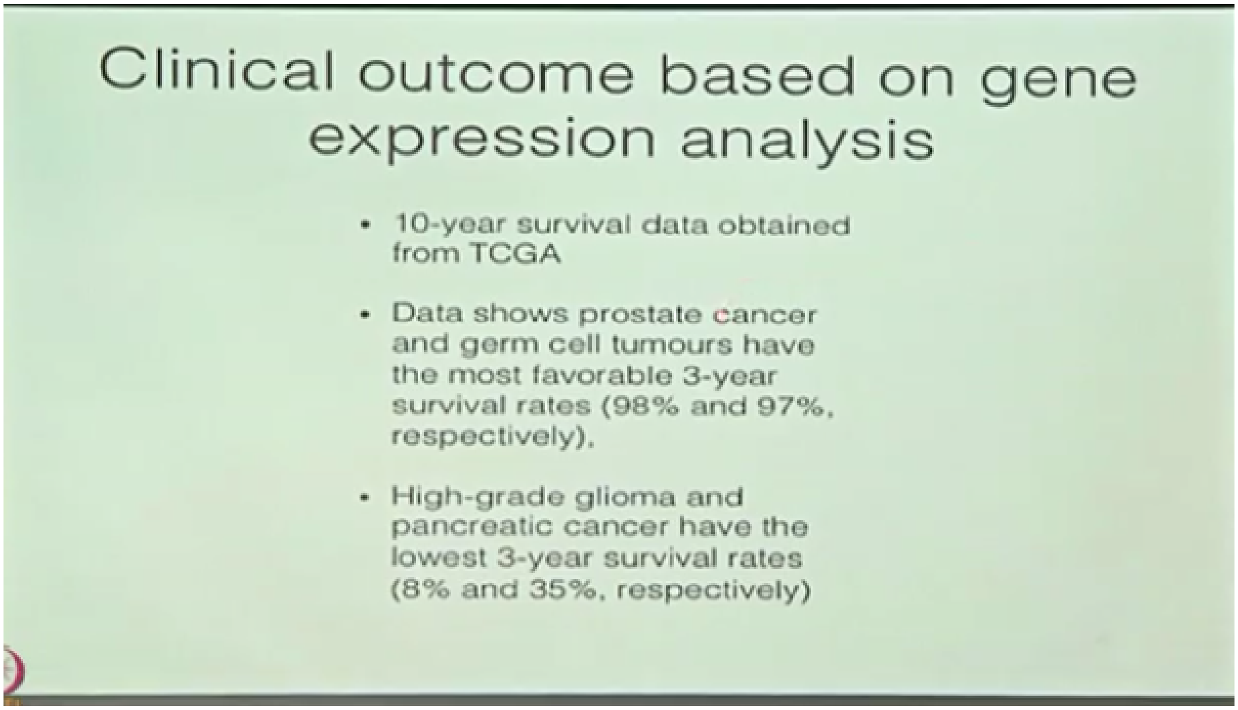


PCA plot showing the individual differences in the genome-wide global expression profiles among the 17 cancer types in 9666 patients

When we look at this each colour and each figure corresponds to a particular type of cancer, you can see how much variation there is, in each individual cancer that means even though people like me say this is a moderately differentiated or a great 2 hepatocellular carcinoma and you take two of those they look different.

Not only that you will see that there is also a spill over into other side, cancers of other types, in fact if you go only by the transcriptomics profile you will say that this resembles, this cancer more than it does it parent cancer, which raises a question in my mind which is so far I haven’t been able to answer, if that’s the case then why does it look like that?

(Refer Slide Time: 10:27)



Clinical outcome based on gene expression analysis

- 10-year survival data obtained from TCGA
- Data shows prostate cancer and germ cell tumours have the most favorable 3-year survival rates (98% and 97%, respectively).
- High-grade glioma and pancreatic cancer have the lowest 3-year survival rates (8% and 35%, respectively)

So the 10 year survival data also was available on cancer genome atlas, it shows prostate cancer and germ cell tumors to have the most favorable 3 years survivals, that's there if you have to have cancer those are the good cancers to get, you will do well with them because of therapies available, and…

Unidentified Speaker: It's a little so the TCGA specifically for some people with else rare disease, right, so I think that's a Tuesday that for people with that disease.

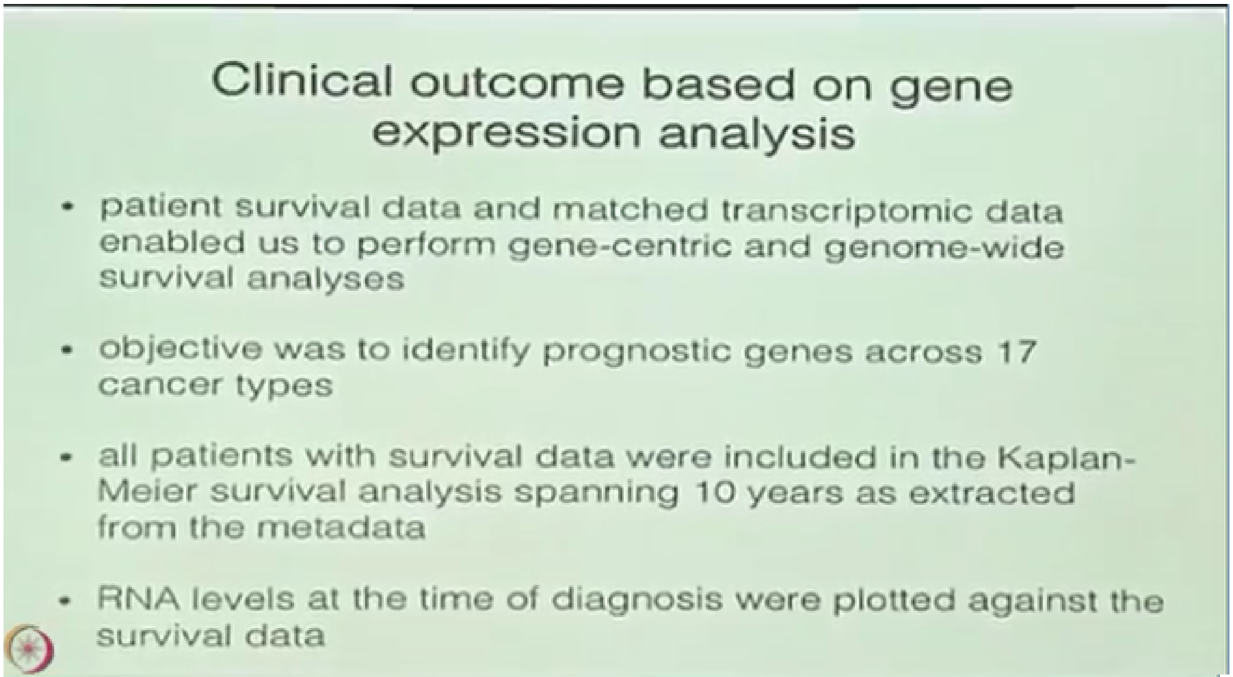
Dr. Sanjay Navani: Yeah

Unidentified Speaker: And why not _11:00_

Dr. Sanjay Navani: Yeah, I don't know whether it's only end stage disease because the RNA-Seq data on the genome atlas is obtained at the point of diagnosis, and they follow-up to the event of death which is what interested us, but this is where your point is very valid, they don't say that he died of prostate cancer, they just say that he died, now it may have been a myocardial infarction, but because we had an end point of death, and we had the RNA-Seq levels at the start, that was the reason for using this data.

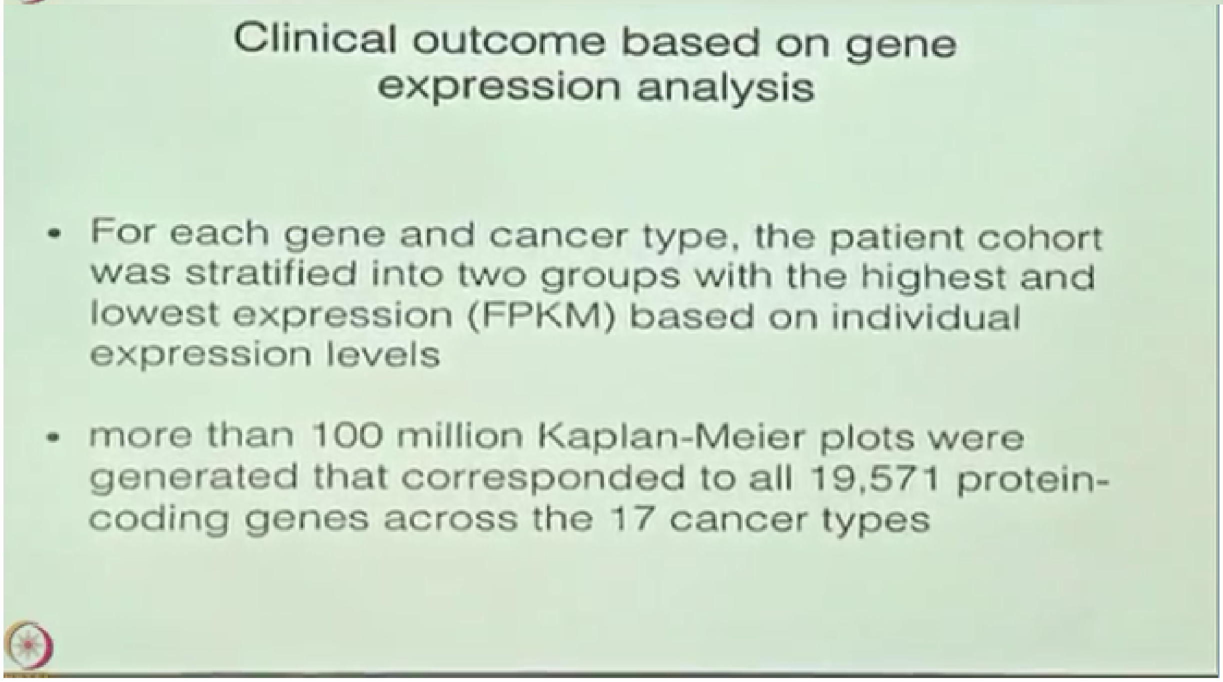
Okay, so what we did was actually a huge exercise of Kaplan Meier curves which has look for survival for each gene, for RNA-Seq data from each gene, and the RNA levels at the time of diagnosis were plotted against the survival data, that's something I was just saying.

(Refer Slide Time: 12:21)



Clinical outcome based on gene expression analysis

- patient survival data and matched transcriptomic data enabled us to perform gene-centric and genome-wide survival analyses
- objective was to identify prognostic genes across 17 cancer types
- all patients with survival data were included in the Kaplan-Meier survival analysis spanning 10 years as extracted from the metadata
- RNA levels at the time of diagnosis were plotted against the survival data

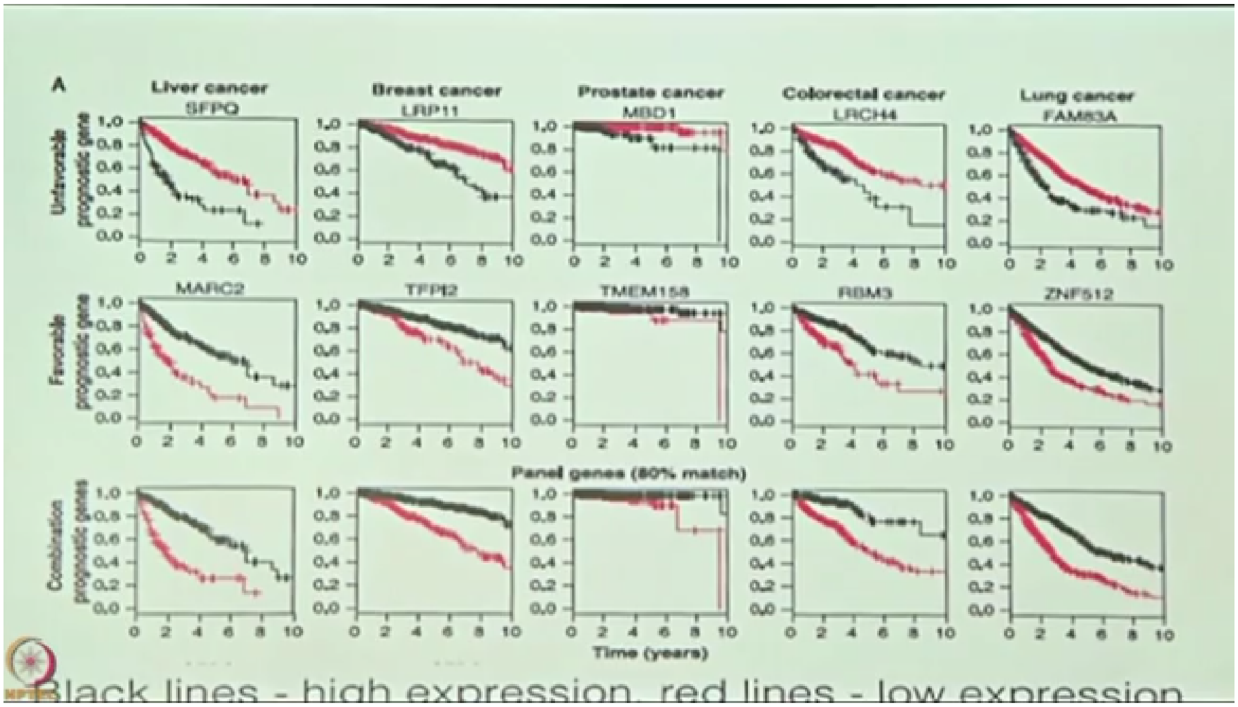


Clinical outcome based on gene expression analysis

- For each gene and cancer type, the patient cohort was stratified into two groups with the highest and lowest expression (FPKM) based on individual expression levels
- more than 100 million Kaplan-Meier plots were generated that corresponded to all 19,571 protein-coding genes across the 17 cancer types

We stratified the RNA-Seq data in each patient into those expressing the highest and those expressing the lowest and correlated that with the outcome, so the basic point was to see if this is the highest and the patient is doing badly then that's not a good deal, that's finally what we wanted to say, and for this exercise they were more than a 100 million Kaplan Meier plots that were generated, I don't think anybody saw all of them, it was just the machine okay,

(Refer Slide Time: 13:02)

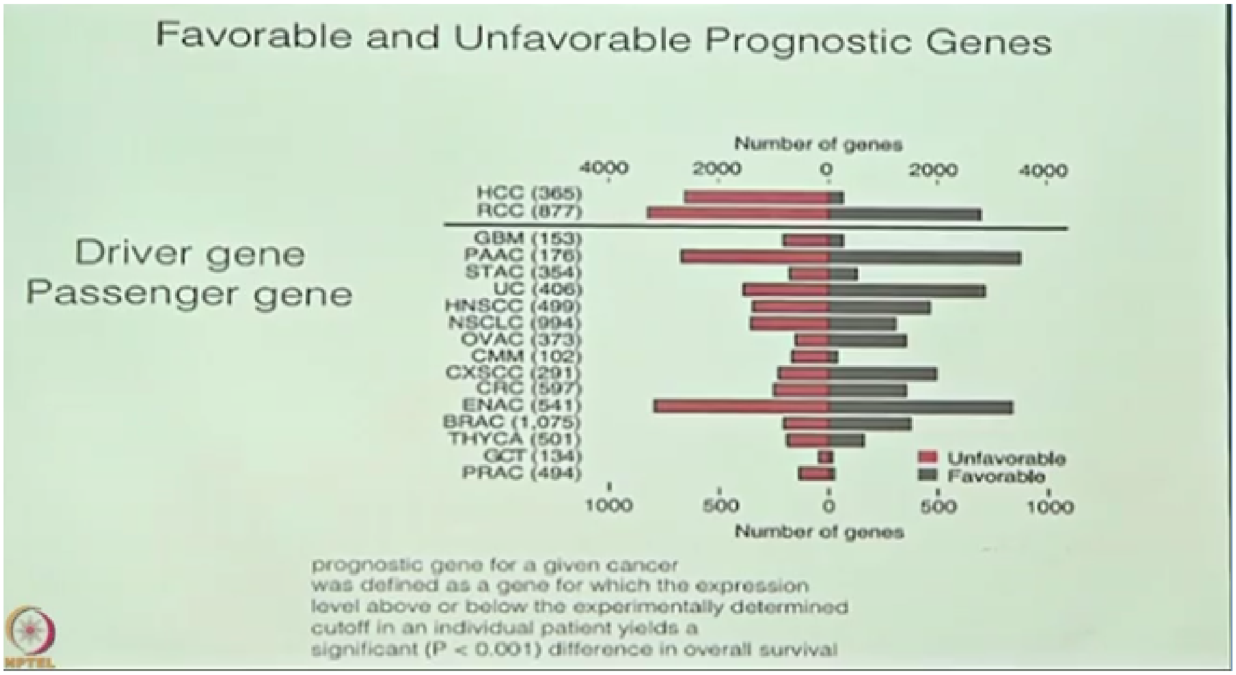


Black lines - high expression, red lines - low expression

so let me give you an example of what favorable and unfavorable genes we found, so let's look at the different cancers on top right here, black lines mean high expression, red lines mean no expression, so this is being expressed high and therefore you see these events happening faster and faster until this time until the patient is gone, therefore that classifies it as an unfavorable prognostic indicator.

This gene MARC 2 the events seem to be happening more slowly, if that is present and therefore that's a favorable prognostic indicator. Finally if you combine with a panel that's what you get, let me put it in a different way,

(Refer Slide Time: 1:01)



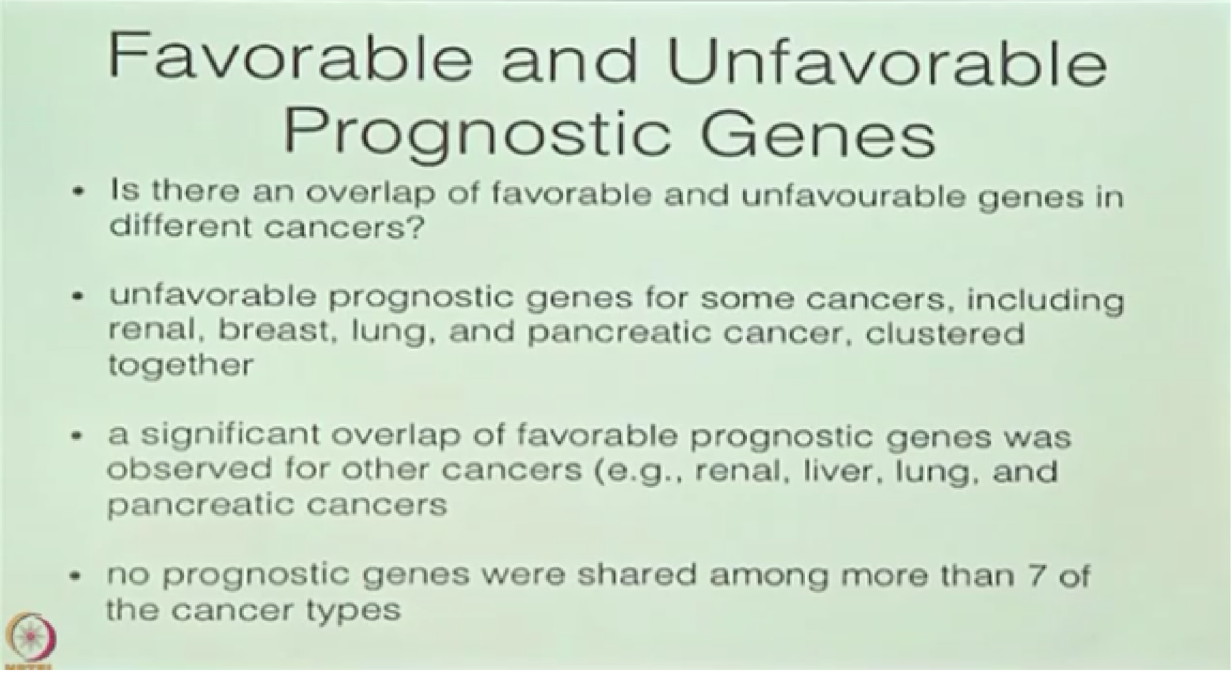
Favorable and Unfavorable Prognostic Genes

Driver gene
Passenger gene

prognostic gene for a given cancer was defined as a gene for which the expression level above or below the experimentally determined cutoff in an individual patient yields a significant ($P < 0.001$) difference in overall survival

the number of prognostic genes were classified into favorable and unfavorable as I just told you.

Hepatocellular carcinomas and renal cell carcinoma which are displayed at the top had the maximum number of prognostic genes in the study, in which we found a correlation, what did we call a prognostic gene? Why did we say that this has got some prognostic effect? Because the expression level was above, the experimentally determined cutoff in an individual patient that is a statistical analysis, so I am just reading it off the chap there, be with less than 0.001.
(Refer Slide Time: 14:58)

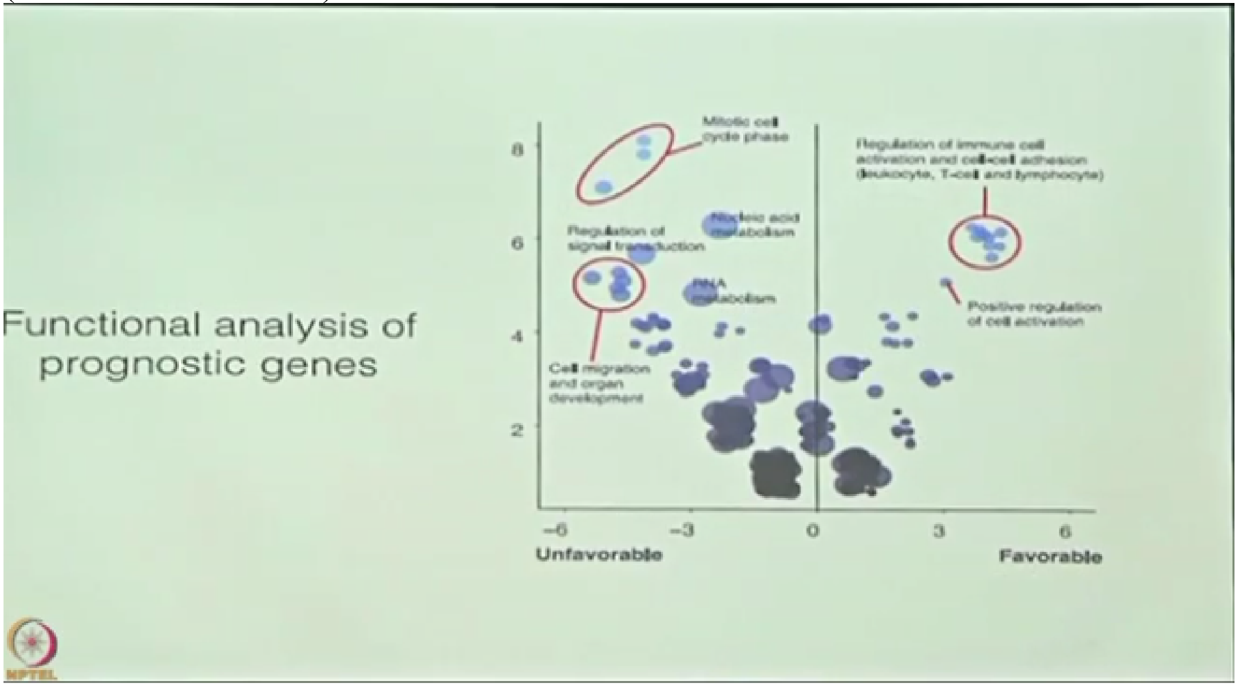


Favorable and Unfavorable Prognostic Genes

- Is there an overlap of favorable and unfavourable genes in different cancers?
- unfavorable prognostic genes for some cancers, including renal, breast, lung, and pancreatic cancer, clustered together
- a significant overlap of favorable prognostic genes was observed for other cancers (e.g., renal, liver, lung, and pancreatic cancers
- no prognostic genes were shared among more than 7 of the cancer types

Now in favorable and unfavorable genes did we fine more favorable, same favorable genes in more than 1 cancer, and unfavorable genes in more than 1 cancer? Yes, there is a big overlap, some unfavorable genes for some cancers like lung cancer, pancreatic cancer clustered together, favorable prognostic genes were seen in liver, lung, but there was an overlap of those genes there, and finally no prognostic genes were shared in more than 7 of the cancers which we thought was significant, because we didn't get it across the board saying favorable and unfavorable for everything.

(Refer Slide Time: 15:51)

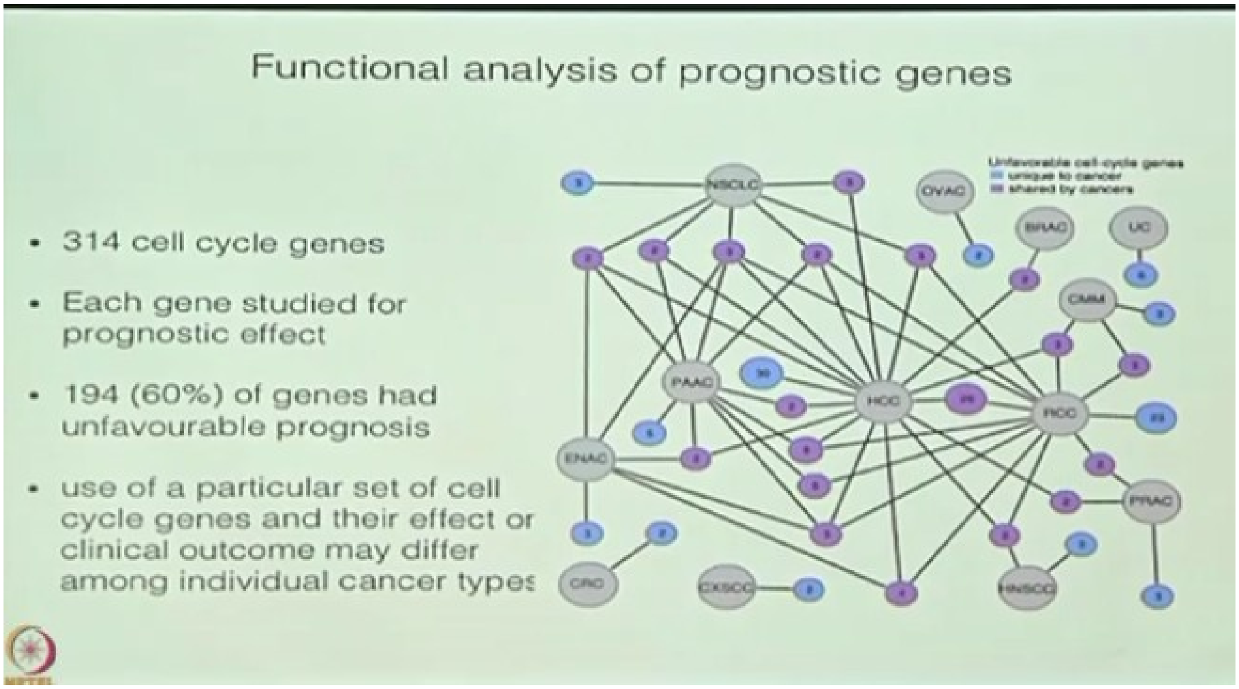


Functional analysis of prognostic genes

Take a look at this chart, these are the unfavorable indicators, these are the favorable indicators, in the unfavorable indicators the most impressive one was the mitotic cell or the cell cycle phase, which we know also for a fact, we've seen that as well even before these tools were not available that tumors which are rapidly dividing, they will be an unfavorable indicator so that much was proved.

On the favorable side it was mainly the regulation of immune cell activation,

(Refer Slide Time: 16:35)

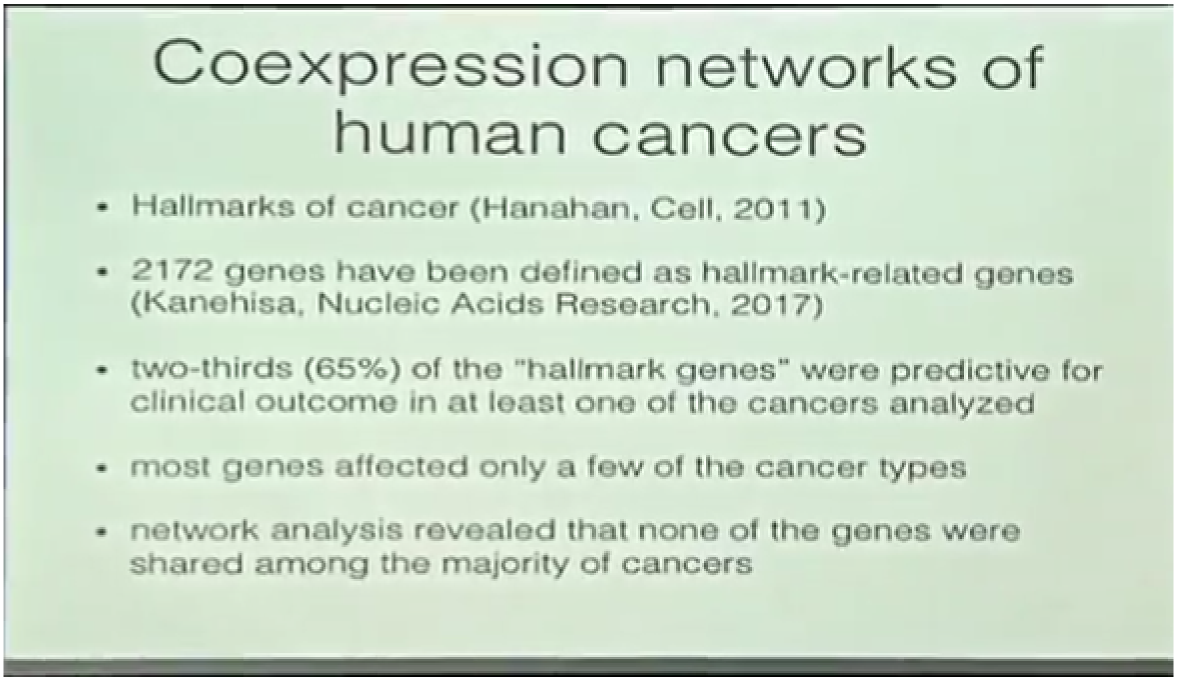


Functional analysis of prognostic genes

- 314 cell cycle genes
- Each gene studied for prognostic effect
- 194 (60%) of genes had unfavourable prognosis
- use of a particular set of cell cycle genes and their effect or clinical outcome may differ among individual cancer types

because the mitotic cell cycle was found to be significant, all 314 cell cycle genes were studied, because it was that significant, and each gene was studied separately for a prognostic effect, now if you ask me for this was the big surprise for me, you know what my concept as a diagnostic pathologist is for mitotic index, you either count the mitosis on the slide or there is a marker called KI67 mib1 which everybody saw as by, but which does not work in all cases, so that's just one prognostic gene from the cell cycle, therefore all cell cycle genes may not apply to all cancers, and there was a publication in 2011 very famous publication which was called hallmarks of cancer, in that there were 2000 odd genes which were defined as hallmark genes of cancer, it's the biggest work of its kind up to date.

When we studied those hallmark genes in our data two thirds of them, 65% of them were predictive for the clinical outcome in at least one cancer,
(Refer Slide Time: 18:02)

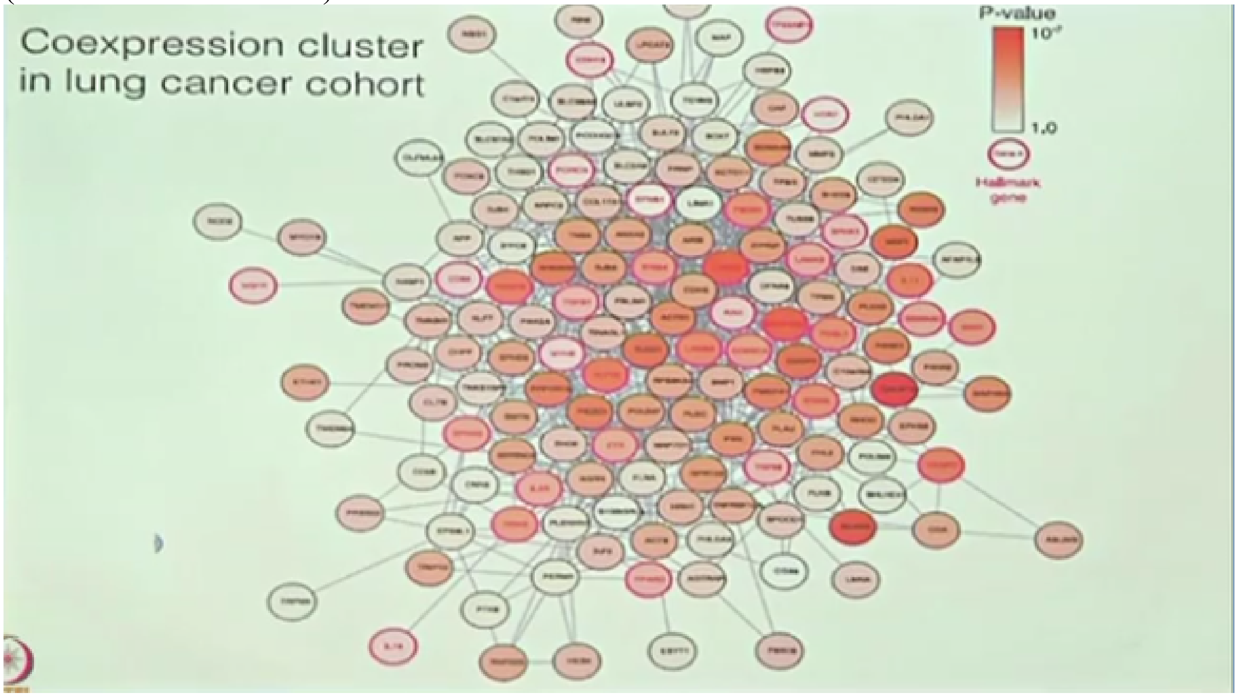


Coexpression networks of human cancers

- Hallmarks of cancer (Hanahan, Cell, 2011)
- 2172 genes have been defined as hallmark-related genes (Kanehisa, Nucleic Acids Research, 2017)
- two-thirds (65%) of the "hallmark genes" were predictive for clinical outcome in at least one of the cancers analyzed
- most genes affected only a few of the cancer types
- network analysis revealed that none of the genes were shared among the majority of cancers

so that was verified what was earlier reported at least by our study. Most genes affected only a few of the cancer types, all cancers were not affected from the hallmark genes, and the network analysis showed that most of those genes were not shared,

(Refer Slide Time: 18:30)



Coexpression cluster in lung cancer cohort

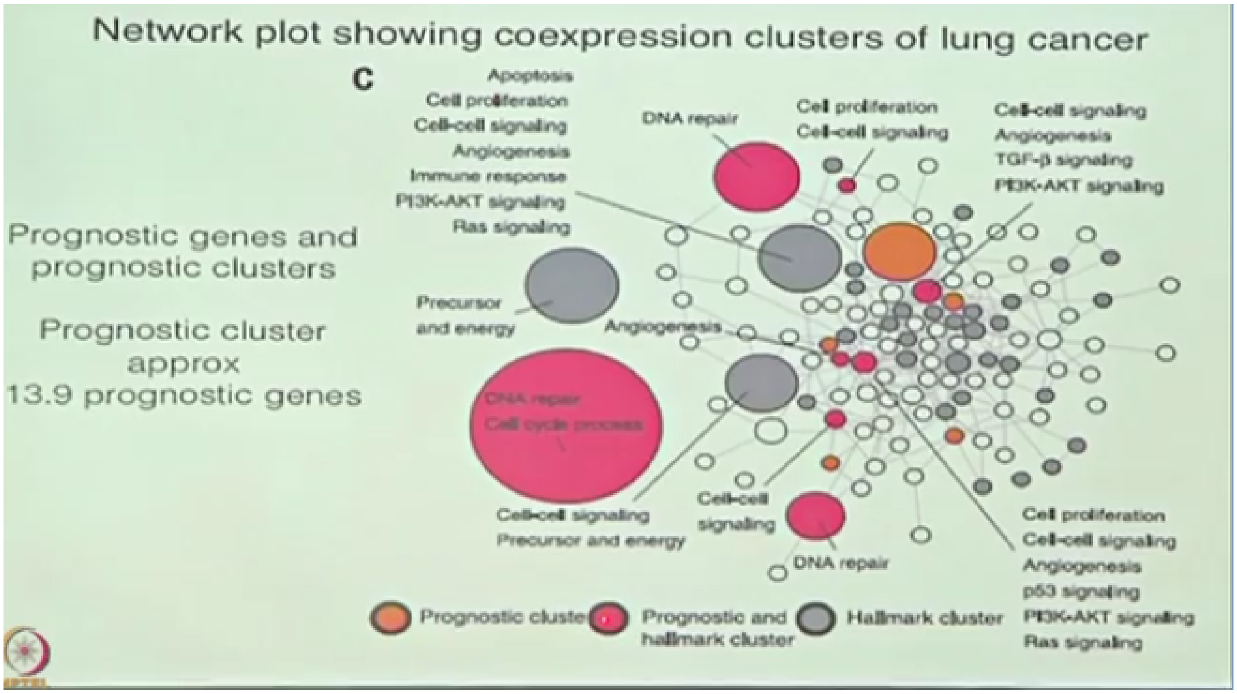
so the next step was to take a cancer example from lung, and to see you see what I'm, in case you've lost track of it, we started out trying to say that these are prognostic genes, we at the end of all this stuff, we said yeah these are the prognostic genes, then as we discussed in the earlier lecture let's go back and say what did other people say about it, so that's how the hallmarks of cancer paper came in, so we are studying, we are trying to find out what are the prognostic genes in cancer.

We’ve done all the stuff, we have done all this data analysis and we say yeah we’ve got something, now is this true or are we backing up the wrong tree? And there is very little work in this area, so is there any other work is the hallmarks of cancer papers, two of them to be exact, so pull them out, what did they say are the hallmark genes? So what they said at the hallmark genes and what we are saying, is it matching? So 65% of them matched, so they went entirely wrong, or entirely right, what we found was that most of those genes which they had identified it affected only a few of the cancer types, so we begin to think that maybe they didn’t get all of them, and the network analysis showed that none of these genes were shared by the cancers which maybe meant that they were looking for specific genes and individual cancers, not necessarily all the genes, so am I clear? Is it better now? Okay.

Now what we did then was, we went into a specific lung cancer example, in that the statisticians they always create something very beautiful, attractive, so this is one of those examples, but don’t get carried away by that, let’s try to understand what it means, if a gene is circled, these are all genes, I’m sorry I couldn’t enlarge it enough for you to read actually what’s written but that’s a name of a gene in there, and if it’s got this red circle around it, it means that it was the Hallmark paper that forceded, and we studied it along with others.

If it doesn’t have the red mark it means that it came up as the prognostic gene in our study, so if you look at the middle these are called the hub genes right at the center, and they have a greater, all of them were thought to be prognostic, they all came up as prognostic, but during plotting this you get some genes which are in the hub, and you get some genes which are in the periphery, and there is a greater likelihood of these genes in the hub being having a prognostic effect rather than once at the periphery, and therefore it’s a speculation that when prognostic genes affect the cancer that 50 of them or 100 of them or 500 of them, who are the really bad guys, who are the drivers, and who are the passengers, so it’s tempting to speculate that these guys in the middle are the drivers, and the once at the periphery are the passengers, but it’s only a speculation and that’s where you should stay for now.

Then another beautiful diagram, it looks like popcorn, doesn’t it? Yeah, so what the statisticians did was they said you are talking about these prognostic genes 1-1 gene and 20 genes you have got there, you talk about a prognostic cluster, all of which are closely related, they are so many genes which are related, which are doing DNA repairs, cell cycle processes, you get them together, make it a prognostic cluster, so that’s what we did, we made this a prognostic cluster,
(Refer Slide Time: 23:06)



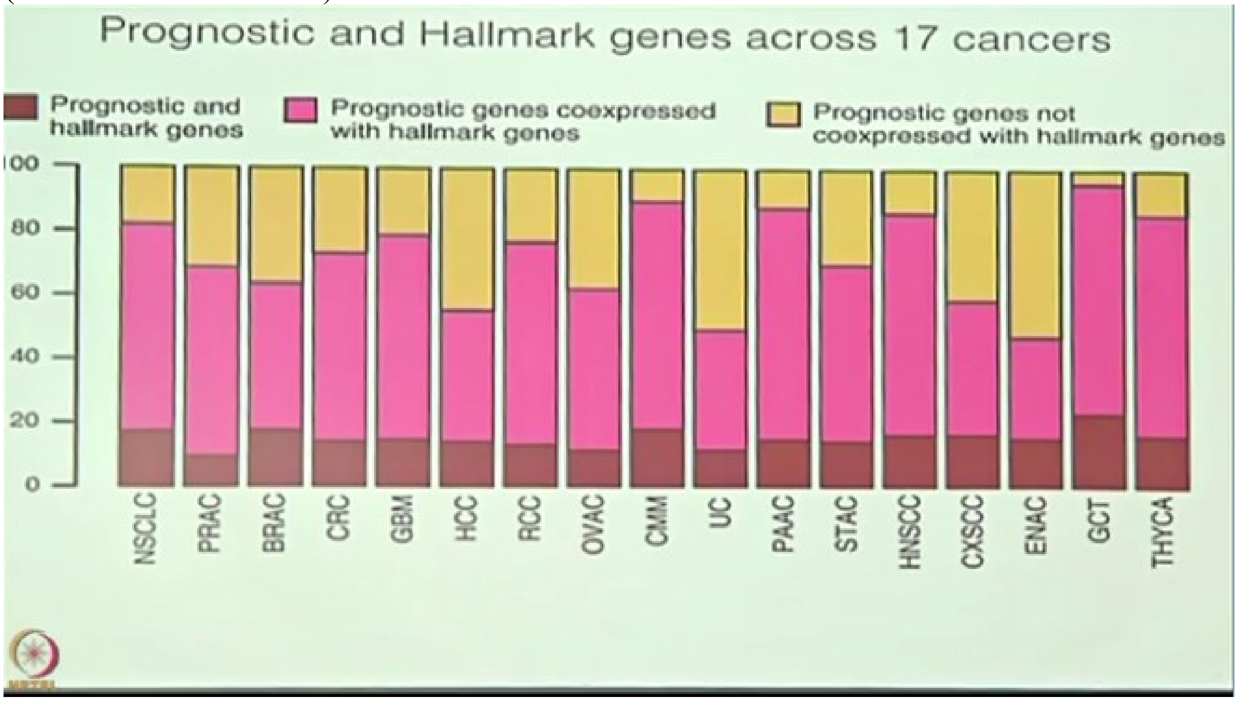
Network plot showing coexpression clusters of lung cancer

Prognostic genes and prognostic clusters

Prognostic cluster approx 13.9 prognostic genes

this is many genes together and in this diagram the reverse is true, if you look at the periphery, the large ones are the ones which have a greater chance of having an effect, the inner ones not so many, and also the differences are highlighted are the grey ones are the hallmark clusters, they were published 10, you know I'm sorry, 7, 8 years ago, and the prognostic and hallmark clusters are our work which has a superimposed on the hallmark we agreed that these are all important. And then there is a third group in which we say this is a prognostic cluster, but hallmark hasn't talked about it.

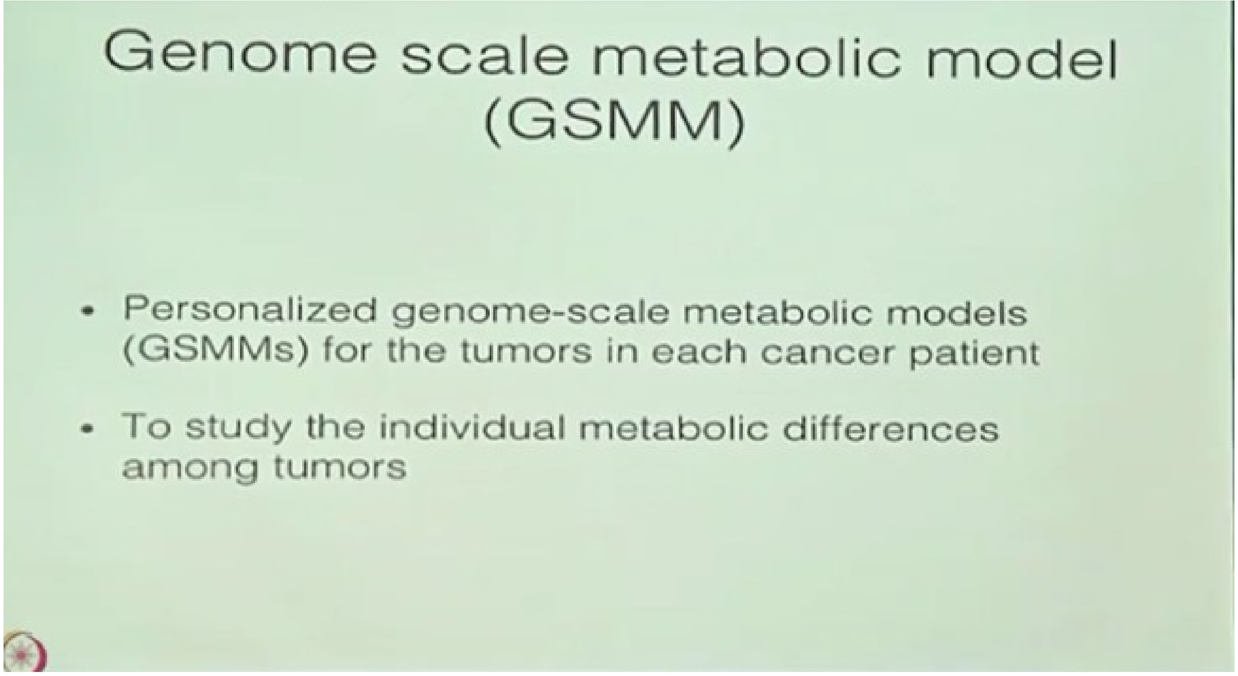
(Refer Slide Time: 23:56)



Prognostic and Hallmark genes across 17 cancers

A slightly different way of putting this, these are the prognostic and hallmark genes, we agree, these are the prognostic genes which are co-expressed with the hallmark genes, and these are the different prognostic genes which we found in our study, which the hallmark papers have not mentioned.

(Refer Slide Time: 24:32)



Genome scale metabolic model (GSMM)

- Personalized genome-scale metabolic models (GSMMs) for the tumors in each cancer patient
- To study the individual metabolic differences among tumors

Now with all this information bringing it to the end of all most my talk, is it possible to generate a personalized model for an individual cancer for treatment, something that is refer to now as a genome scale metabolic model, if you construct a full genome scale model for this cancer and for the next cancer, let's take two liver cancers for which all the proteomics are known or transcriptomics are known. We will be able to compare and say how they are different, are you with me or no?

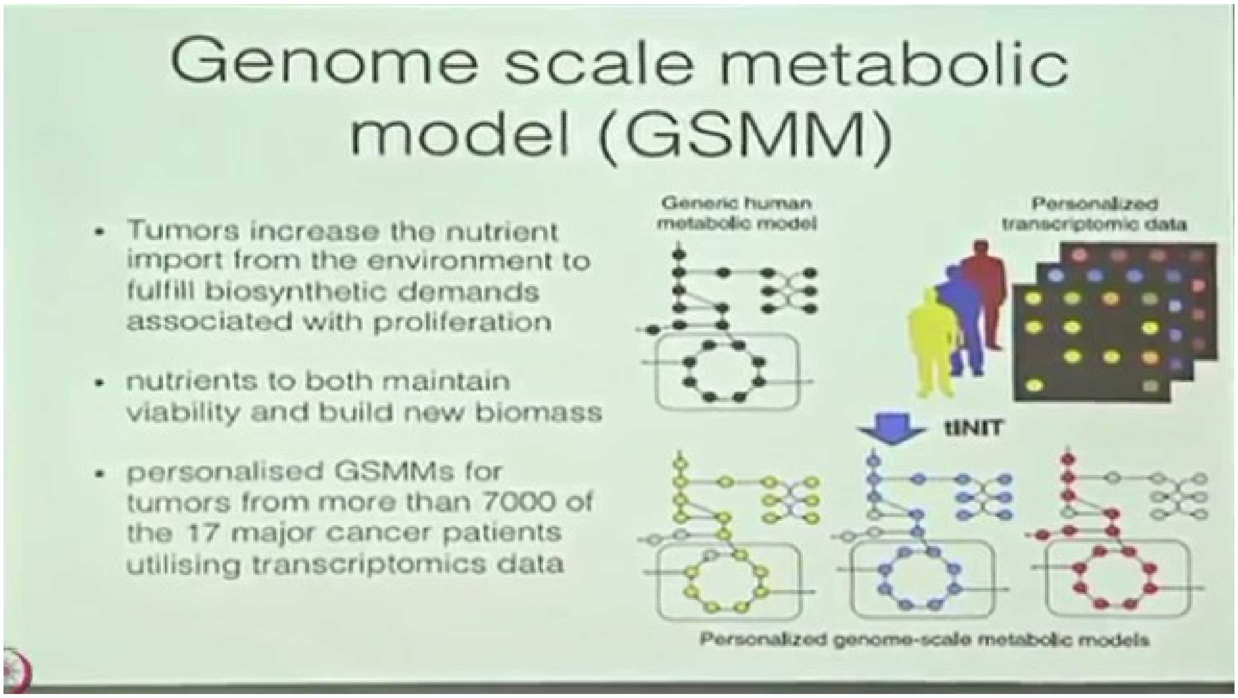
Unidentified Speaker: _25:20_

Dr. Sanjay Navani: No, this is, it's quite simple there is no need to get into genetic differences between individuals at this stage, we are now talking about genetic differences between cancers, which look the same, which are off the same type, so hepatocellular cancer which looks the same, what does the transcriptomics data say on it, that's a question I'm trying to answer right now, so let's just finish that, are you with me?

Unidentified Speaker: Yes.

Dr. Sanjay Navani: Yeah, okay, so that's we carried out a personalized genome scale metabolic model for tumors from more than 7,000 of the 17 major cancer patients, and we expected something different because cancer cells pull in more nutrients from the surrounding

(Refer Slide Time: 26:36)



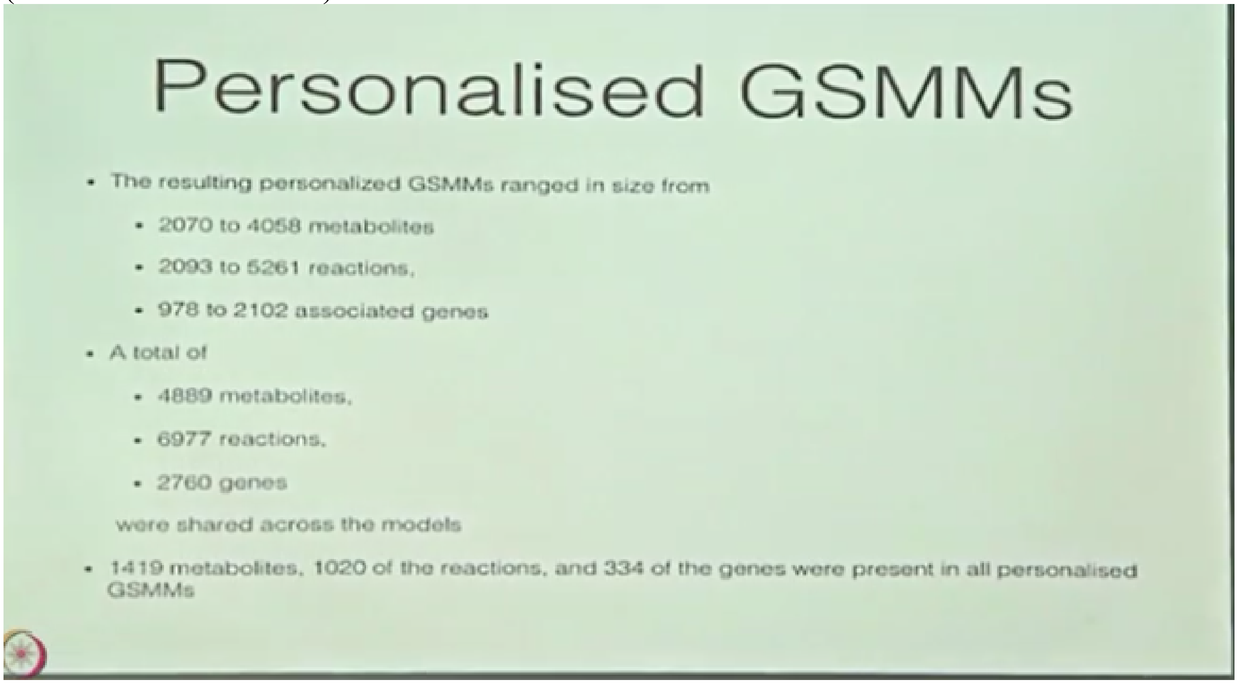
Genome scale metabolic model (GSMM)

- Tumors increase the nutrient import from the environment to fulfill biosynthetic demands associated with proliferation
- nutrients to both maintain viability and build new biomass
- personalised GSMMs for tumors from more than 7000 of the 17 major cancer patients utilising transcriptomics data

Generic human metabolic model
Personalized transcriptomic data
tINIT
Personalized genome-scale metabolic models

and they build up more of a biomass,

(Refer Slide Time: 26:40)



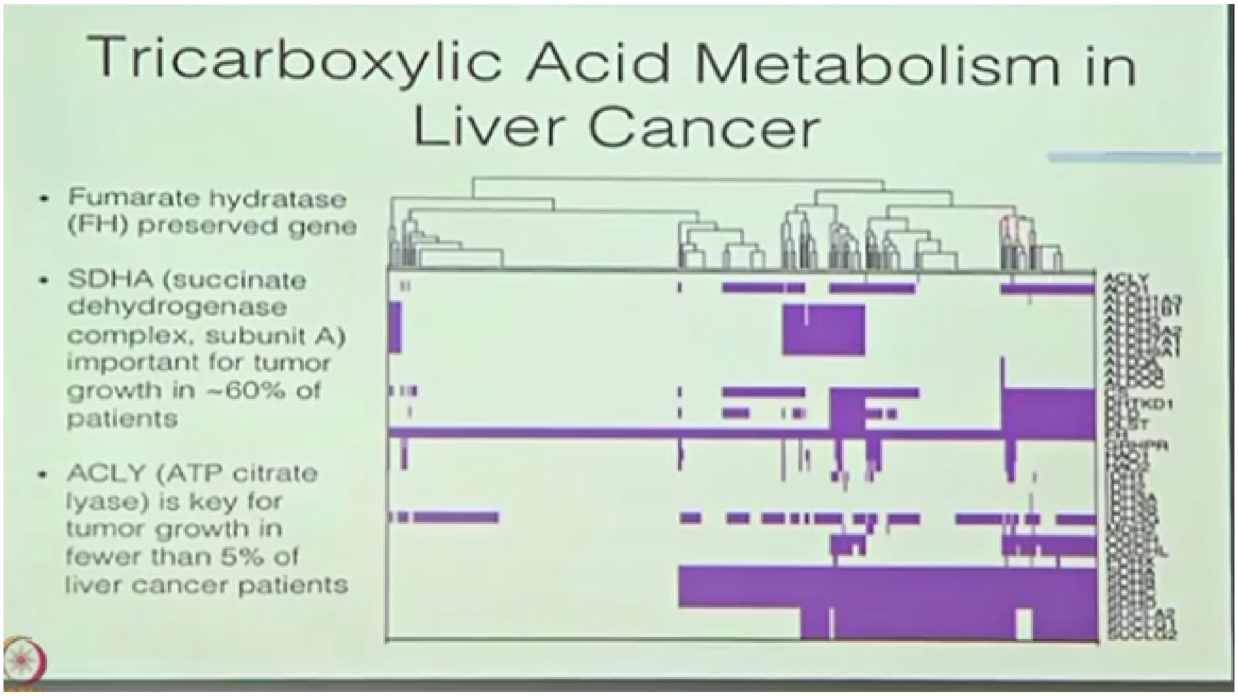
Personalised GSMMs

- The resulting personalized GSMMs ranged in size from
  - 2070 to 4058 metabolites
  - 2093 to 5261 reactions,
  - 978 to 2102 associated genes
- A total of
  - 4889 metabolites,
  - 6977 reactions,
  - 2760 genes

  were shared across the models
- 1419 metabolites, 1020 of the reactions, and 334 of the genes were present in all personalised GSMMs

this is some of the statistics let it throughout, I won't go into all that except to say that 1400 metabolites or 1000 reactions and 334 of all the genes were present in all the personalized models, it was common.

(Refer Slide Time: 27:07)

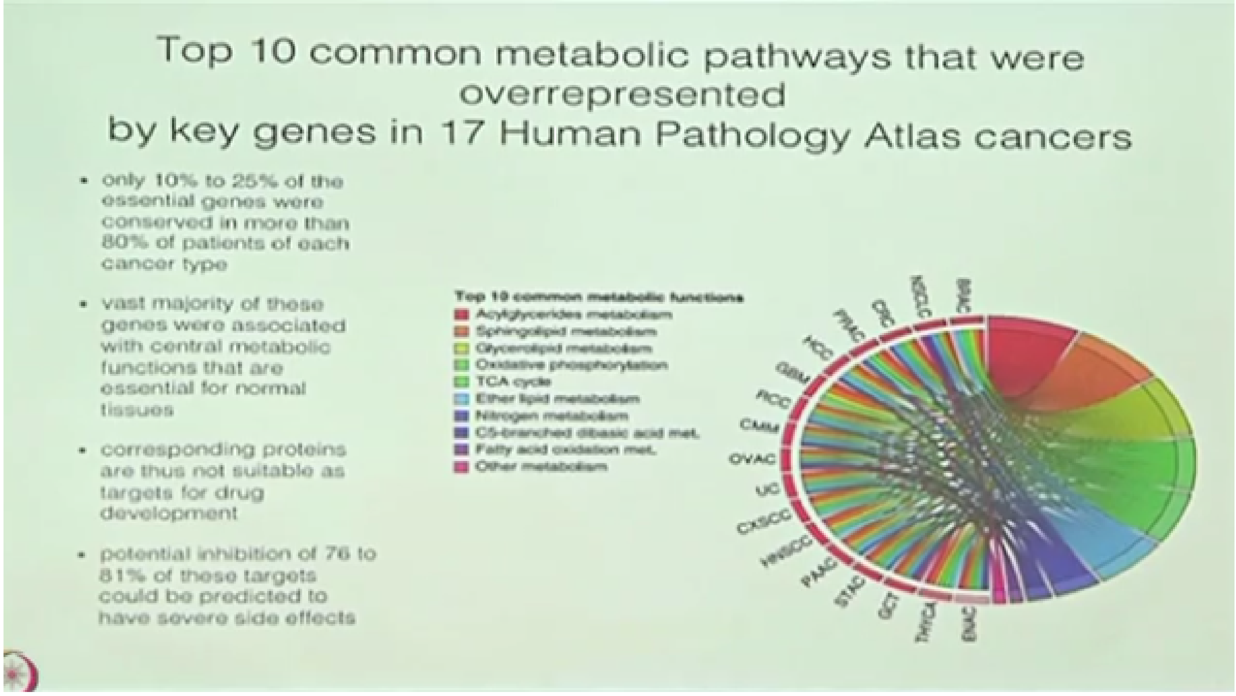


Tricarboxylic Acid Metabolism in Liver Cancer

- Fumarate hydratase (FH) preserved gene
- SDHA (succinate dehydrogenase complex, subunit A) important for tumor growth in ~60% of patients
- ACLY (ATP citrate lyase) is key for tumor growth in fewer than 5% of liver cancer patients

Then we looked at these are all liver cancer patients and I am looking at elements of tricarboxylic acid metabolism, so I want to tell you that in that FH or fumarate hydratase was found in all liver cancer tumors. Then I want to tell you that ACLY right there on the top, see those small bars there, it was found in less than 5% of liver cancer patients, and finally succinate dehydrogenate complex unit A was found in 60%, the point I'm trying to make is that there is sufficient difference amount cancers of the same type which underscores the aid for a personalize model, okay.

(Refer Slide Time: 28:00)



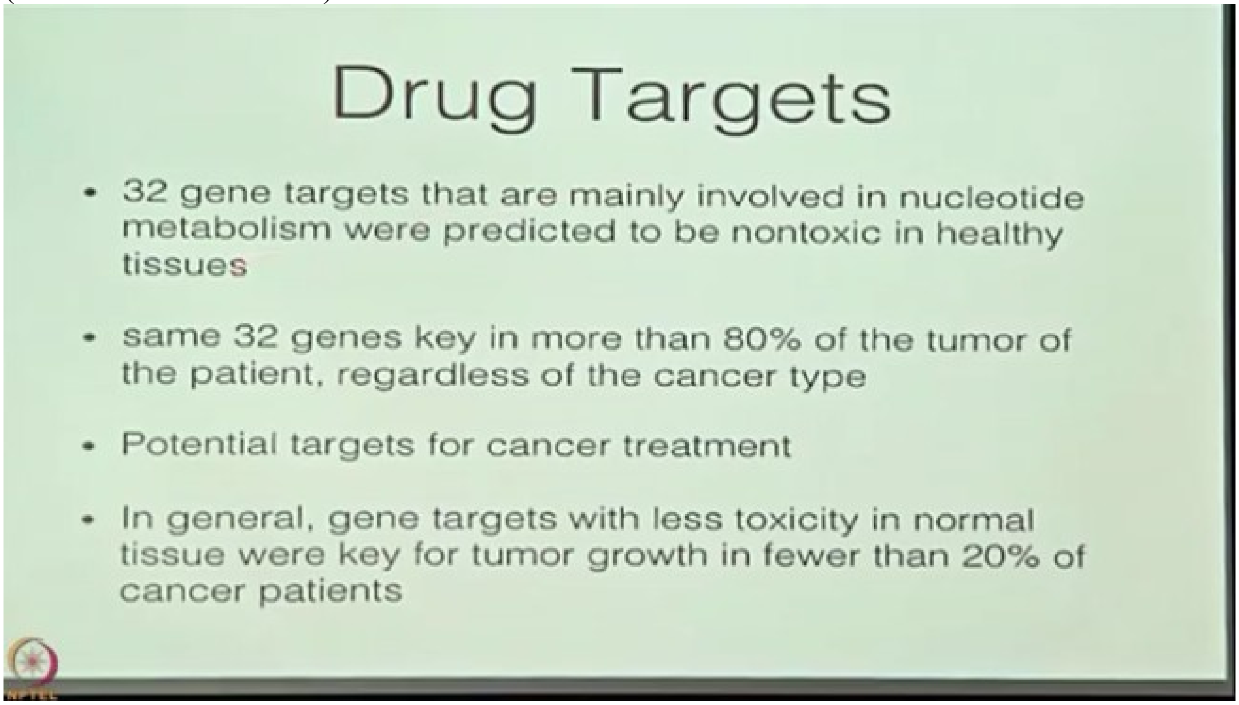
Top 10 common metabolic pathways that were overrepresented by key genes in 17 Human Pathology Atlas cancers

- only 10% to 25% of the essential genes were conserved in more than 80% of patients of each cancer type
- vast majority of these genes were associated with central metabolic functions that are essential for normal tissues
- corresponding proteins are thus not suitable as targets for drug development
- potential inhibition of 76 to 81% of these targets could be predicted to have severe side effects

One more impressive chart for you to understand, this thing over here the most common genes that were commonly expressed were of the most common metabolic functions, and they were all expressed in several of these cancers.

Now if you think of a drug target, because these are common metabolic functions we have also occurring in normal cells, therefore if you give anything to these patients as I outlined before, there was a possibility that about 80% of these targets would have side effects, so you hit the cancer, but you also hit the normal cell. And now the concept of what chemotherapy does to a patient, I think you can begin to appreciate, why they have so many problems? Okay.

(Refer Slide Time: 29:05)

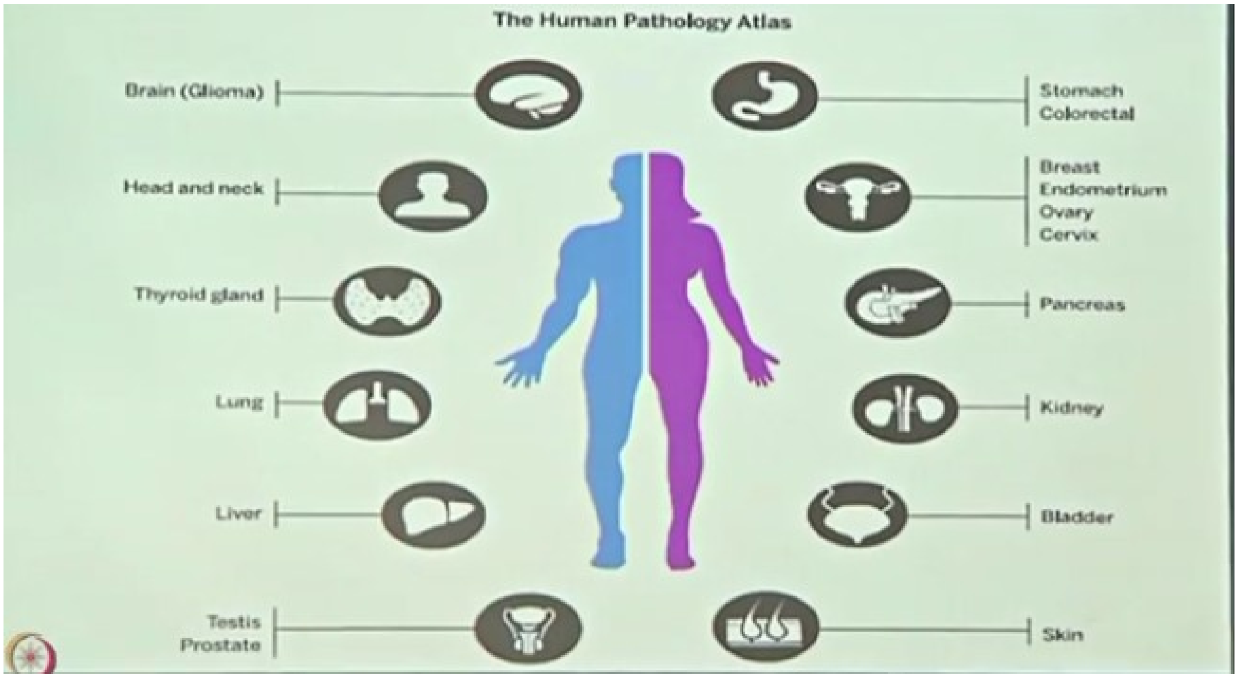


Drug Targets

- 32 gene targets that are mainly involved in nucleotide metabolism were predicted to be nontoxic in healthy tissues
- same 32 genes key in more than 80% of the tumor of the patient, regardless of the cancer type
- Potential targets for cancer treatment
- In general, gene targets with less toxicity in normal tissue were key for tumor growth in fewer than 20% of cancer patients

Now the last point, 32 gene targets that were mainly involved in nucleotide metabolism look like potential targets, they are expressed in more than 80% of the tumors of the patient regardless of the cancer type, and they are potential targets because they will not affect the normal cells, okay,

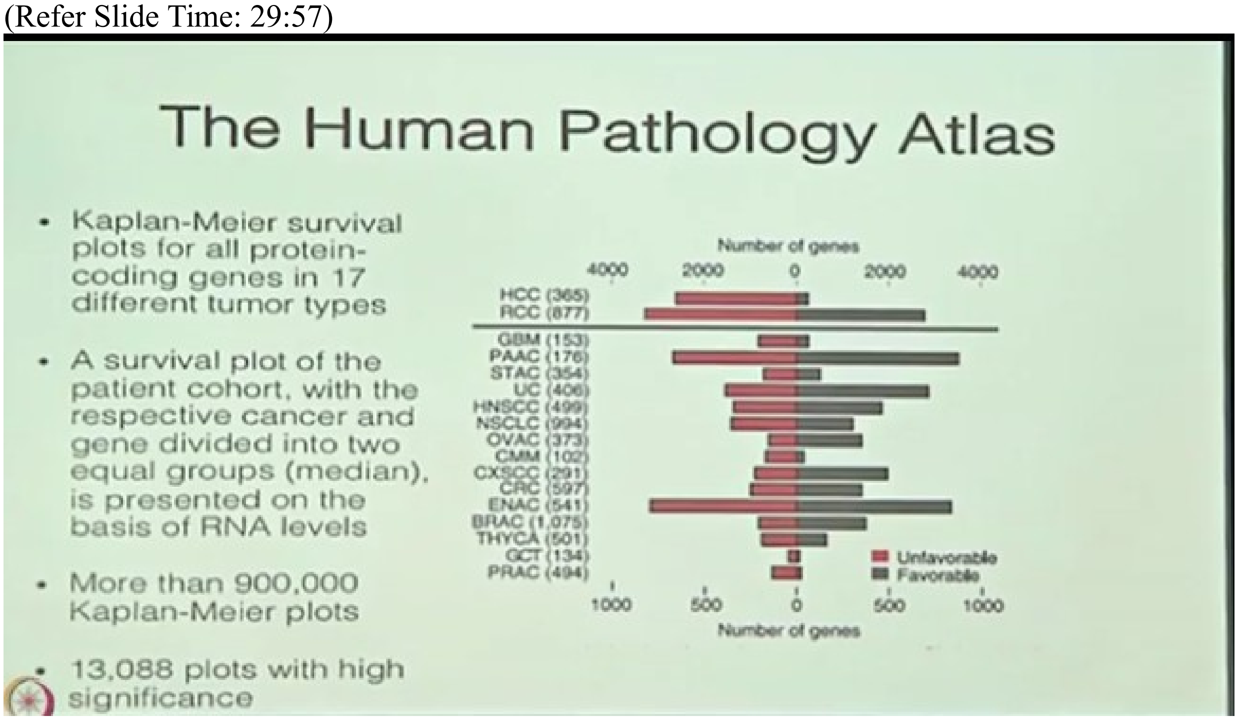
(Refer Slide Time: 29:37)



The Human Pathology Atlas

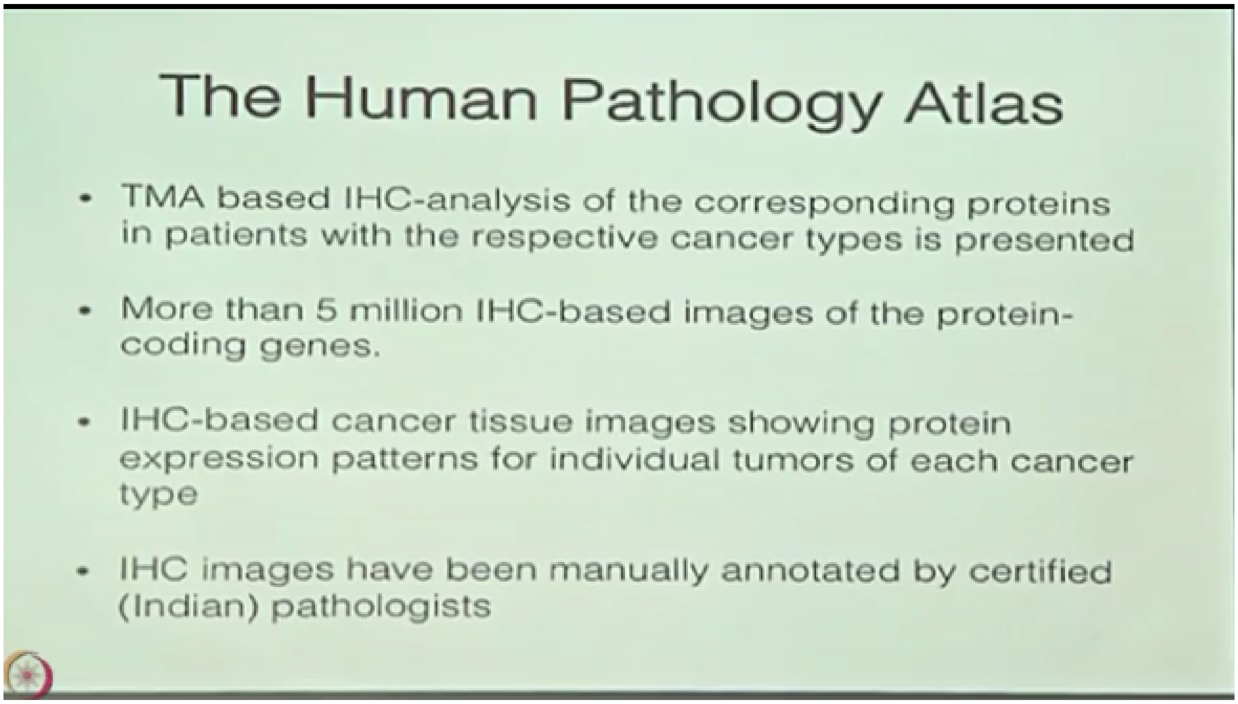
so if you go to the protein atlas there is a new section called it's well 6 months old now, but the human pathology atlas which will give you access to all the Kaplan Meier plots you want to see anything, the significant plots, the insignificant plots, all the data is up there,

(Refer Slide Time: 29:57)



The Human Pathology Atlas

- Kaplan-Meier survival plots for all protein-coding genes in 17 different tumor types
- A survival plot of the patient cohort, with the respective cancer and gene divided into two equal groups (median), is presented on the basis of RNA levels
- More than 900,000 Kaplan-Meier plots
- 13,088 plots with high significance

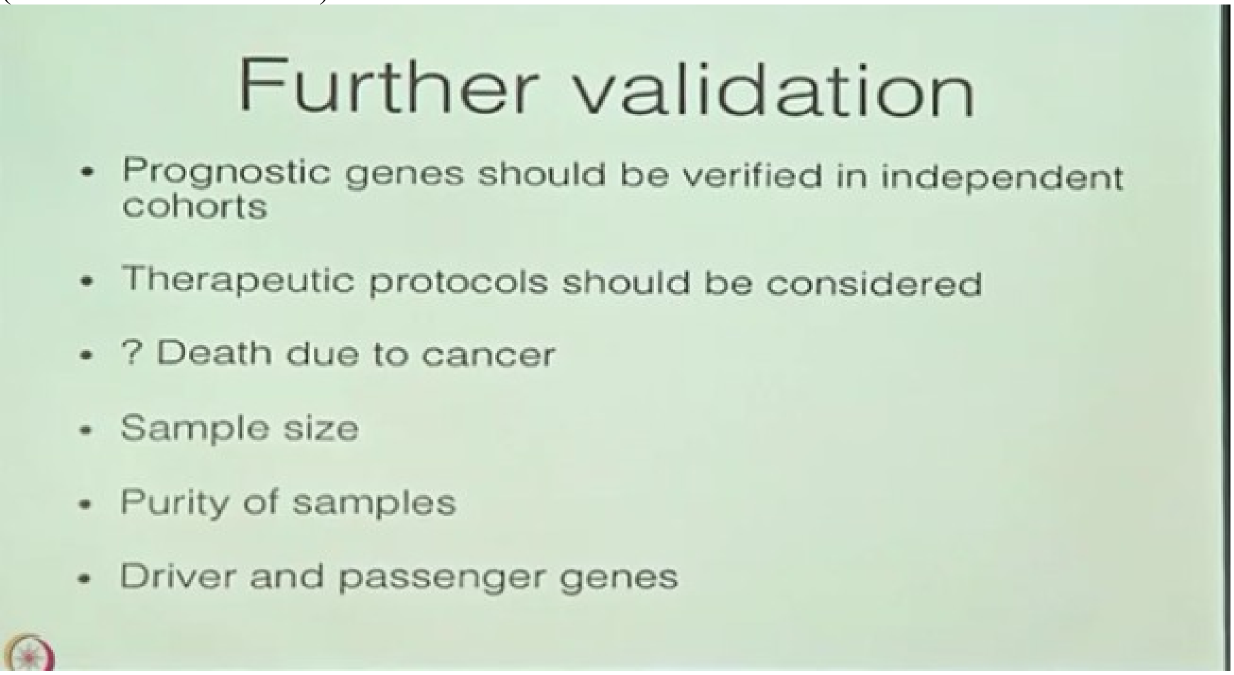


The Human Pathology Atlas

- TMA based IHC-analysis of the corresponding proteins in patients with the respective cancer types is presented
- More than 5 million IHC-based images of the protein-coding genes.
- IHC-based cancer tissue images showing protein expression patterns for individual tumors of each cancer type
- IHC images have been manually annotated by certified (Indian) pathologists

you will also get the survival of the patients if you wish to check the, more than 5 million IHC images for cancer can be seen there, most of it annotated by us here in India.

(Refer Slide Time: 30:21)



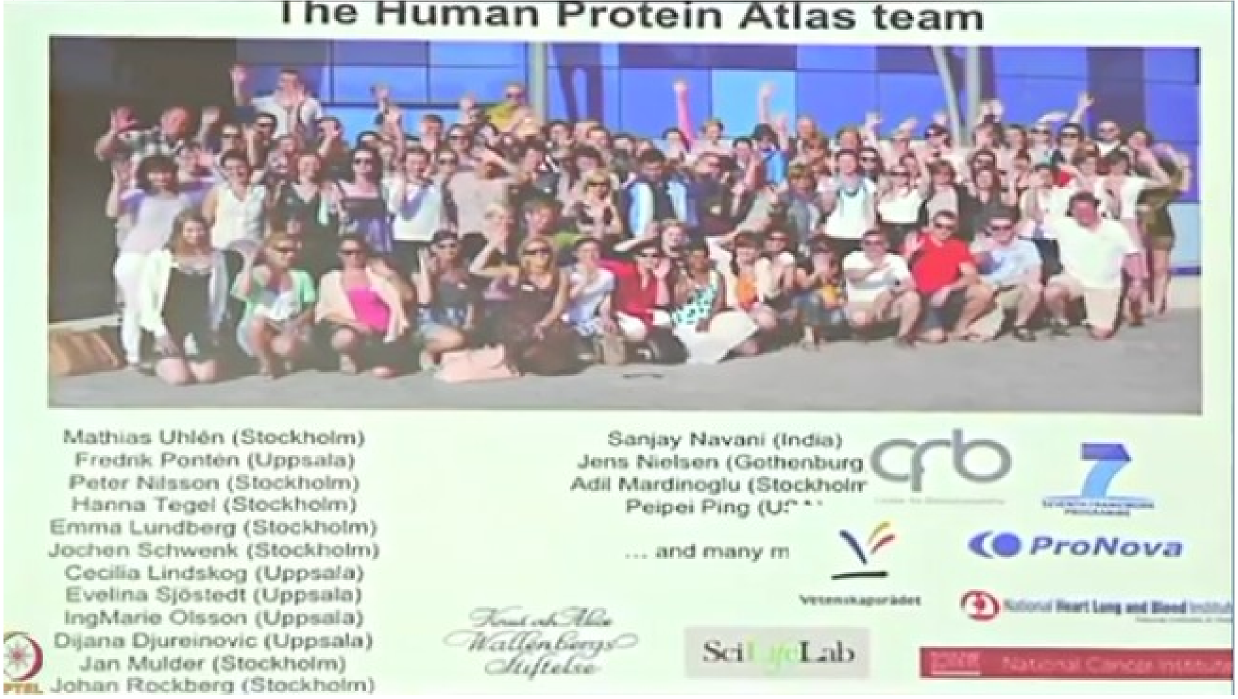
Further validation

- Prognostic genes should be verified in independent cohorts
- Therapeutic protocols should be considered
- ? Death due to cancer
- Sample size
- Purity of samples
- Driver and passenger genes

A few points just to leave in your mind thus I finish, these prognostic genes as we heard in the morning, they have to be verified in an independent cohort, secondly the death whichever the question that Joshua raised earlier, we are assuming that the death was due to the cancer, but we don’t know that data is not available.

Another point which was discussed earlier in the morning, how pure was the sample? People spoke about not being fixed on time, technical issues, is that also confounding the data, and finally out of all this prognostic genes are all causing the cancer or some guys doing it and the others are just following, and what are those? I'll just like to leave with that thanks to the people who made it possible for us to do the job here in India in particular this man right there,

(Refer Slide Time: 31:34)



The Human Protein Atlas team

Mathias Uhlén (Stockholm)
Fredrik Pontén (Uppsala)
Peter Nilsson (Stockholm)
Hanna Tegel (Stockholm)
Emma Lundberg (Stockholm)
Jochen Schwenk (Stockholm)
Cecilia Lindskog (Uppsala)
Evelina Sjöstedt (Uppsala)
IngMarie Olsson (Uppsala)
Dijana Djureinovic (Uppsala)
Jan Mulder (Stockholm)
Johan Rockberg (Stockholm)

Sanjay Navani (India)
Jens Nielsen (Gothenburg
Adil Mardinoglu (Stockholm
Peipei Ping (U

... and many m

crb
7
ProNova
Vetenskapsrådet
National Heart Lung and Blood Institute
Knut och Alice Wallenbergs Stiftelse
SciLifeLab
National Cancer Institute
PTEL

Mathias Uhlen who is the director, really quite a person I enjoyed meeting and interacting with, I think you know when you associate with people from other places, other disciplines, other countries, you get exposed to many things which may not necessarily be true from the country that you are working from.

What I envied him mostly for an which I had told him quite frequently was his capacity to think big, not just in numbers you see when you start thinking big, you really think big on all levels, I have never been able to get over how he gave me the job, I met him for 15 minutes, but it was of course the previous work up, you know more younger people in the organization met me, I put my forward my ideas, everything happened, I gave a presentation and then finally I met him for 15 minutes and he said okay, I think your idea looks good, I've just one question, have you ever done this thing before, scan all these images, send them to India, look them on the computer, work on the software, but I said no it's just an idea.

In the research field people look for background, you can come and say anything, but what have you actually done? Now fortunately for me this kind of thing I had never been done, so new ideas were okay, if the idea is good let's give it a shot, we were supposed to do 2 million images in a year, and when I first heard that I didn't know whether I should say yes, but I did, but the moment I said yes, he said okay like a good senior person, he said okay I'm sure you can do it Sanjay, why don't we do a small experiment, why don't you just do 250,000 images for the first year, if you do those well and we get good results, you are able to maintain the quality control, there are many things, but internet has to work, the pathologist must understand, big exercise, then from the next year you do all 2 million, I thought that was fair, but I don't think he and I knew what we were talking about, because nothing had happened, so I went out, I had pathologist who had never seen a single slide of IHC in their lives, all this stuff is done by guys and girls who had never seen any IHC ever, and why did that I hire them? Because they wanted to learn it, and these are people from here, we are not talking about a different country, I'd just like all of you to remember that.

So after we started it, we finished 250,000 images in one and a half months, and Mathias called me and said do you want to continue? I told him you bet, and that's how the whole thing happened, so as time has gone by you know at one point in time the numbers use to be very important to me, 15 million images with quality control is that, all that you know, everything finally passes, the only thing that's remained and which always makes me feel very good when I think about it is that there were Indian pathologists who didn't know anything about IHC, who did this job, and it makes me feel very happy, because I never expected that, so I just want you to remember that, that it's very important to have enthusiasm, to have a positive outlook and to say clearly when you can do things and when you cannot do them, it's alright, thank you for your attention.

Unidentified Speaker: Sir, I'm treat with the gene that the metabolism gene that suggest targets, especially because you know, remember the targets with cancer even before and all the stuff where DNA metabolism?

Dr. Sanjay Navani: Yeah, yeah.

Unidentified Speaker: 5, 4 years so...

Dr. Sanjay Navani: Yeah, yeah.

Unidentified Speaker: So did those showed up I mean?

Dr. Sanjay Navani: Yeah, yeah, they showed up in fact there was, in fact there was the main group, there was the main group, so the cell cycle genes because there was a group that separated from the rest, all the 314 genes in the cell cycle were evaluated on an individual basis, and 60% of them showed a correlation with the cancers, unfavorable prognosis, but not all of them affected the same cancer, so there were different genes in the cell cycle itself which had different impacts on the cancer.

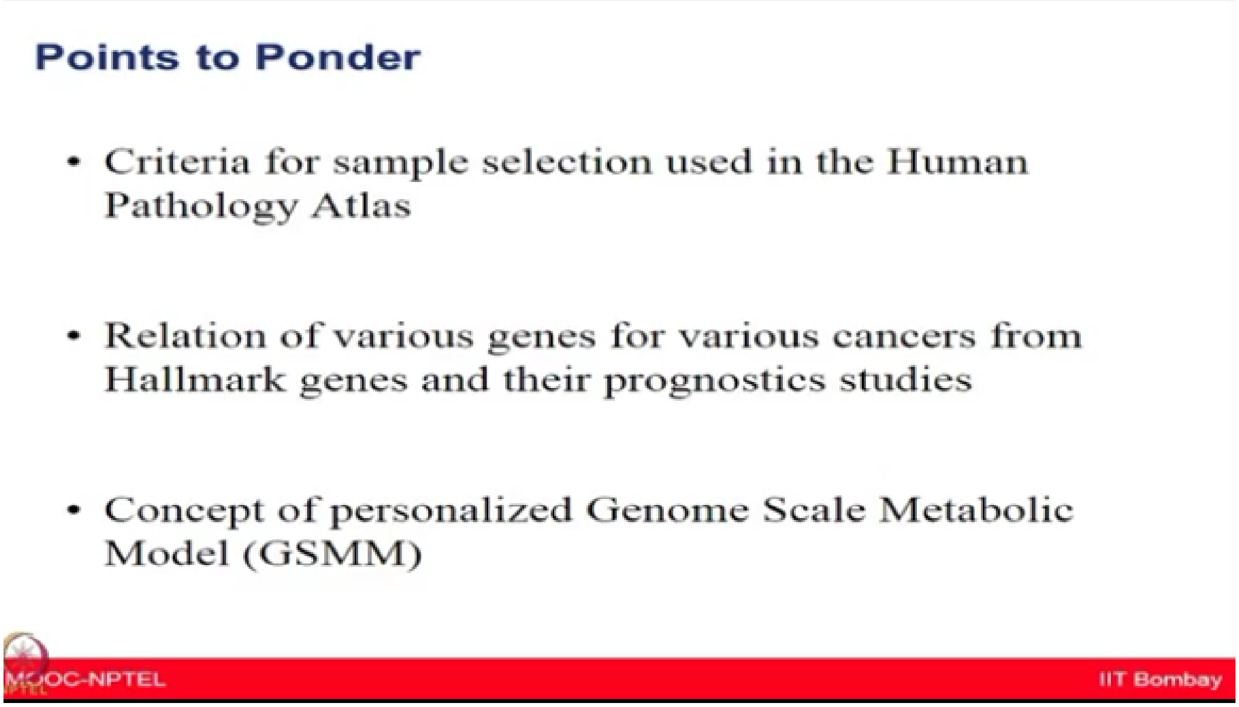
Unidentified Speaker: Yeah.

Dr. Sanjay Navani: So they were very much apart of the group, also published.

Unidentified Speaker: _36:57_

Dr. Sanjay Navani: Yeah, yeah, that was also described by the hallmark group, so all of that is confirmed and now there is and by our study I mean confirmed by our study and now there are additional prognostic genes which we've brought up, which we feel should be evaluated further.

(Refer Slide Time: 37:26)



Points to Ponder

- Criteria for sample selection used in the Human Pathology Atlas
- Relation of various genes for various cancers from Hallmark genes and their prognostics studies
- Concept of personalized Genome Scale Metabolic Model (GSMM)

MOOC-NPTEL IIT Bombay

Sanjeeva Srivastava: So I'm sure by now you got a very good understanding of the human protein atlas project and especially human pathology atlas project. The human pathology atlas was created as part of the human protein atlas program to explore the prognostic role of each protein coding gene in 17 different cancers, it really mega project, and this project the HP project shows the impact of protein levels for survival of patients with cancer, it uses transcriptomics and antibody waste profiling to provide a standalone resource for cancer position medicine, I must say that you should look into this really enriched resource for your own research where you can get so much data and information for all the possible proteins in various cancer type.

In the latest versions of HPA the survival scatter plot also show the clinical status for all the individuals in the patient cohorts, all the data which is presented are also made publically available in a very interactive open accelerator base to allows one to study the impact of individual proteins on clinical outcome in major human cancers, we are now moving almost towards the end of the course and I hope you are enjoying not only these lectures, but also the information available and resources available for you to conduct your own research even if you want to do some bio-pharmaceutics work just on sitting on your you know place, on your computers, you can do a lot just by looking at available data and from these resources, so I hope you are really going to make best use of this, and I'll see you again in the next lecture. Thank you.

(Refer Slide Time: 39:26)

Next lecture....

Experimental design and Statisical analysis I

MOOC-NPTEL

IIT Bombay

Prof. Sridhar Iyer

NPTEL Principal Investigator
&
Head CDEEP, IIT Bombay

Tushar R. Deshpande
Sr. Project Technical Assistant

Amin B. Shaikh
Sr. Project Technical Assistant

Vijay A. Kedare
Project Technical Assistant

Ravi. D Paswan
Project Attendant

Teaching Assistants

Apoorva Venkatesh

Nikita Gahoi

Shalini Aggarwal

Bharati Sakpal
Project Manager

Bharati Sarang
Project Research Associate

Nisha Thakur
Sr. Project Technical Assistant

Vinayak Raut
Project Assistant