Interactomics: Basics and Applications
Prof. Sanjeeva Srivastava
Dr. Sanjay Navani
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 53
The Human Pathology Atlas: A Pathology Atlas of the Human Transcriptome – II

In last lecture you got a 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 3 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 lecturer and continue discussion about Human Pathology Atlas project. At this time of big data, there are a 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 a total of 11000 patients for about 9600 patients, which was the study pool. We looked at the global gene expression patterns, for all the protein and coding genes.

Now, here you have to be a bit careful; I am using these words protein and coding genes what I am basically telling you is only 1 protein per gene. I am 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 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.

Student: Sir, how are you getting all these different normal tissues of healthy?

From healthy people.

Student: Yeah.

So, I tried to define healthy; I mean one source was autopsies.

Student: Ok.

And the other source was people who had been biopsied, but who were non cancer that was the control that was used.

Student: Ok.

Which so, every biopsy did not did not expect a diagnosis of cancer they were.

Student: Yeah.

Some of those patients and some of them we were done to rule out a cancer.

Student: Right.

So, therefore, they were not entirely normal.

Student: We do not know (Refer Time: 03:33), there is no way here limited number of patients you might to do a biopsy without a cancer right.

Well they the swedes had it, they had a biobank; very well maintained bio bank I have to say. So, it was a struggle and that is why.

Student: They are introducing potential biases.

Of course.

Student: Or if called autopsy of a healthy person.

Of course, sure. But it was when there was no pathologic issue, seen in that tissue under the microscope or core or correlated.

Student: Yeah.

I am sure the patients I am sure must have had many other problems.

Student: Yeah.

So, normal I mean all of us have many problems.

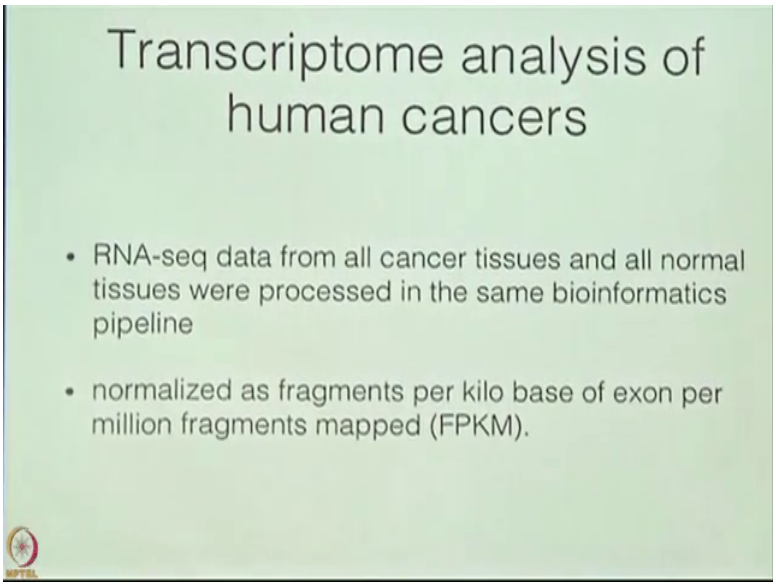
Student: Yes.

So, we are not normal to. So, that was the best that was possible under that.

Student: Yeah.

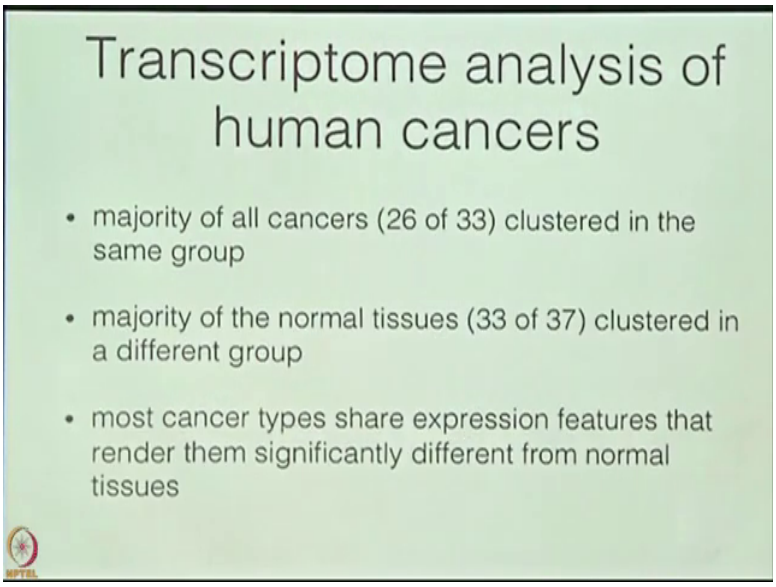
Circumstance. And all the RNA seq data both from the cancer, as well as the normal tissues, they were processed in the same pipeline.

(Refer Slide Time: 04:27)



And they were given normalized according to the FPKM. So, that was how it was expressed.

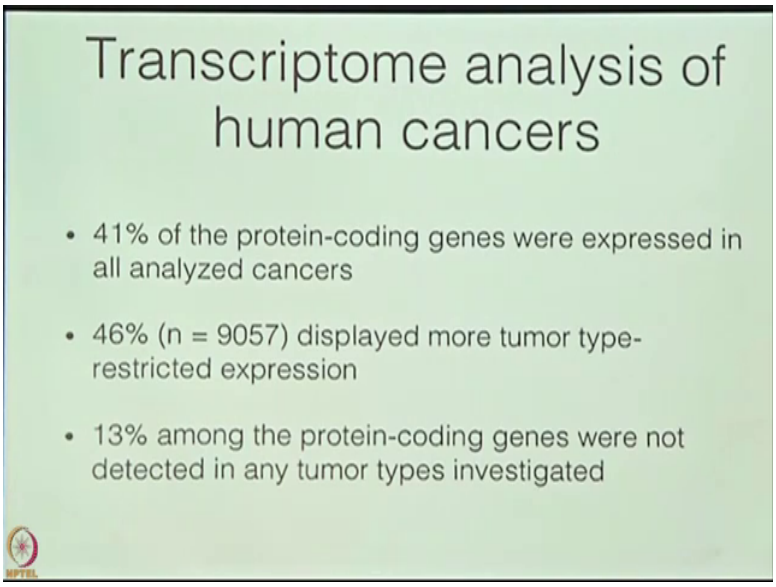
(Refer Slide Time: 04:36)



Then 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 a different group.

And the conclusion was of course, of 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: 05:08)



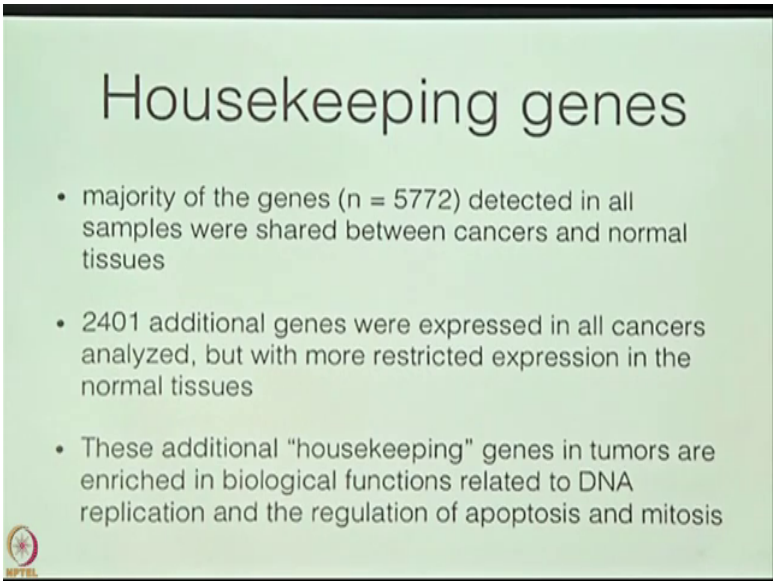
Out of all the protein encoding genes 41 percent were present in all cancers. So, a breast cancer was not necessarily that different from gallbladder cancer; there was a large overlap.

Secondly, 46 percent and this I consider 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 percent of the protein and coding genes we are not seen at all.

Now, that is a big question mark where are those genes you have 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

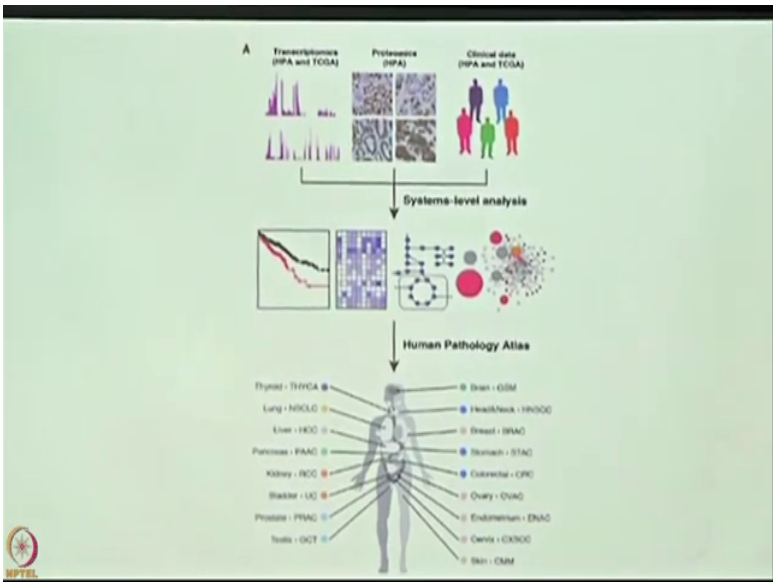
only before birth? But I will not get into that. Now the housekeeping genes, were detected in all samples both cancers as well as normal tissues.

(Refer Slide Time: 06:47)



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 than the cancer cell. But what we cannot forget is they also present in normal tissues that was how the program worked out.

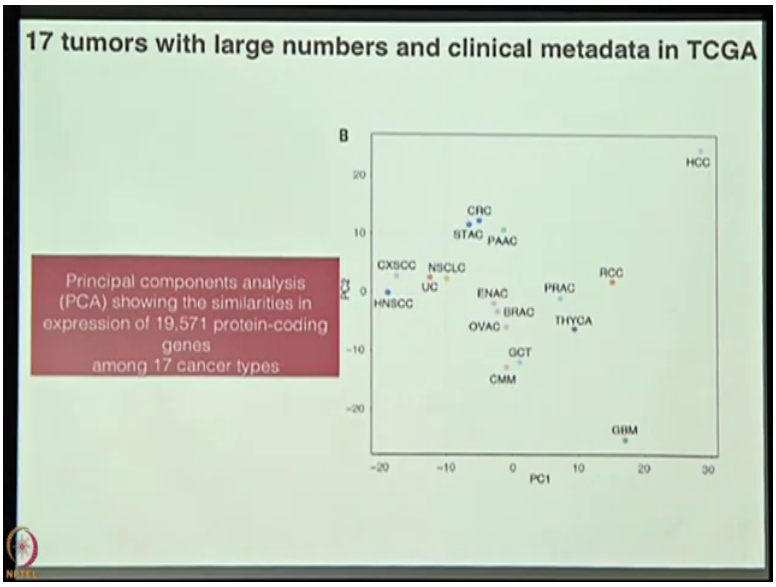
(Refer Slide Time: 07:11)



Student: How are you determined that (Refer Time: 07:17).

So, the transcriptomics was taken from The Cancer Genome Atlas and the HPA the HPA contains the immunohistochemistry 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 you 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: 07:47)

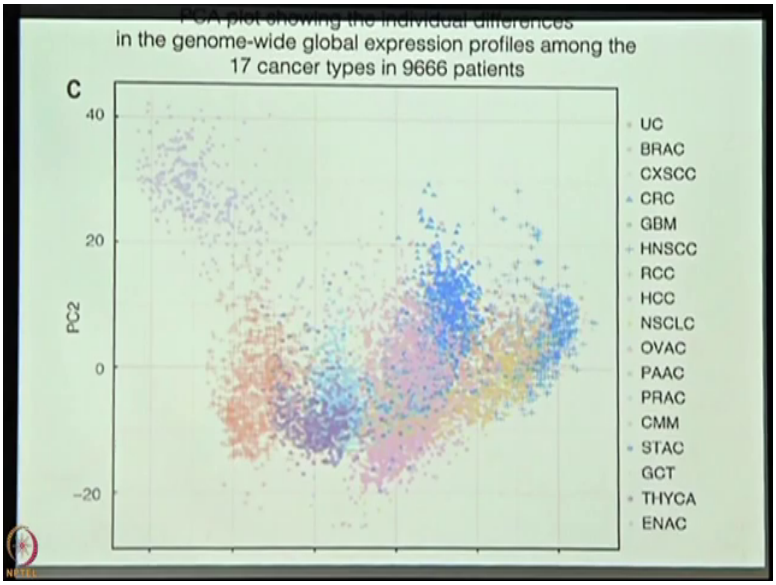


We then narrowed the search to 17 tumors with large numbers. And clinical metadata from the original 20, 33 we came down 37 we came down to 17; because these cancers had adequate numbers. And as you can see these cancers, which are grouping here are gastrointestinal cancers; this is colorectal carcinoma that is stomach adenocarcinoma that is pancreatic adenocarcinoma. This group which is coming here are the squamous cell carcinoma.

So, its 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. There were only two very far outliers; one was hepatocellular carcinoma very different from everything else and

glioblastoma multiforme, which is a brain tumor which looked quite different from everything else.

(Refer Slide Time: 09:01)



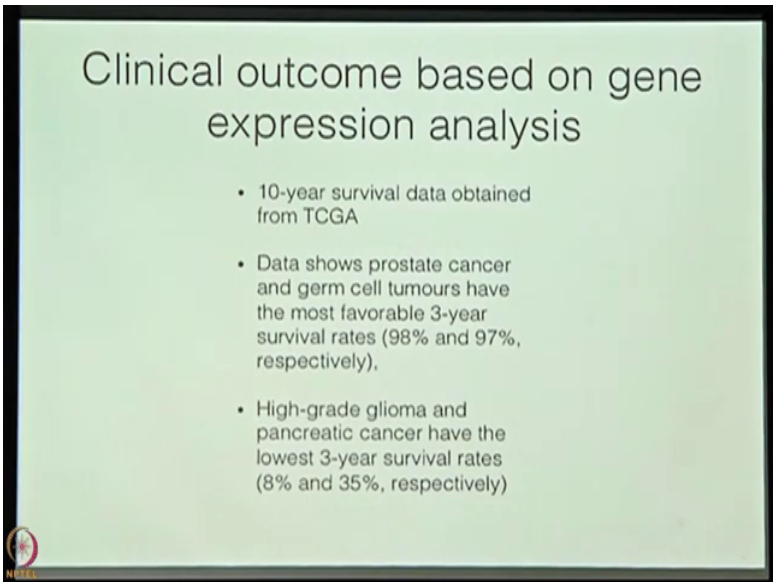
The surprises are yet to come. When we look at this each color and each figure corresponds to a particular type of cancer. You can see how much variation there is in each individual cancer; that means, even though people like me say this is a moderately differentiated or a great two hepatocellular carcinoma.

And you take two of those they look different. Not only that you will see that there is also a spill over into other sides 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 its parent cancer.

Student: Hm.

Which raises a question in my mind which is so, far I have not been able to answer if that is the case then why does it look like that.

(Refer Slide Time: 10:04)



So, the 10 year survival data also was available on The Cancer Genome Atlas. It shows prostate cancer and germ cell tumors to have the most favorable 3 year survival. That is 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.

Student: (Refer Time: 10:27) know for that. So, the TCGA specifically focus on people end staged diseased right.

Yeah.

Student: So, you I think their's choose statement from people who where end stage diseased.

Yeah.

Student: It may not be furious it.

I do not.

Student: Done like that you know.

Yeah, I do not know whether its only end stage disease; because the RNA seq data on the.

Student: Yeah.

On the genome atlas is obtained at the point of diagnosis and they follow to the event of death which is what interested us; but and this is where your point is very valid, they do not say that he died of prostate cancer. They just say that he died now it may have been a myocardial infarction.

Student: Um.

But because we had an endpoint of death.

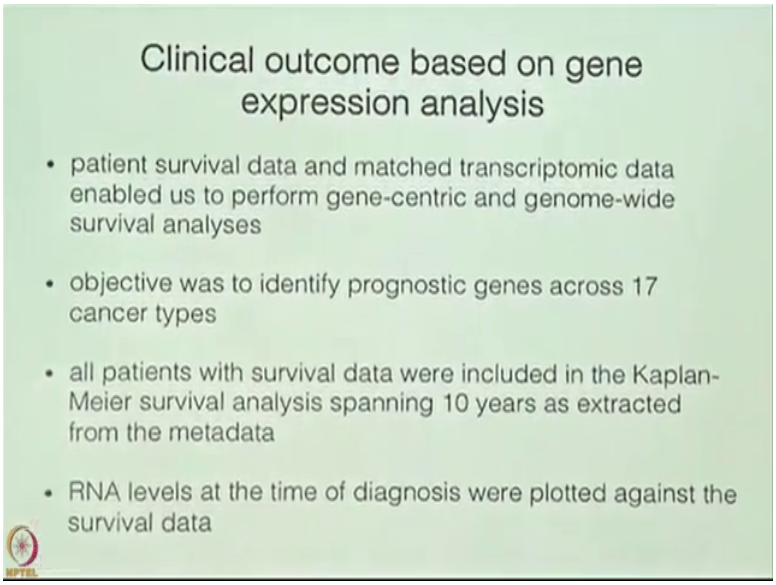
Student: As a.

And we had the RNA seq levels at the start.

Student: Um.

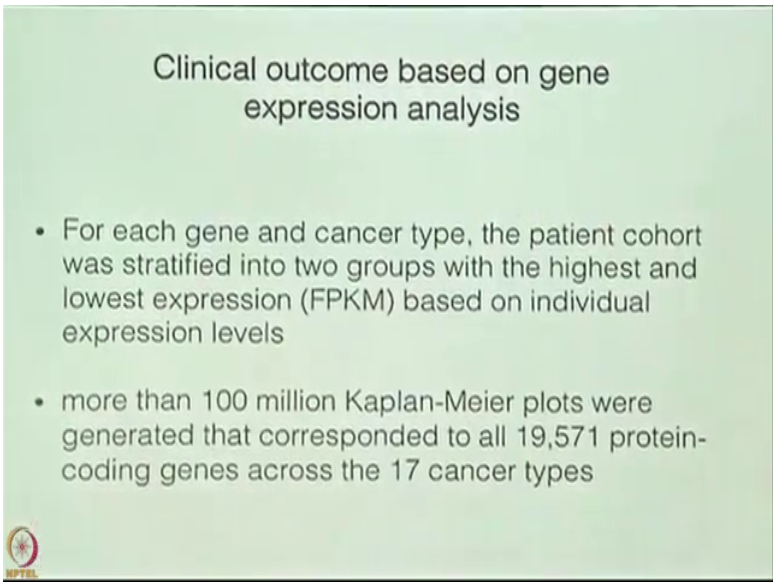
That was the reason for using this data. So, what we did was actually a huge exercise of Kaplan Meier curves. Which has looks for survival for each gene for RNA seq data from each gene and the RNA levels.

(Refer Slide Time: 11:50)



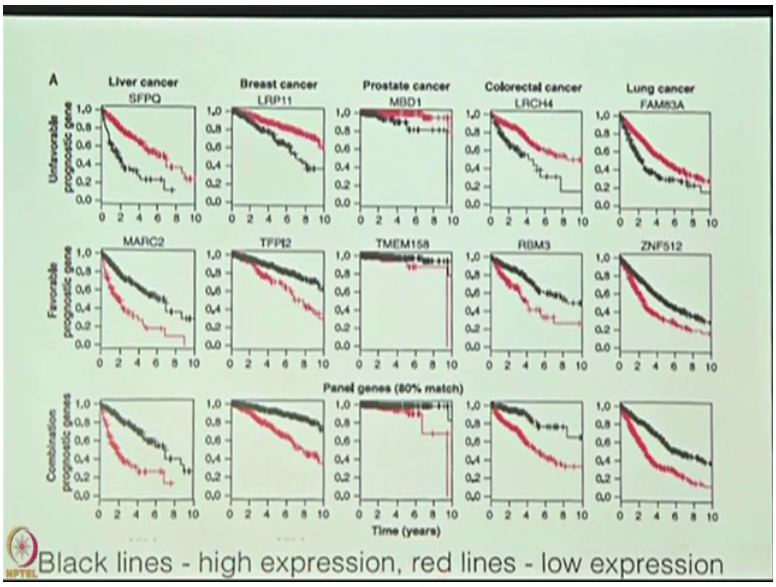
At the time of diagnosis, were plotted against the survival data that is something I was just saying.

(Refer Slide Time: 11:56)



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 is not a good gene that is finally, what we wanted to say. And for this exercise there were more than hundred million Kaplan Meier plots that were generated; I do not think anybody saw all of them it they was just the machines ok.

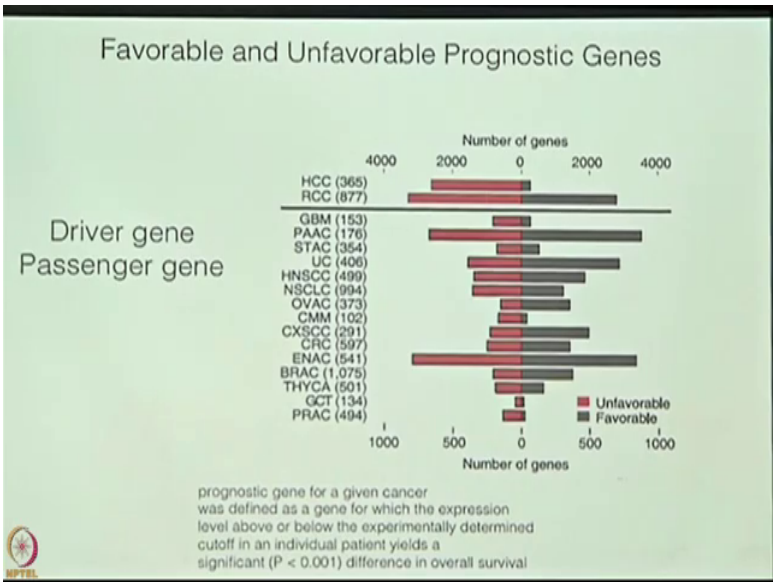
(Refer Slide Time: 12:41)



So, let me give you an example of what favorable and unfavorable genes we found. So, let us look at the different cancers on top right here black lines mean high expression red lines mean low 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 mark 2, the events seem to be happening more slowly. If that is present and therefore, that is a favorable prognostic indicator. Finally, if you combine with a panel that is what you get.

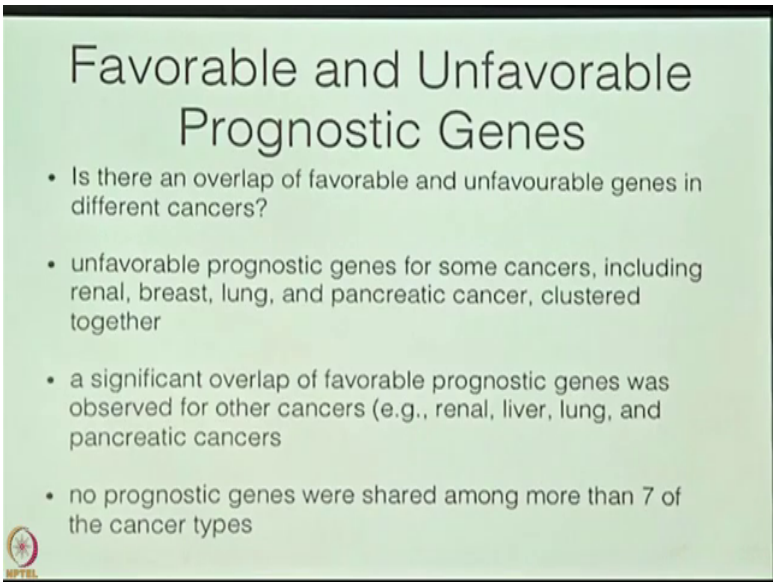
(Refer Slide Time: 13:37)



Let me put it in a different way. The number of prognostic genes were classified into favorable. And unfavorable as I just told you. Hepatocellular carcinoma and renal cell carcinoma, which are displayed at the top at 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 chart there P with less than 0.001.

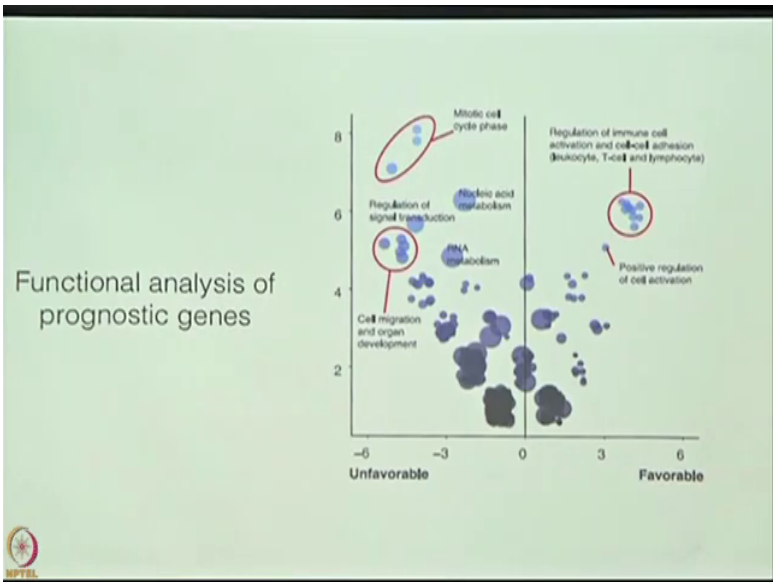
(Refer Slide Time: 14:37)



Now, in favorable and unfavorable genes; did we find more favorable same favorable genes in more than one cancer and unfavorable genes and more than one cancer? Yes, there is a big overlap. Some cancer 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 did not get it across the board saying favorable and unfavorable for everything.

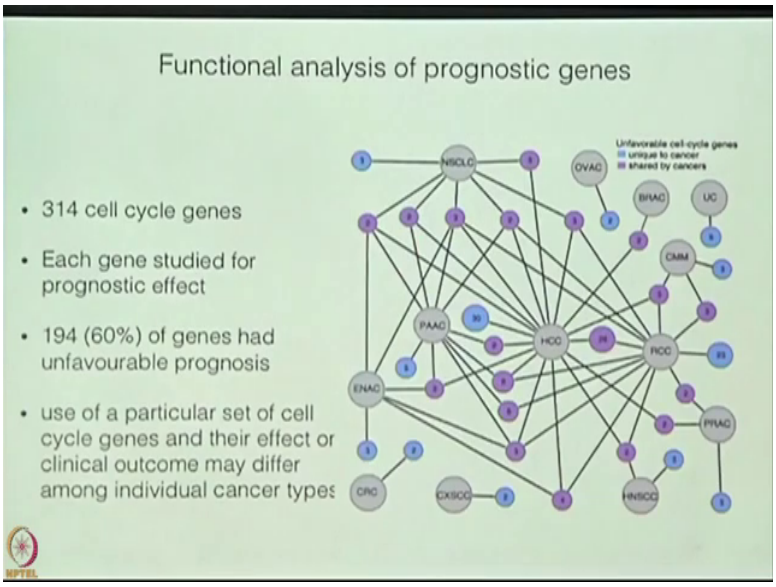
(Refer Slide Time: 15:28)



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 have 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. This because the mitotic cell cycle was found to be significant; all 314 cell cycle genes was studied because it was that significant.

(Refer Slide Time: 16:11)

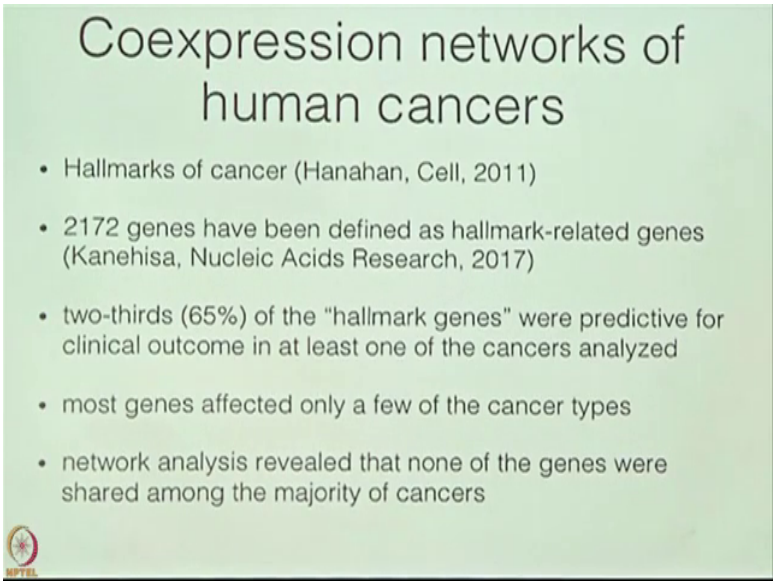


And each gene was studied separately for a prognostic effect. Now, if you ask me for this was a big surprise for me. You know what, my concept as a diagnostic pathologist is for mitotic index you either count the mitosis on this slide or there is a marker called K i 67 MIB-1 which everybody show as by.

But which does not work in all cases. So, that is just 1 prognostic gene from the cell cycle. Therefore, all cell cycle genes may not apply to all cancers. Now, 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.

Its the biggest work of its kind up to date. When we studied those hallmark genes in our data two thirds of them 65 percent of them were predictive for the clinical outcome.

(Refer Slide Time: 17:39)



In at least one cancer 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: 18:07)

So, the next step was to take a cancer example from lung and to see you see what I am in case you have 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 us go back and say what did other people say about it.

So, that is 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 have done all this stuff we have done all this data analysis and we say we have got something. Now is this true or are we barking 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 though they said are the hallmark genes and what we are saying is it matching? So, 65 percent of them matched. So,

they want 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 began to think then maybe they did not get all of them. And the network analysis showed that none of these genes were shared by the cancers, which may be meant that they were looking for specific genes in individual cancers not necessarily all the genes. So, am I clear is it better now?. 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 do not get carried away by that let us try and understand what it means. If a gene is circled these are all genes I am sorry, I could not enlarge it enough for you to read actually what is written, but that is the name of a gene in there. And if its got this red circle around it, it means that it was the hallmark paper that first said it and we studied it along with others.

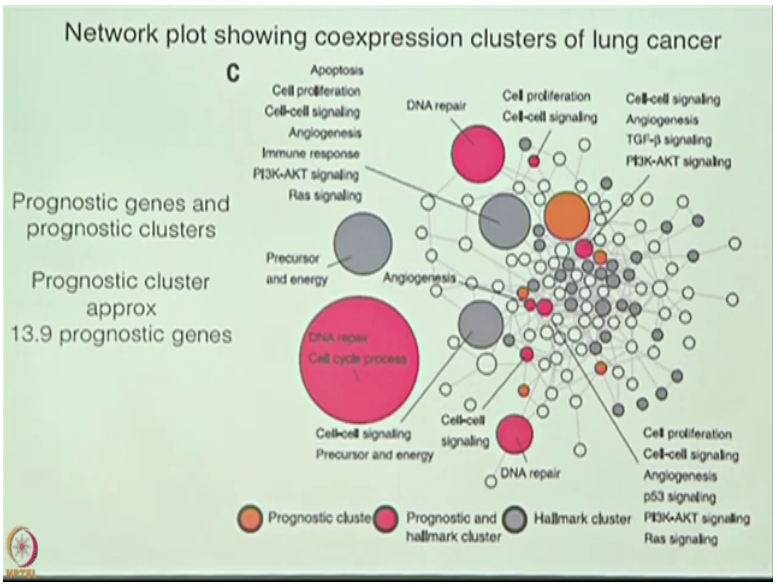
If it does not have the red mark, it means that it came up as a prognostic gene in our study. So, if you look at the middle, these are called the hub genes right at the center. And they have a great up all of them were thought to be prognostic they all came up as prognostic.

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, its a speculation that when prognostic genes affect a cancer that 50 of them a 100 of them or 500 of them, who are the really bad guys who are the drivers and who are the passengers.

So, its tempting to speculate; that these guys in the middle are the drivers and the ones at the periphery are the passengers; but its only a speculation that is where it should stay for now. Then and on the beautiful diagram it looks like popcorn does not it. So, what the statisticians did was they said you are talking about these prognostic genes 11 gene and 20 genes. You have got there you talk about a prognostic cluster all of which are closely related.

There are so many genes, which are related, which are doing DNA repair cell cycle processes you get them together make it a prognostic cluster. So, that is what we did.

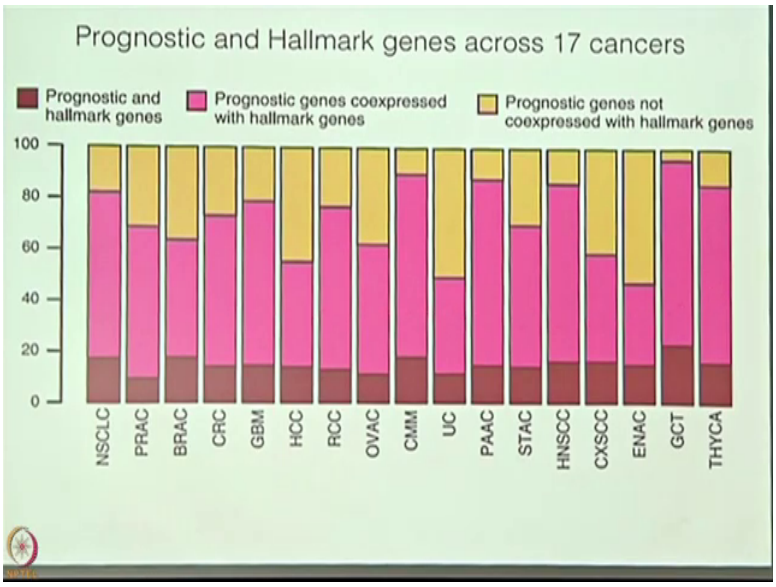
(Refer Slide Time: 22:34)



We made this a prognostic cluster 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 gray ones are the hallmark clusters they were published 10 you know I am sorry 7, 8 years ago. And the prognostic and hallmark clusters are our work, which has is superimposed on the hallmark we are agreed that these are all important.

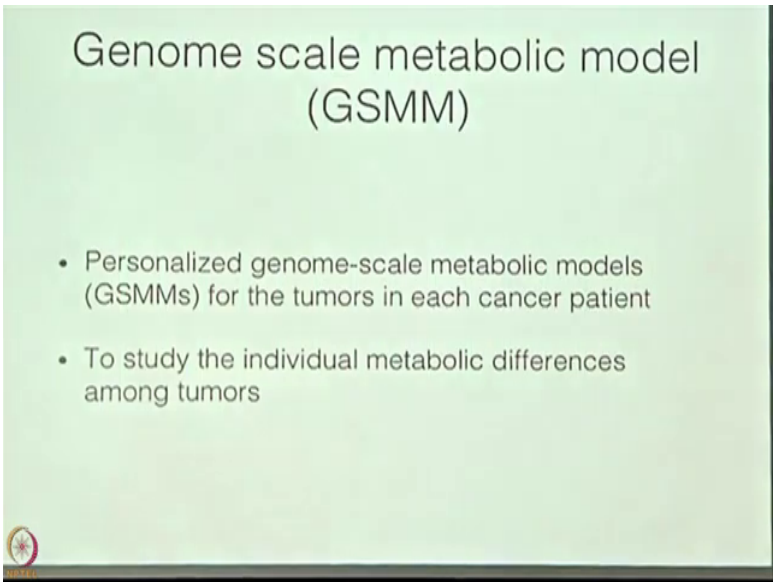
And then there is a third group, in which we say this is a prognostic cluster, but hallmark has not talked about it.

(Refer Slide Time: 23:34)



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: 24:07)



Now, with all this information, bringing it to the end of almost my talk. Is it possible to generate a personalized model for an individual cancer for treatment? Something that is referred 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 us 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 are you with me or no?

Student: (Refer Time: 24:55) sir no on genes equal to (Refer Time: 24:57) differentiating genetic differentiation need not tumor in a different origin. So, in that case a how we will develop below that particular equal to that particular marker (Refer Time: 25:15).

No, this is its 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 of the same type.

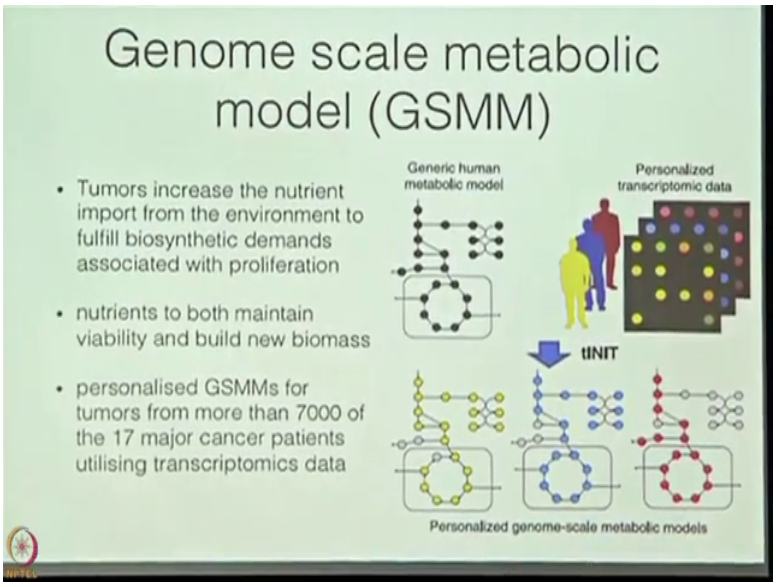
Student: Ah.

So, hepato cellular cancer, which looks the same what does the transcromic transcriptomic data say on it? That is the question, I am trying to answer right now. So, let us just finish that. Are you with me?

Student: Yes.

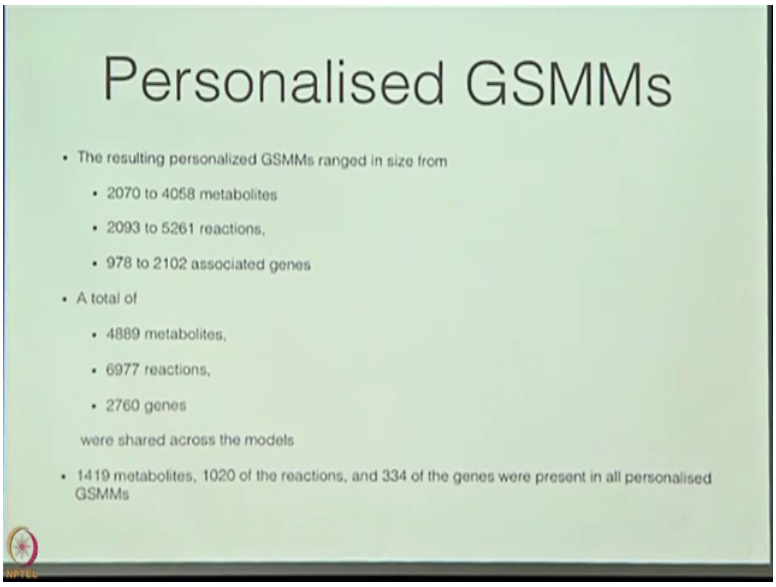
So, that is we carried out a personalized genome scale metabolic model for tumors from more than 7000 of the 17 major cancer patients.

(Refer Slide Time: 26:07)



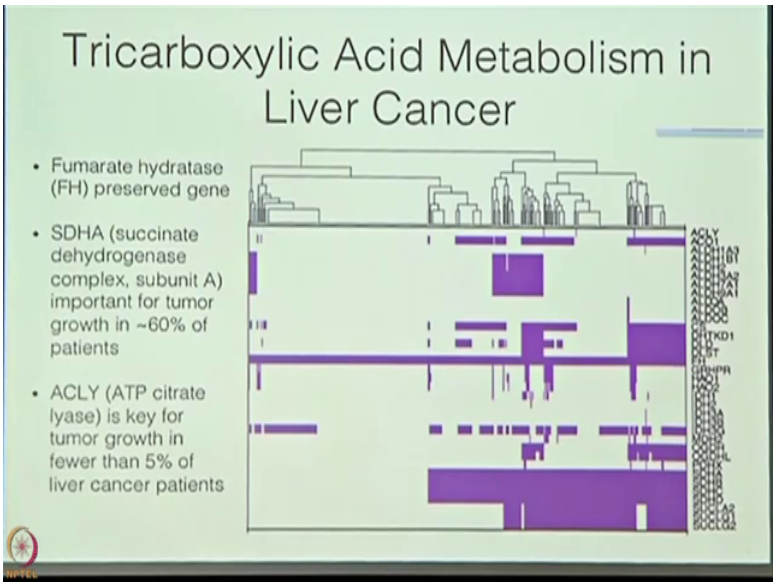
And we expected something different; because cancer cells pull in more nutrients from the surrounding and they build up more of a biomass.

(Refer Slide Time: 26:20)



This is some of the statistics that it throughout. I will not go into all that except to say. That 1400 metabolites 1000 reactions and 334 of all the genes are present in all the personalized models it was common.

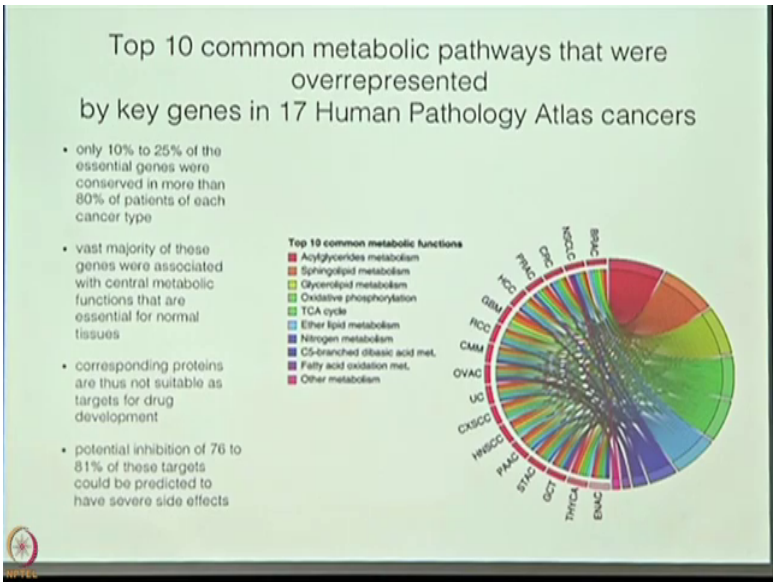
(Refer Slide Time: 26:41)



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 a fumarate hydratase was found in all liver cancer tumors.

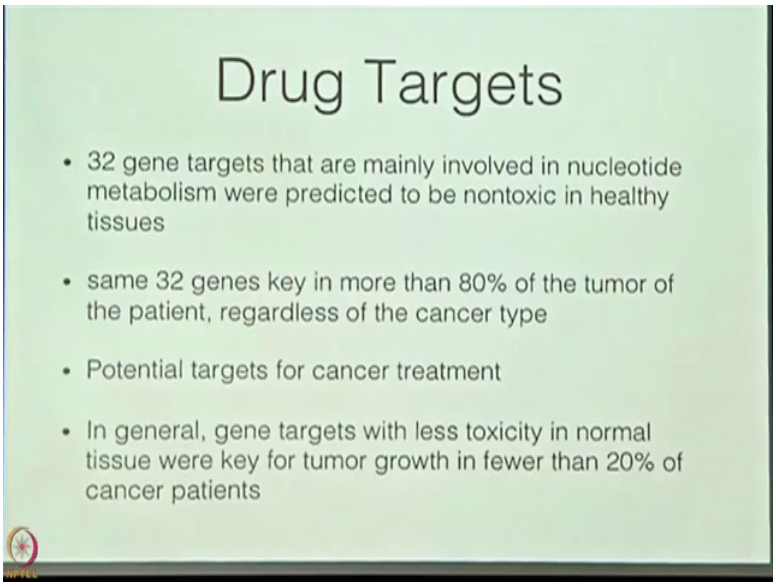
Then I want to tell you that ACLY right there on the top; see those small bars there it was found in less than 5 percent on liver cancer patients. And finally, succinate dehydrogenase complex unit A was formed in 60 percent. The point I am trying to make is, that there is sufficient difference among cancers of the same time, which underscores the need for a personalized model ok.

(Refer Slide Time: 27:36)



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 they are also occurring in normal cells.

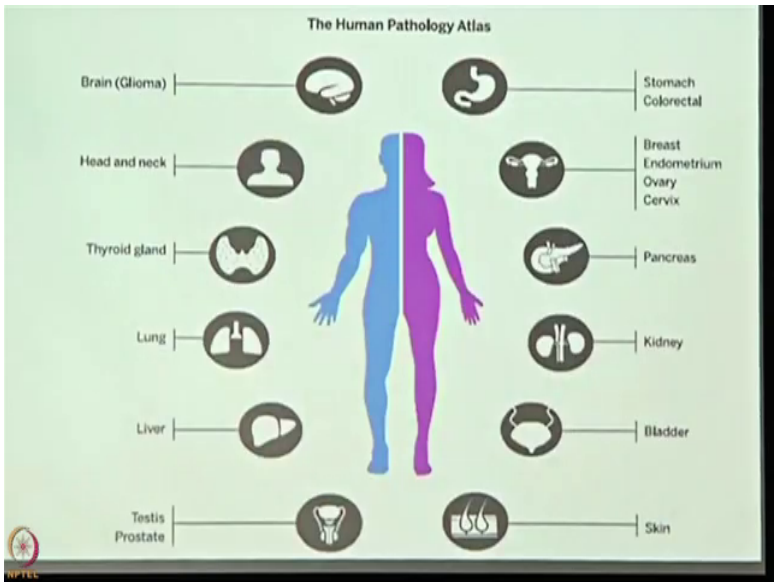
(Refer Slide Time: 28:16)



Therefore, if you give anything to these patients as I outlined before, there was a possibility that about 80 percent of these targets would have side effects.

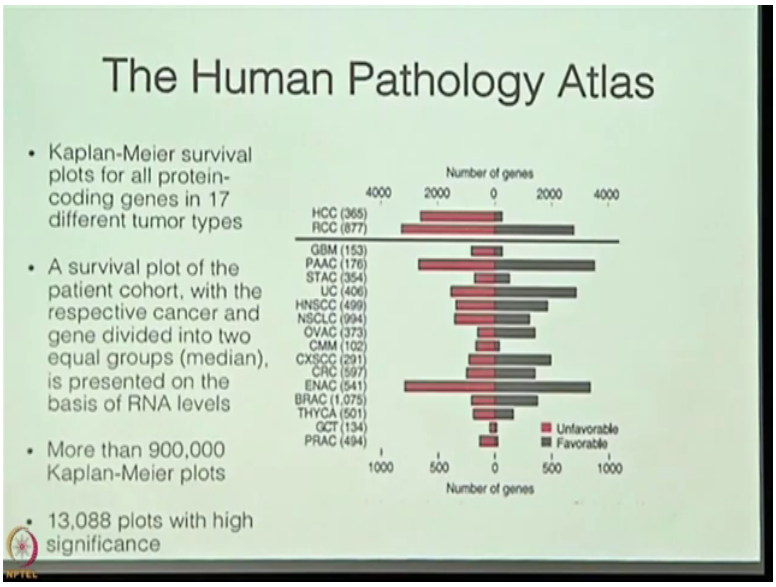
So, you hit the cancer, but you also hit the normal cell. And now the concept of what chemotherapy does to a patient? I think you can begin to appreciate why they have so many problems ok. Now, the last point 32 gene targets that were mainly involved in nucleotide metabolism look like potential targets. They are expressed in more than 80 percent of the tumor of the tumors of the patient regardless of the cancer type. And they are potential targets because they will not affect the normal cells ok.

(Refer Slide Time: 29:17)



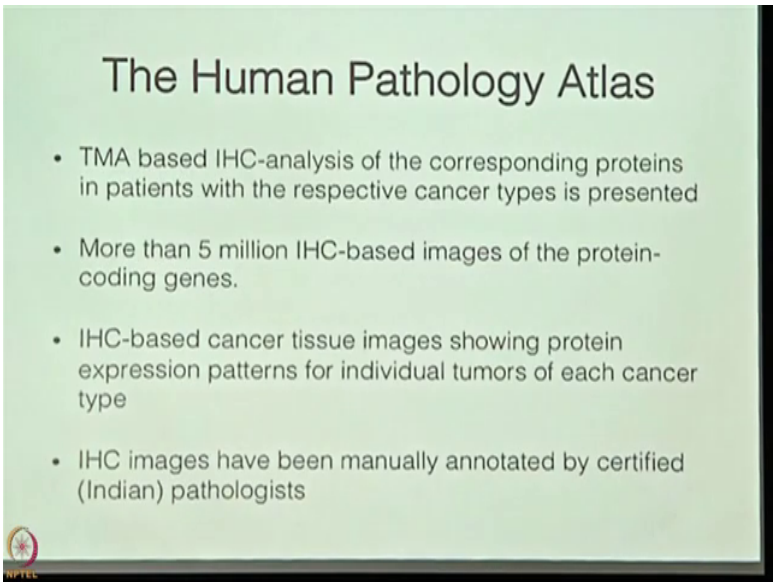
So, if you go to the protein atlas there is a new section called its well 6 months old now. But the Human Pathology Atlas which will give you access to all the Kaplan Meier plots. You want to see anything the significant plots the insignificant plots all the data is up there.

(Refer Slide Time: 29:35)



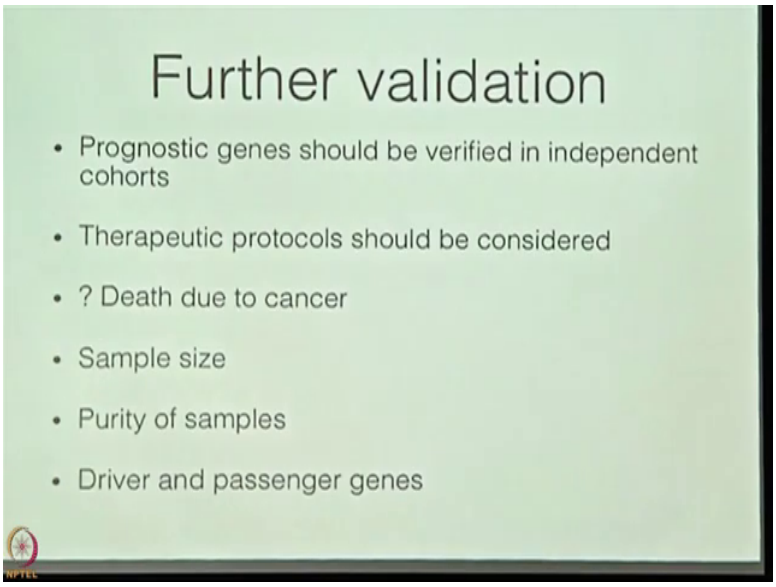
You will also get the survival of the patients if you wish to check that.

(Refer Slide Time: 29:40)



More than 5 million IHC images for cancer can be seen the most of it annotated bias here in India.

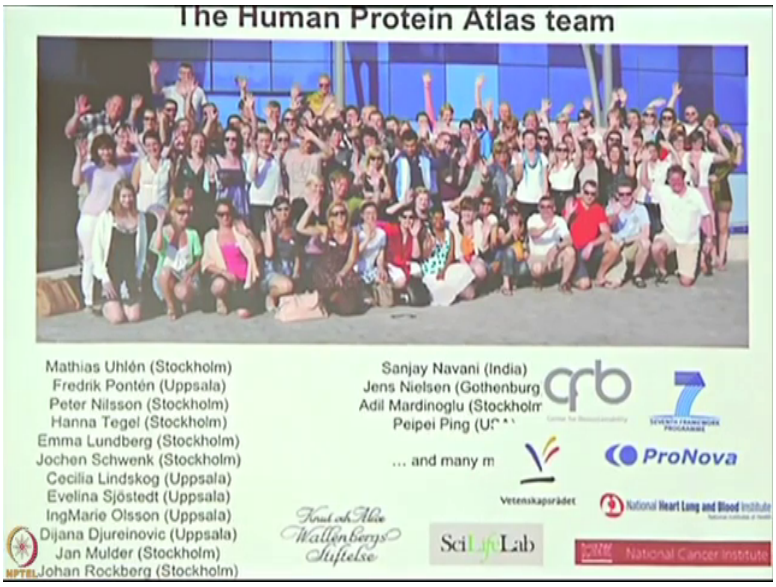
(Refer Slide Time: 29:54)



A few points just to leave in your mind as I finish. These prognostic genes as we heard in the morning, they have to be verified in an independent cohort. Secondly, the death which was a question that Joshuva raised earlier, we are assuming that the death was due to the cancer, but we do not 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 these prognostic genes are all causing the cancer or some guys doing it and the others are just following and what are those? I will just like to leave with a thanks to the people who made it possible for us to do the job here in India.

(Refer Slide Time: 31:09)



In particular this manner right there, Mathias Uhleon whose 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 and which I 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 that was of course, a previous workup; the you know more younger people in the organization met me, I put forth my ideas everything happened I gave a presentation. And then finally, I met him for 15 minutes and he said ok, I think your idea

looks good. I have 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?

I said no, its 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 had never been done. So, new ideas were ok, if the idea is good let us give it a shot. We were supposed to do 2 million images in a year and when I first heard that, I did not know whether I should say yes, but I did.

But the moment I said yes, he said like a good senior person. He said I am sure you can do it Sanjay, why do not we do a small experiment why do not; you just do 250000 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. The 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 do not think he and I knew what we were talking about; because nothing had happened. So, I went out I hired pathologists, who had never seen a single slide of IHC their life's. All this stuff is done by guys and girls who would never seen any IHC ever and why did I hire them.

Because they wanted to learn it. And these are people from here we are not talking about a different country. I just like all of you to remember that. So, after we started it we finished 250000 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 is how the whole thing happened.

So, as time has gone by, you know at one point in time the numbers used to be very important to me, 15 million images with quality control this that. All that you know everything finally, passes the only thing that is remained and which always makes me feel very good. When I think about it is that there were Indian pathologists who did not know anything about IHC, who did this job and it makes me feel very happy; because it I never expected that.

So, I just want you to remember that that its 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 its alright, thank you for your attention.

Student: Questions [FL] sir, the I mean treat where the gene that the will be (Refer Time: 35:33) genes that suggest the targets. Especially because you know number of the major targets the cancer even before of the (Refer Time: 35:41) DNA in the (Refer Time: 35:41).

Yeah.

Student: 5 for USA (Refer Time: 35:45).

Yeah.

Student: So, that the those show up I mean (Refer Time: 35:51).

They showed up. In fact, there was. In fact, they were the main group they were the main group.

Student: Right (Refer Time: 35:57).

So, it the cell cycle genes, because there was a group that separated from the rest

Student: Yeah.

All the 314 genes in the cell cycle, were evaluated on an individual basis.

Student: Ah.

And 60 percent of them showed a correlation with the cancers, unfavorable prognosis; but not all of them affected the same cancer.

Student: Hm.

So, there were different genes in the cell cycle itself.

Student: Hm.

Which had different impacts on the cancer.

Student: Yeah.

So, they were very much a part of the group also published.

Student: (Refer Time: 36:35) which actual to the x 1.

Yeah.

Student: Purely a nonvegetarian can found.

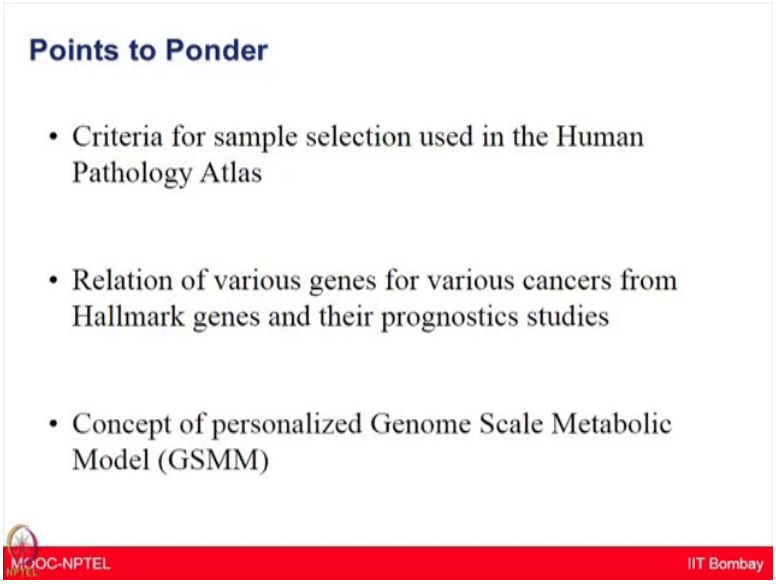
That was also described by the hallmark group. So, all of that is confirmed and now there is and by our study, I mean confirmed by our study. And now there are additional prognostic genes.

Student: Yeah.

Which we have brought up, which we feel should be evaluated further.

Student: Yeah.

(Refer Slide Time: 37:58)



So, I am 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. Its 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 based profiling to provide a standalone resource for cancer precision medicine. I must say that you should look into this a verily enriched a 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 plots also show the clinical status for all the individuals in the patient cohorts. All the data which is presented are also made publicly available in a very interactive open access database, to allow ones 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 bioinformatics work just from 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 will see you again in the next lecture.

Thank you.