Introduction to Proteogenomics

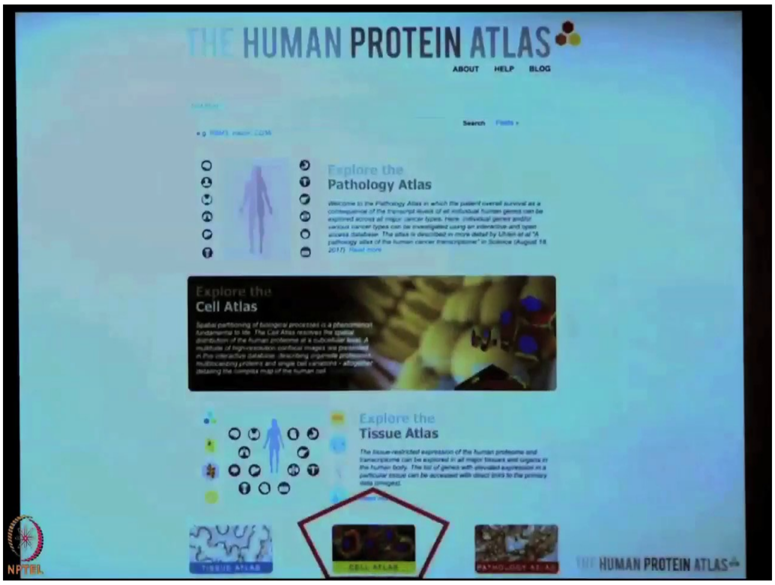
Dr. Sanjeeva Srivastava
Dr. Fredrick Ponten
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 37
Human Protein Atlas-II

Welcome to MOOC course on Introduction to Proteogenomics. In today's lecture which the second lecture of Dr. Fredrick Ponten we are going to continue the discussion about the Human Protein Atlas project. Dr. Ponten will talk about two other domains of HPA, the cell atlas and pathology atlas. The cell atlas provides high resolution insights into the spatio-temporal distribution of proteins within human cells whereas; the pathology atlas contains mRNA and protein expression data from 17 different forms of human cancer.

He will also tell us about integrated omics aspect regarding each of the 17 cancer types with mRNA and protein expression data including the genes associated with prognosis. He will also talk about protein localization data, which is a part of cell atlas and how it is derive from more than 60 different cell lines. Professor Ponten will further talk about the brain atlas project which will give information of the region specific gene expression of human, pig and mouse. Finally, he will talk about the blood atlas and how genes or blood cell types are related to each other. So, let us welcome Professor Ponten for his today's lecture.

(Refer Slide Time: 01:49)



THE HUMAN PROTEIN ATLAS

ABOUT HELP BLOG

Search Fields ▸

Explore the
Pathology Atlas

Welcome to the Pathology Atlas in which the patient overall survival as a consequence of the transcript levels of all individual human genes can be explored across all major cancer types. Here, individual genes and/or various cancer types can be investigated using an interactive and open access database. The atlas is described in more detail by Uhlen et al "A pathology atlas of the human cancer transcriptome" in Science (August 18 2017) Read more

Explore the
Cell Atlas

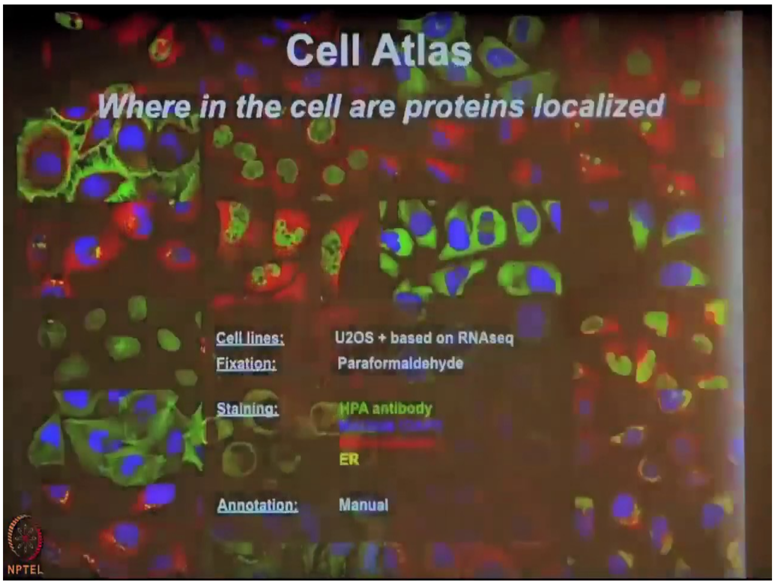
Spatial partitioning of biological processes is a phenomenon fundamental to life. The Cell Atlas resolves the spatial distribution of the human proteome at a subcellular level. A multitude of high-resolution confocal images are presented in the interactive database, describing organelle proteomes, multilocalizing proteins and single cell variations - altogether detailing the complex map of the human cell

Explore the
Tissue Atlas

The tissue-restricted expression of the human proteome and transcriptome can be explored in all major tissues and organs in the human body. The list of genes with elevated expression in a particular tissue can be accessed with direct links to the primary data (images).

So, the next part is the cell atlas or the sub cell atlas.

(Refer Slide Time: 01:55)



Cell Atlas

Where in the cell are proteins localized

Cell lines: U2OS + based on RNAseq

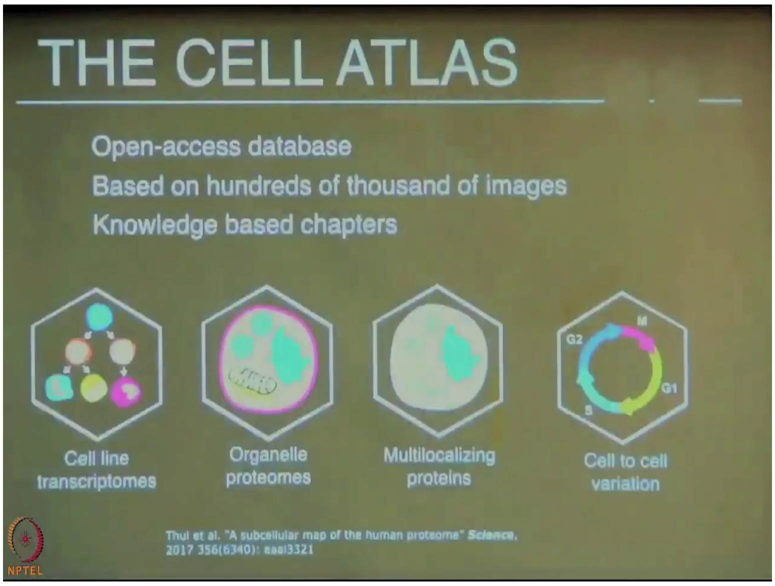
Fixation: Paraformaldehyde

Staining: HPA antibody
[illegible]
[illegible]
ER

Annotation: Manual

And, this is image based on immunofluorescence and here we use RNA seq to select the cell line which has a highest expression of each gene and we also use U2OS cell line for every just have a standard throughout all antibodies and genes. And, here we then the this is also manual annotation at this point, but were working on more automated annotation and of course, these immunofluorescence I think everybody agrees produces extremely beautiful artistic images, but they are also very informative and I will just show you two examples, it is a transcription factor.

(Refer Slide Time: 02:39)



THE CELL ATLAS

Open-access database
Based on hundreds of thousand of images
Knowledge based chapters

Cell line transcriptomes

Organelle proteomes

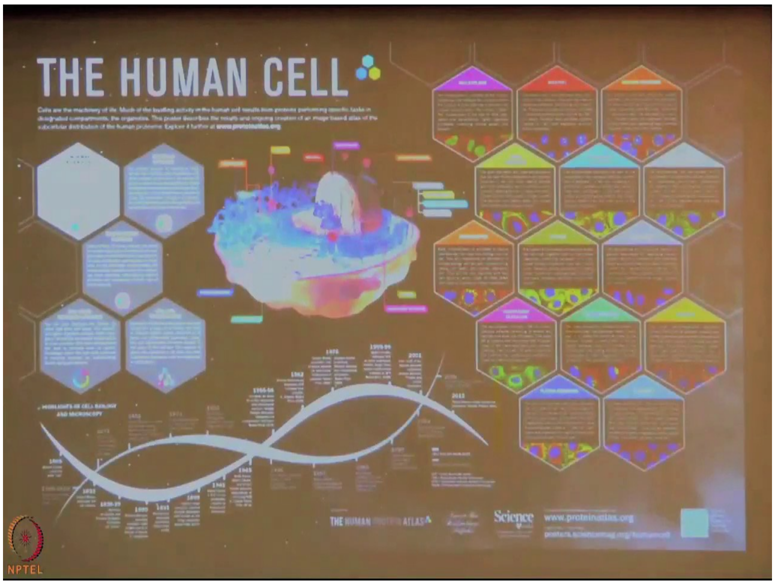
Multilocalizing proteins

Cell to cell variation

Thul et al. "A subcellular map of the human proteome" *Science*, 2017 356(6340): eaal3321

But, also the cell atlas we have produced these knowledge based chapters where you can read about the cell line transcriptomics, the different organelle proteomes, multilocalizing proteins which are extremely interesting the proteins which are which then can be in both for instance golgi and maybe the ribosomes etcetera which there is a lot more to a lot more research need to be done about. And, also about the cell to cell variation the proteins are vary during the cell cycle or show other forms of variability also very interesting proteins.

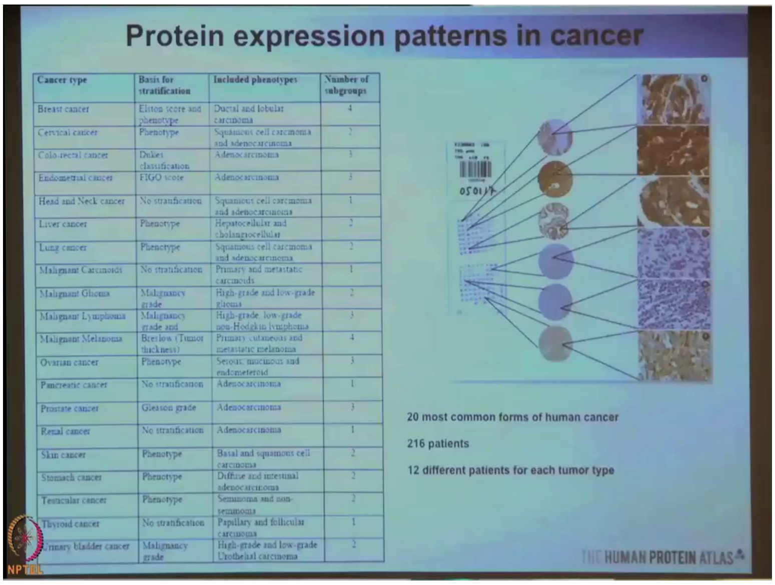
(Refer Slide Time: 03:16)



And, also here we were successful and had a paper in science and also a poster in science etcetera and these posters if you are interested you can just contact me and I will send them to you because we have lots of them left. The third part done is the pathology atlas and that is what is in a sense the being closest to me since I am a pathologist and here we started off with having millions of millions of images of immunohisto-chemically stained cancer, but not any real data more as examples of how all our different proteins are expressed in cancer.

And, what we had were tissue microarrays that included tell 12 individually different tumors reached tumor type. So, 12 breast cancers, 12 colon cancers, 12 gliomas etcetera and, trying to include the kind of prototypic types of tumors.

(Refer Slide Time: 04:03)



Protein expression patterns in cancer

Cancer type	Basis for stratification	Included phenotypes	Number of subgroups
Breast cancer	Elston score and phenotype	Ductal and lobular carcinoma	4
Cervical cancer	Phenotype	Squamous cell carcinoma and adenocarcinoma	2
Colo-rectal cancer	Dukes classification	Adenocarcinoma	3
Endometrial cancer	FIGO score	Adenocarcinoma	3
Head and Neck cancer	No stratification	Squamous cell carcinoma and adenocarcinoma	1
Liver cancer	Phenotype	Hepatocellular and cholangiocellular	2
Lung cancer	Phenotype	Squamous cell carcinoma and adenocarcinoma	2
Malignant Carcinoids	No stratification	Primary and metastatic carcinoids	1
Malignant Glioma	Malignancy grade	High-grade and low-grade glioma	2
Malignant Lymphoma	Malignancy grade and	High-grade, low-grade non-Hodgkin lymphoma	3
Malignant Melanoma	Breslow (Tumor thickness)	Primary cutaneous and metastatic melanoma	4
Ovarian cancer	Phenotype	Serous, mucinous and endometeroid	3
Pancreatic cancer	No stratification	Adenocarcinoma	1
Prostate cancer	Gleason grade	Adenocarcinoma	3
Renal cancer	No stratification	Adenocarcinoma	1
Skin cancer	Phenotype	Basal and squamous cell carcinoma	2
Stomach cancer	Phenotype	Diffuse and intestinal adenocarcinoma	2
Testicular cancer	Phenotype	Seminoma and non-seminoma	2
Thyroid cancer	No stratification	Papillary and follicular carcinoma	1
Urinary bladder cancer	Malignancy grade	High-grade and low-grade Urothelial carcinoma	2

20 most common forms of human cancer

216 patients

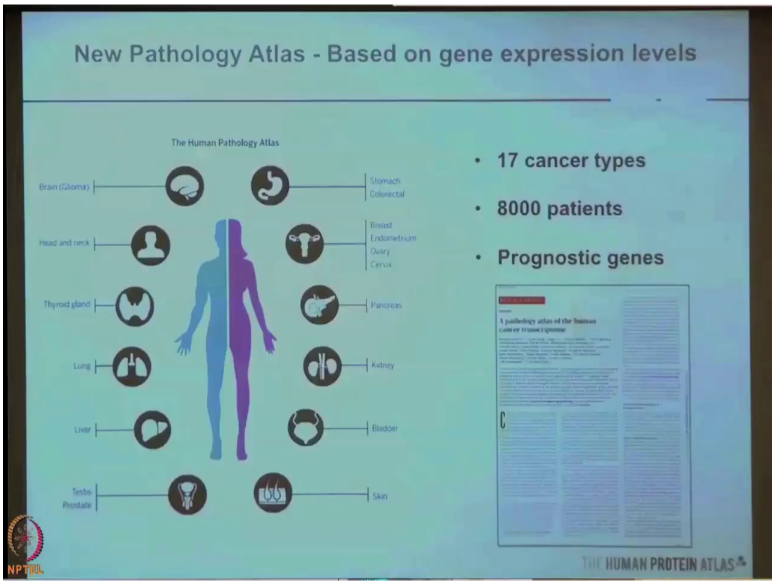
12 different patients for each tumor type

THE HUMAN PROTEIN ATLAS

So, for breast cancers we had mostly ductal cancers, but we also had a few lobular cancers etcetera; gliomas high grade in low-grade etcetera, but no real data just examples. Of course, it is shows you if like I showed you for SATB2 if a protein is expressed in very many colon cancers in no other cancers it could be a good biomarker for colon cancer.

But, we have realized we learned so much from the efforts using RNA sequencing. So, we thought we have to do something cancer based on own RNAseq and then you know about the human that the cancer genome atlas. So, we went and took all the data from the cancer genome atlas and massaged through it.

(Refer Slide Time: 04:50)

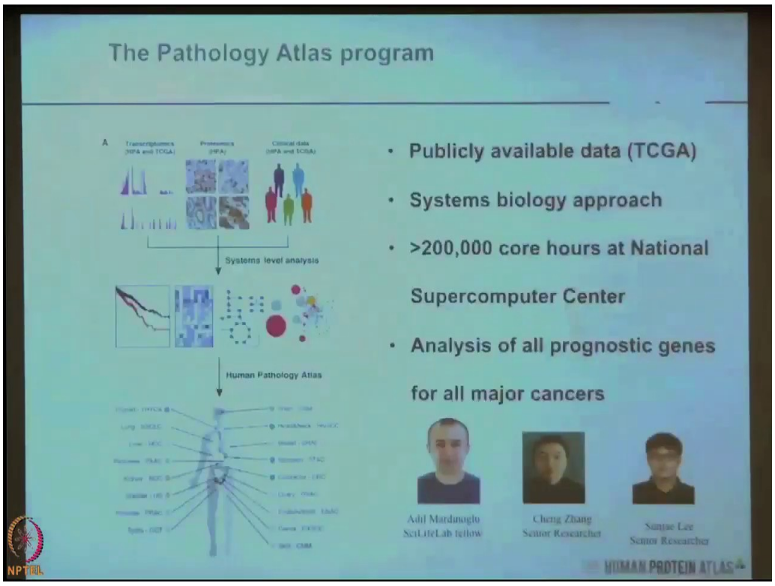


New Pathology Atlas - Based on gene expression levels

- 17 cancer types
- 8000 patients
- Prognostic genes

And, what we did then we defined the cancer types where there were at least 100 patients with full clinical follow up data and where we had then RNAseq data and that turned out to be 8000 patients. And, what our question was which genes correlate to patient's survival for these 17 different cancer types.

(Refer Slide Time: 15:15)

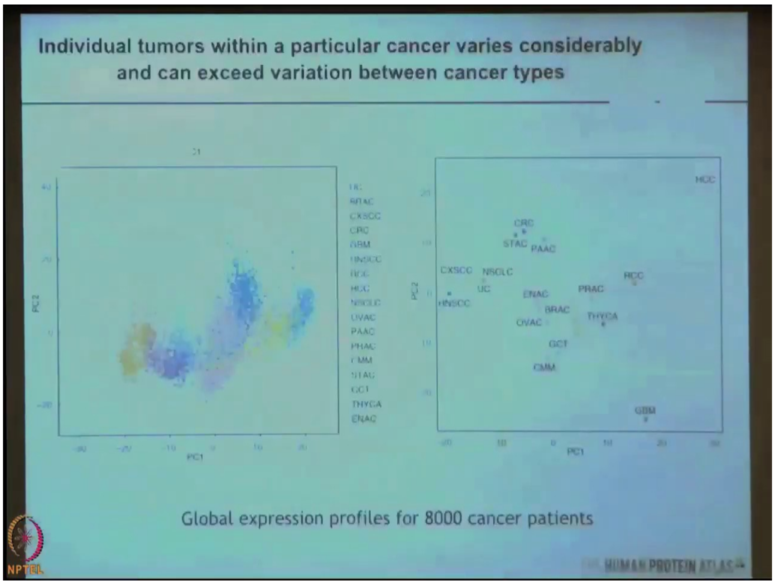


The Pathology Atlas program

- Publicly available data (TCGA)
- Systems biology approach
- >200,000 core hours at National Supercomputer Center
- Analysis of all prognostic genes for all major cancers

And an enormous amount of bio informatics work was done during about 6 months 5 months headed by Adil who is a great computer scientist and his group of bioinformaticians. Spent lots of time at the national supercomputer center had the assistance biology approach all the data was already available, but it was not kind of put together. So, we massaged it or they massage it and I will show you just two or three slides of the some of the summary of what we learned from that effort and all this data is of course, available on the protein atlas the pathology part.

(Refer Slide Time: 05:48)



Individual tumors within a particular cancer varies considerably and can exceed variation between cancer types

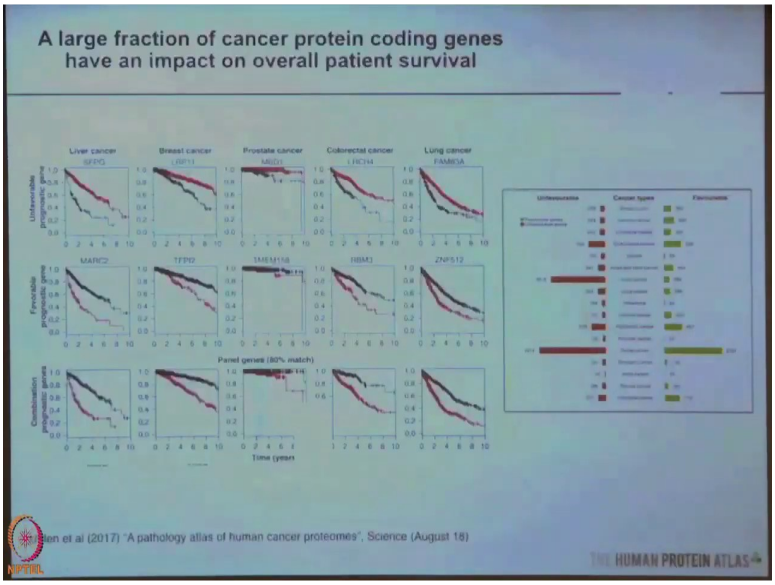
Global expression profiles for 8000 cancer patients

So, one of the more interesting things was this is all the 8000 patients and this is an all the 20000 genes and how they are expressed in a PCA plot and you can see then you can see the 17 different color codes, but what you cannot see is that the cancers form clouds which are highly overlapping and what this means is of course, that if I have a prostate cancer that can be more similar to my wife’s breast cancer than to another man’s prostate cancer.

And, so, we start to think the cancer might not be so organ of course, for surgery it is very important there is a very big difference, but maybe when it comes to more to more biological approaches to treat cancer so on, one has to think of other ways of classifying cancer not just based on anatomy. One type of cancer stands out pretty much this is hip hop to cellular carcinoma liver cancer especially type of cancer, glioma stands out a bit too, but not as much as one would expect. But, so, many of the cancers can if one looks at means this is an the mean instead of looking at all the 8000 patients.

You can actually see some clusters here are the gastrointestinal tract tumors, here is the hormone driven cancers, breast cancers and epithelial cancer, ovarian cancer which also is pretty close to prostate cancer. Here you have glioblastoma you see the cancers which have quite a lot of influence of squamous cell carcinoma etcetera. So, actually the different tumors do separate, but if you look at individual tumors they are very diffuse clouds actually.

(Refer Slide Time: 07:26)



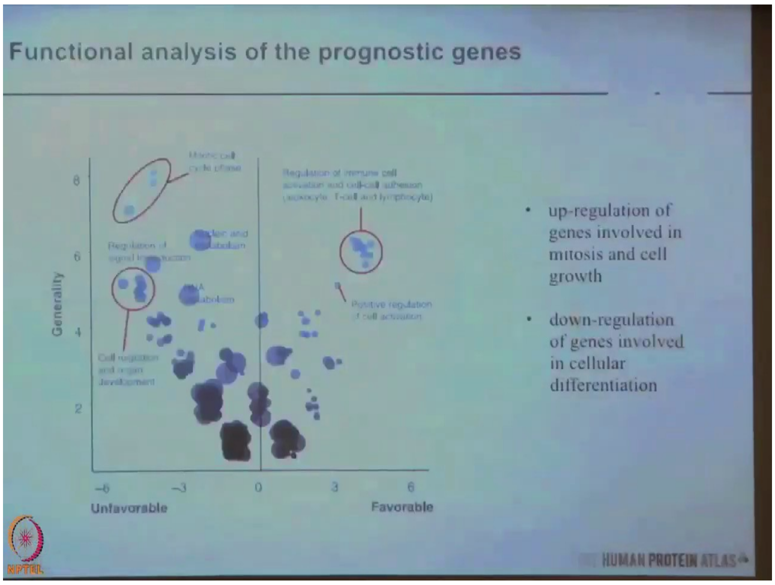
A large fraction of cancer protein coding genes have an impact on overall patient survival

...len et al (2017) "A pathology atlas of human cancer proteomes", Science (August 18)

What we also saw was that many of our genes are much higher proportion than thought of before, their expression level did correlate to patient survival and that there were quite a lot of genes that we call them favorable or good genes where high expression was associated with prolonged survival.

And, many genes were high expression was associated with poor survival and about 3 4 5 600 such both favorable and unfavorable genes for all the different tumor types. So anyone could find very if you do this capital my analysis very highly statistically significant separation between high expression and low expression.

(Refer Slide Time: 08:08)



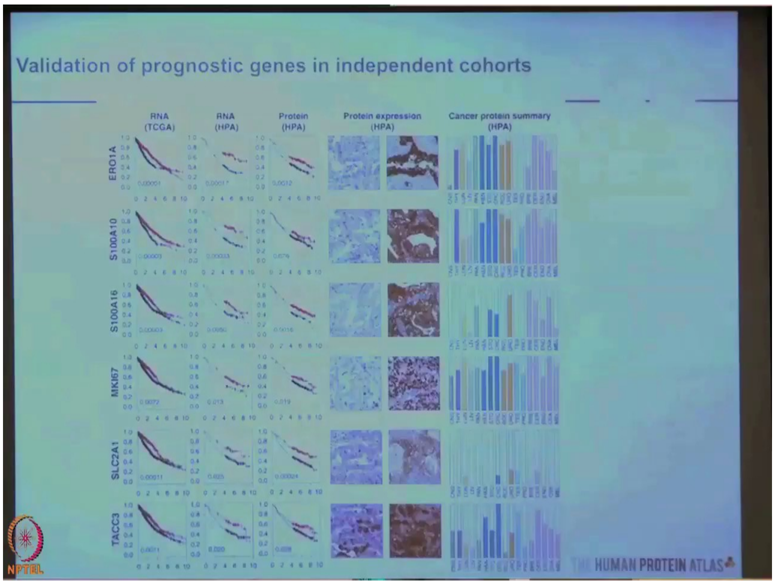
Functional analysis of the prognostic genes

- up-regulation of genes involved in mitosis and cell growth
- down-regulation of genes involved in cellular differentiation

What about what the function of these genes and this is very generalizing very much I know that, but if you look at the ones which are unfavorable genes what do they what do they encode for what type of proteins. Will they code for proteins which are involved in cell cycle regulation, cell cycle progression, cell growth very logical that would then correlate to poor prognosis.

While the proteins that are correlated where high expression is correlated to good prognosis or proteins involved in cellular differentiation and immune response. Also, quite expected from the pathology point of view, but never before shown on the kind of transcriptomics or this global view of all our genes.

(Refer Slide Time: 08:52)

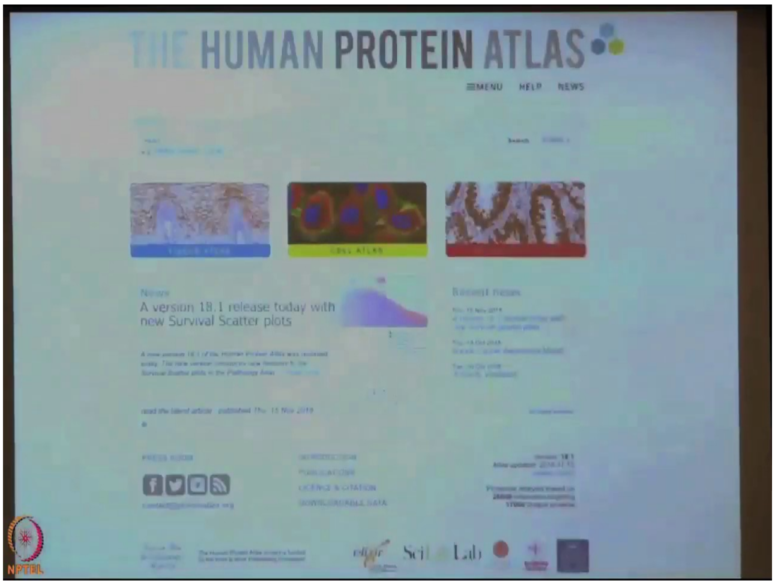


Validation of prognostic genes in independent cohorts

And, lastly just from this paper which then also was published in science is that what is extremely important when you do such a big kind of discovery you all know this about false discovery rates and everything is that you have to validate your results in totally independent cohorts with preferably independent methods.

So, we this is lung cancer and this is our primary data and then we from the TCGA and then we had our own cohort of 400 lung cancer patients and only the candidate proteins that stand a tested in a totally independent cohort are the ones that are actually could be clinically interesting and relevant. And, we could also show that, that was true in the protein level by using tissue microarrays from tumors in these in large patient cohorts ok.

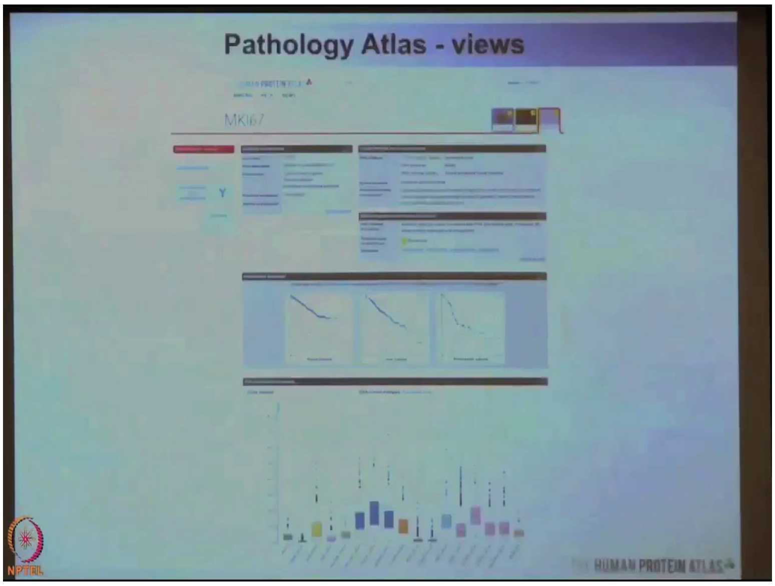
(Refer Slide Time: 09:51)



News
A version 18.1 release today with new Survival Scatter plots

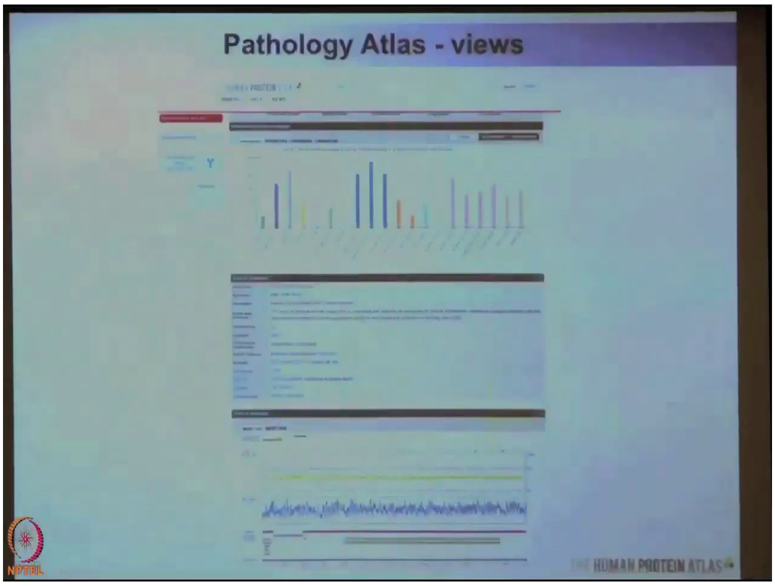
So, I will just show you a few last slides from the newest version of the pathology atlas, it is now called 18.1; we did not take just 2 months ago and where we introduced something we called survival scatter plots.

(Refer Slide Time: 10:14)


Pathology Atlas - views

And, I will show you the views I am from the pathology atlas. So, the pathology atlas contains about 900,000 Kaplan Meier curves survival patient survival curves for all cancer types of these 17 cancer types and all genes where you have a highly significant the difference in survival if you have high or low expression with a median as cutoff. So, these are shown for all audience. I am going to show the TCGA data, the raw data from the RNA sequencing.

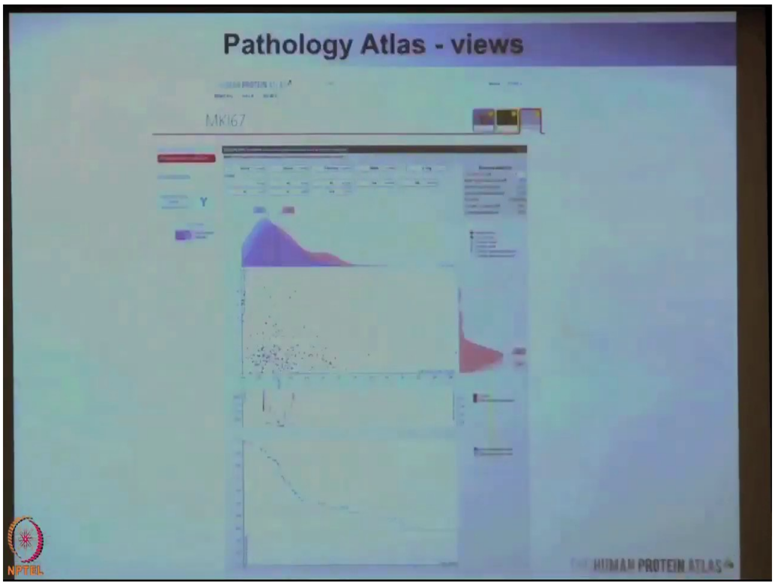
(Refer Slide Time: 11:50)



Pathology Atlas - views

And, when you then go further into the pathology atlas you can see then the protein data, the annotation data, done here in India and then some background on the different genes.

(Refer Slide Time: 11:00)

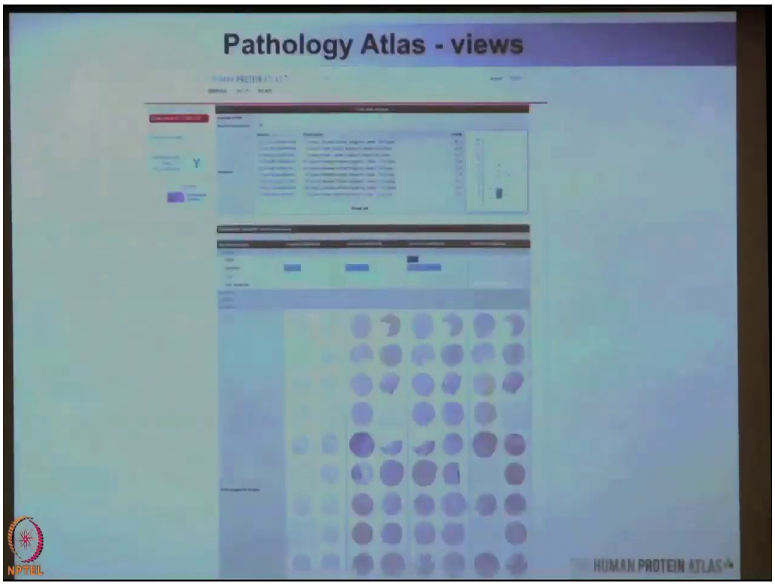


Pathology Atlas - views

What we then introduced was these survival scatter plots, now I will come back to that in just a slide or two, but this is exactly the same data as this. And, this of course, for those who look who work with oncology everybody wants to make KaplanMeyer curve that looks like this that has a high level of separation in the p score that is below 0.0000.

And of course, this looks very seductive while this looks like it is just like a scatter plot where you barely you can see that this is a highly significant gene. And, this is by the way KI67, the most used biomarker for proliferation in cancer and in pancreatic cancer. So, I will come back that in 2 second.

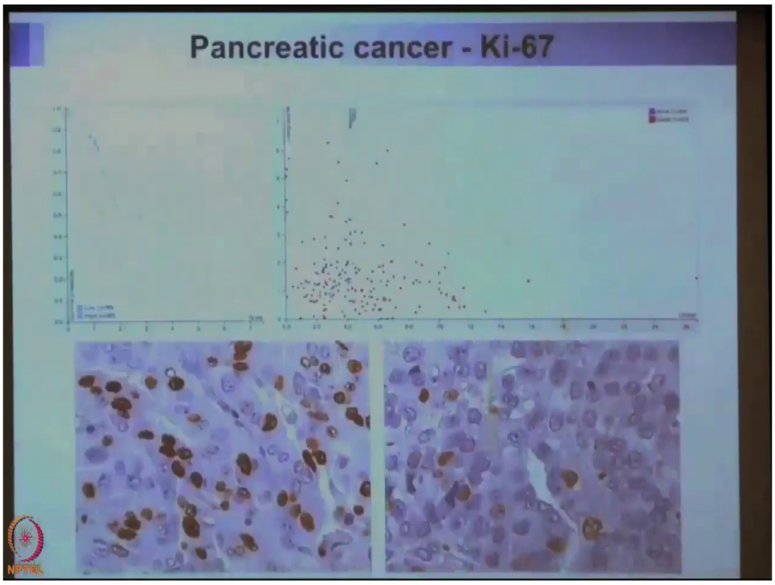
(Refer Slide Time: 11:43)



Pathology Atlas - views

And, we also showing there in the pathology atlas of course, all the immunohistochemical data.

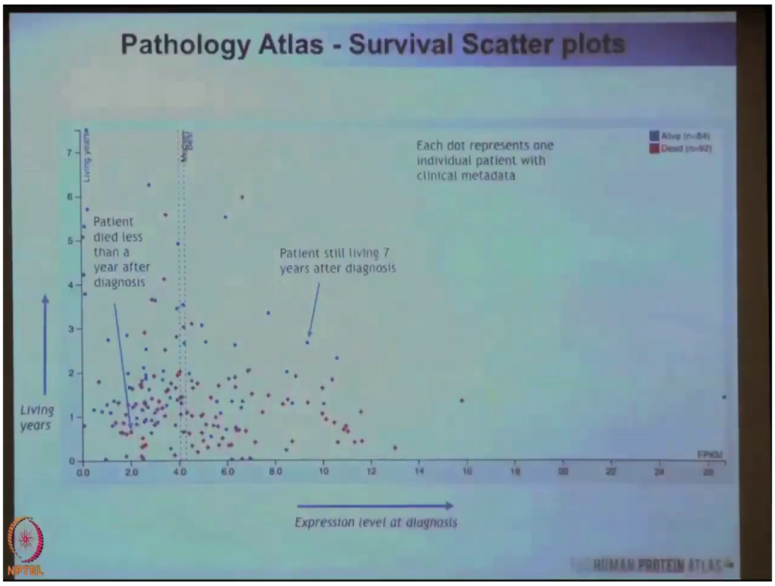
(Refer Slide Time: 11:49)



Pancreatic cancer - Ki-67

And, with where you can go to the full resolution image and this is pancreatic cancer and KI67, this is then the Kaplan-Meier curve where the cutoff is not the medium, but is set at the best highest significance at the lowest p score. This is exactly the same data as this and I will show you what this scatter plot is.

(Refer Slide Time: 12:10)



Pathology Atlas - Survival Scatter plots

Each dot represents one individual patient with clinical metadata

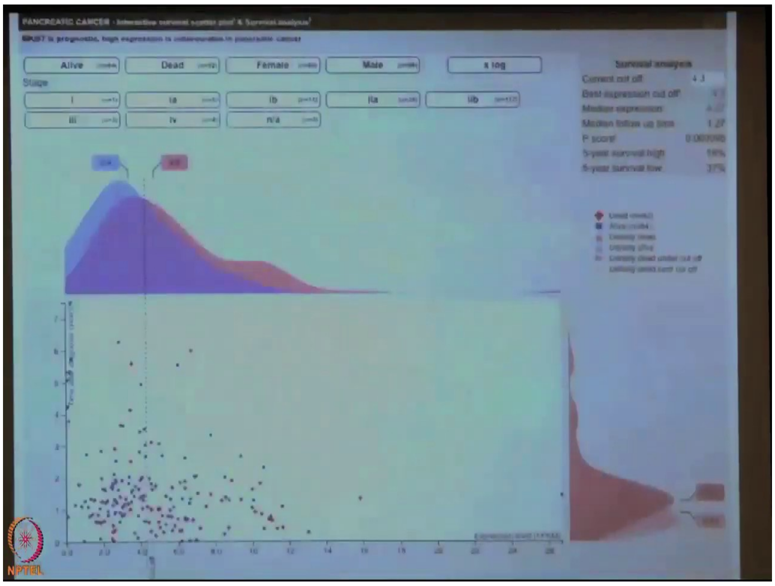
Patient died less than a year after diagnosis

Patient still living 7 years after diagnosis

So, the scatter plot has then living years on the x-axis and then here you can see the level of RNA for KI67 in as in FPKM value and it is 2-dimensional in the sense or 3 or 3-dimensional in the sense that we then have color coded each patients. So, that every blue dot here is a patient that is alive at last time for follow up, every red patient is a patient that has died.

So, you can see here that this well this is a patient that is still living; does it say 7 years after diagnosis that is wrong figure of course, 3 years after diagnosis and this is a patient of red one here this died less than 1 year. You understand how the whole scatter plot works and then we have just plotted in where the best separation is.

(Refer Slide Time: 13:09)

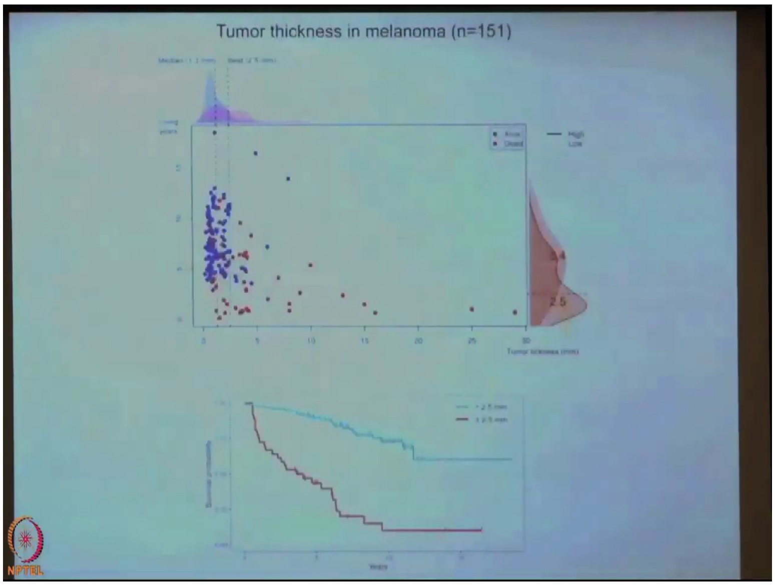


And, we have now then developed this further and this is why we updated their plus 2 months ago by creating these summation curves and making it interactive. So, you can put your own cutoff where you want to and these summation curves then shows you all the blue values added up here and then just smoothened up between the different levels. So, here you can see that this is if you have low levels of KI67 the blue ones you are the surviving patient.

So, low values of FPKM have a higher degree relative degree of surviving patients as compared to the ones with high proliferation where you can see here or up here. So, you can see that this is actually very unfavorable biomarker and I think this curve it is difficult to look at the scatter plot I think this curve is much more easy to get a feeling for both the patient cohort, but also the power or the strength of the biomarker.

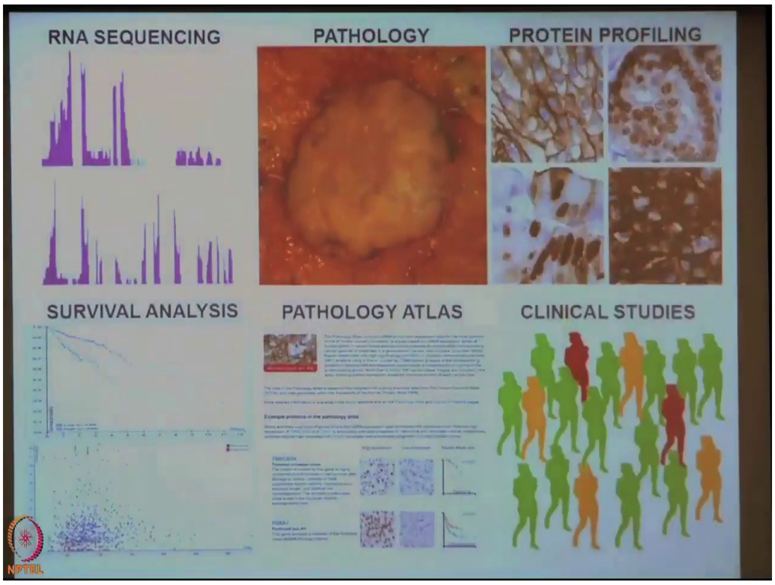
And, here we made a summation of all the dead, all the dead patients and then looked at below and above the cutoff. So, here you can see the patients that have high KI67 above the cutoff they live 0.9 years, while the ones with low proliferation they live 1.3 years. And, that is a kind of nice assessment to get in one curve instead of just showing a couple of Meier curve which is really just fooling you that something looks very good.

(Refer Slide Time: 14:36)



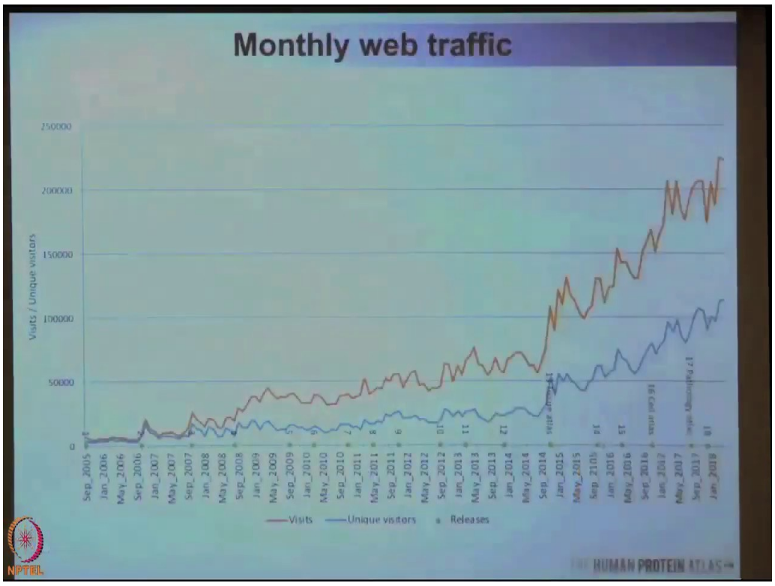
And, I will just show you this like to give you a comparison. The only other continuous biomarker that we have in clinical medicine is tumor thickness for melanoma and here you can see what tumor thickness, what these summation curves can look like in. So, thin melanomas they live 5.4 years, thick melanomas they live 2.5 years. You can see a lot of blue here when you have thin melanomas, lot of red when the melanomas are thicker.

(Refer Slide Time: 15:01)


RNA SEQUENCING
PATHOLOGY
PROTEIN PROFILING
SURVIVAL ANALYSIS
PATHOLOGY ATLAS
CLINICAL STUDIES

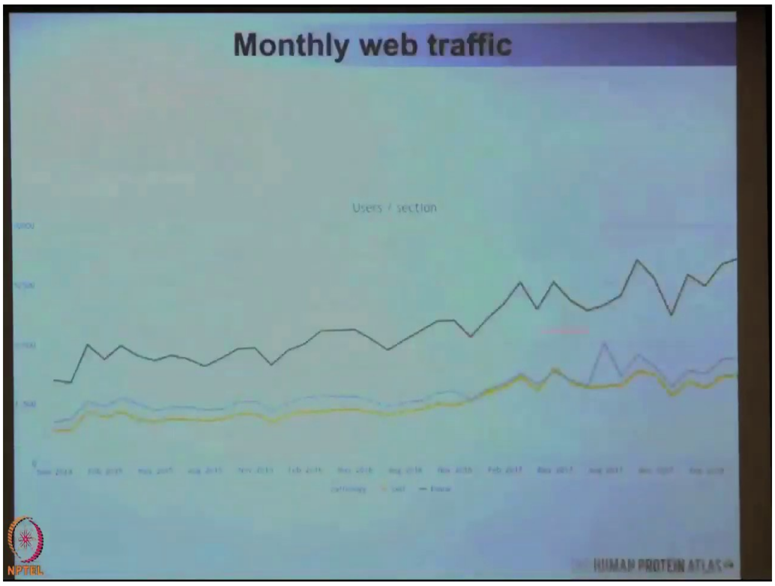
So, what I think I have shown you is that the pathology or the human protein atlas is not least the pathology atlas gives you basic pathology data from RNA sequencing protein profiling data. I think it is a great starting point for clinical studies for biomarker studies. I think there is really a lot of data there that can be utilized for many different forms of research projects.

(Refer Slide Time: 15:25)



I will close by telling you a little bit about how we are doing in the protein atlas. We are doing fine, this is the curve of visits. These are unique visitors and these are visits to the human protein atlas and we have about 300,000 visits per month to the human protein atlas.

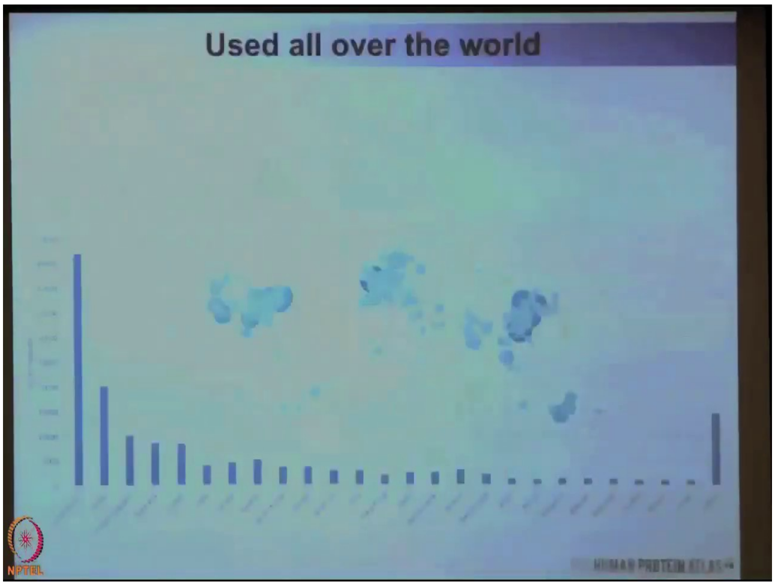
(Refer Slide Time: 15:44)



Monthly web traffic

Most are looking into the normal tissue atlas and looking at the data produced here in India, but also the pathology atlas and cell atlas are coming along quite nicely with a lot of visits every month.

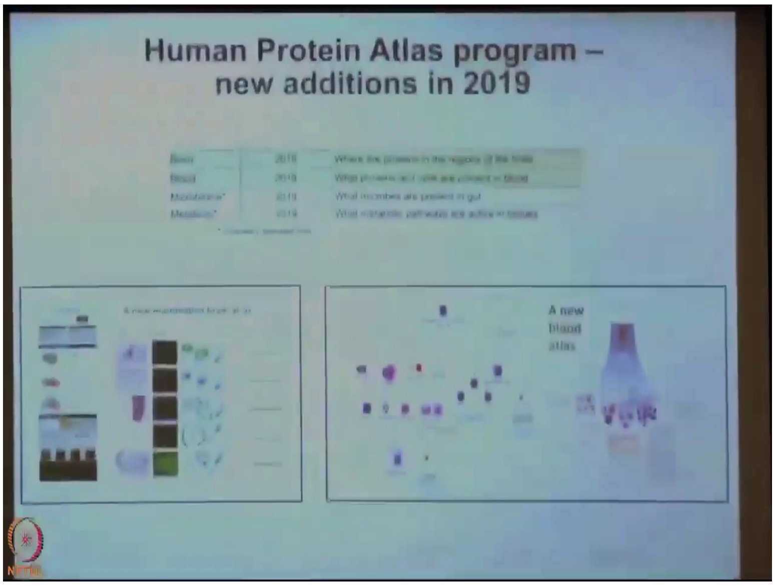
(Refer Slide Time: 15:56)



Used all over the world

The atlas is used all over the world. Dominating of course, is the United States highly dominating followed by China, UK and Germany, but you can see that India is up here we are also Sweden, Canada and some of the other countries are when it comes to using the human protein atlas.

(Refer Slide Time: 16:14)



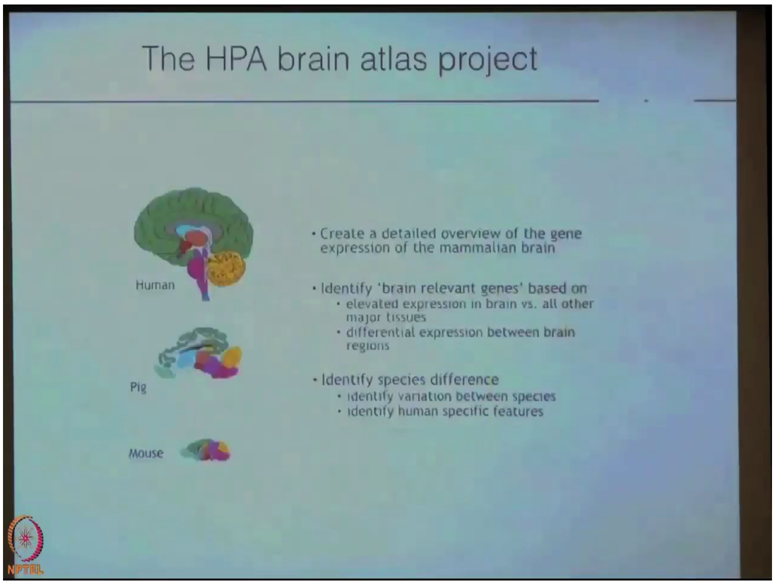
Human Protein Atlas program – new additions in 2019

Brain	2019	Where the proteins in the regions of the brain
Blood	2019	What proteins and cells are present in blood
Microbiome*	2019	What microbes are present in gut
Metabolic*	2019	What metabolic pathways are active in tissues

* [illegible]

So, I will end with what where are we going. So, this is the status today and this is where we are at and we are happy with that, but what are we doing now. So, what we are doing and what would be the most major and biggest update of the protein atlas. So far will be in about 4 or 5 months from now where we will introduce both the brain atlas and a blood atlas, and maybe later on in the fall there will be microbiome and a metabolic atlas being introduced as well.

(Refer Slide Time: 16:44)



The HPA brain atlas project

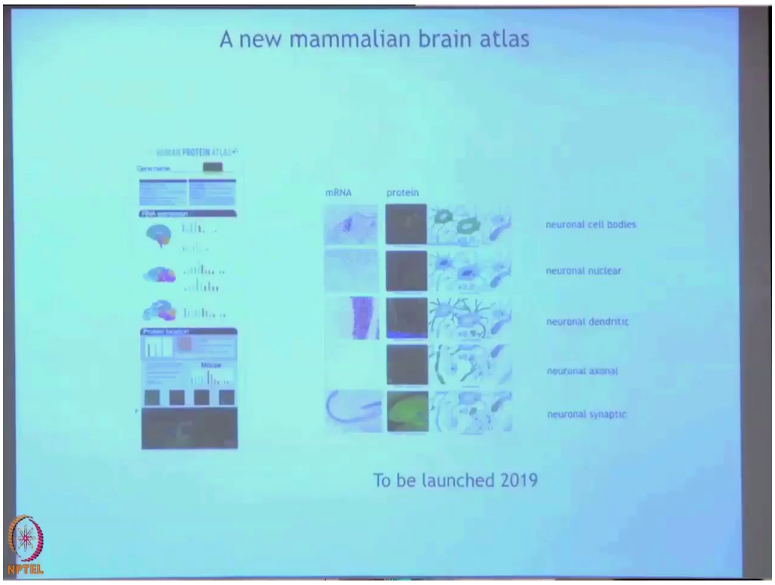
Human
Pig
Mouse

- Create a detailed overview of the gene expression of the mammalian brain
- Identify 'brain relevant genes' based on
  - elevated expression in brain vs. all other major tissues
  - differential expression between brain regions
- Identify species difference
  - identify variation between species
  - identify human specific features

And, I will just show one or two slides about these two atlases the brain and the blood atlas. So, the brain atlas will then be using three different species and this is new them for being the protein atlas we will both use pig data in a collaborate, in a collaboration with the BGI – Beijing Genome Institute and also mouse data that we have on our own.

And, we will then try to identify which genes are specifically expressed in the brain, which are region specific expressed in only the hippocampus or in amygdala or some other brain specific region. But, also try to look at what about the species differences when it comes to overall gene expression in the brain.

(Refer Slide Time: 17:25)

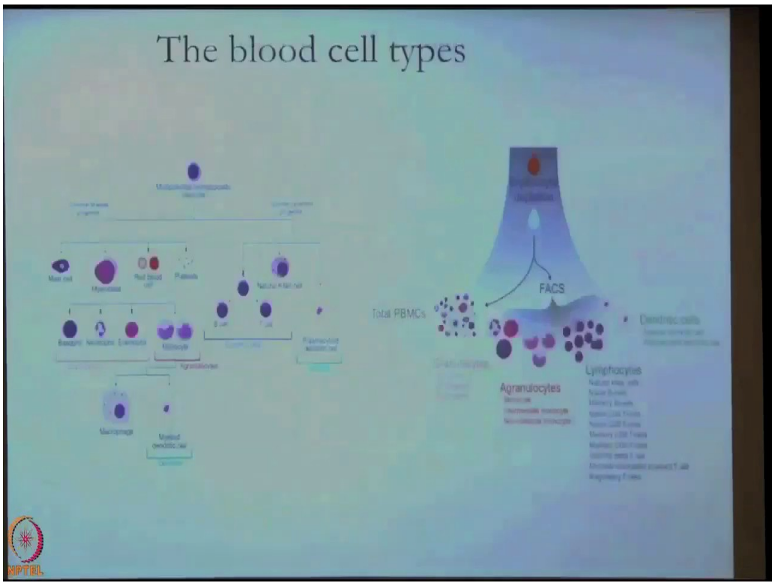


A new mammalian brain atlas

To be launched 2019

This is just the set up with what we are trying to cover with our data and this is a little bit what the data will look like. So, we will be able to both show protein and quantitative RNA measures from three different species and I think this will be a good contribution for neurobiology research.

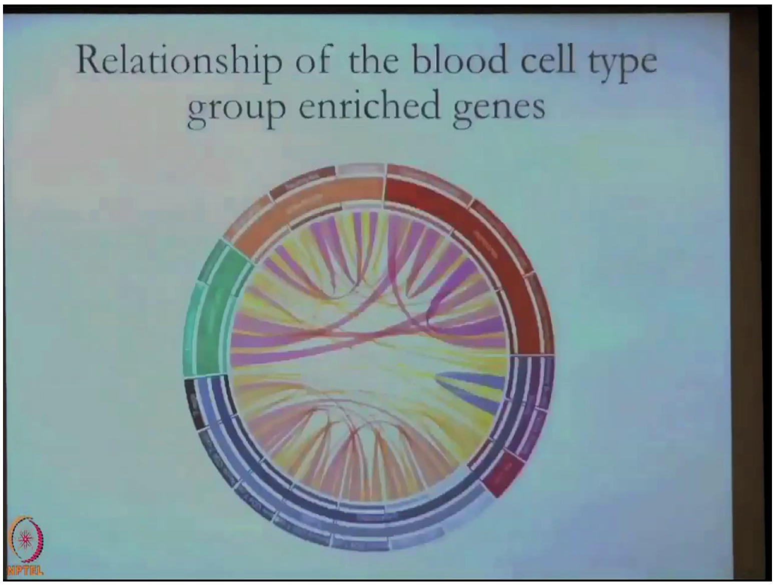
(Refer Slide Time: 17:51)



The blood cell types

The other very interesting part and this is also going to be extremely exciting to see what how the world responds to is that we have then used facts to sort out 18 different subsets of blood cells and doing the full transcriptomic profiling of these, so that one can now start to understand how they relate to each other on a more global scale.

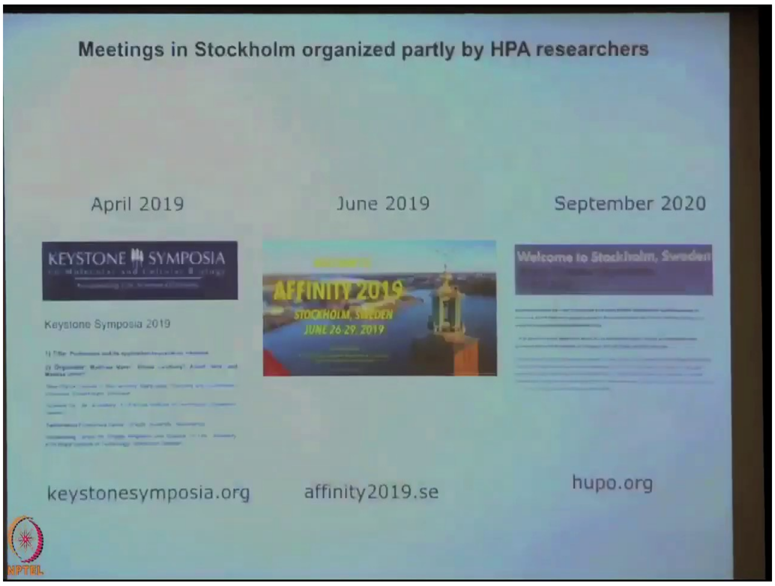
(Refer Slide Time: 18:12)



Relationship of the blood cell type group enriched genes

Showing what genes and how do they relate to each other, different types of blood cells

(Refer Slide Time: 18:20)



Meetings in Stockholm organized partly by HPA researchers

April 2019

June 2019

September 2020

Keystone Symposia 2019

keystonesymposia.org

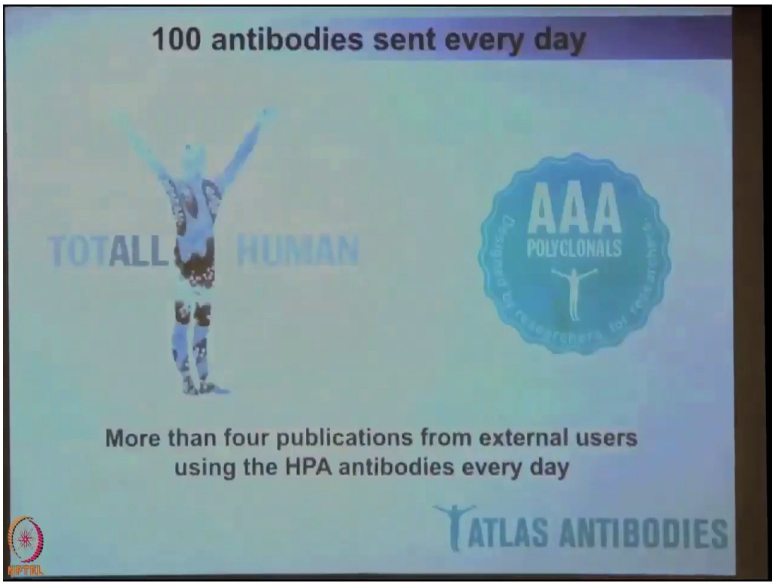
affinity2019.se

hupo.org

KEYSTONE SYMPOSIA
AFFINITY 2019
STOCKHOLM, SWEDEN
JUNE 26-29, 2019
Welcome to Stockholm, Sweden

So, I just want to end by giving you trying to get you to come to Sweden and visit our very cold, but nice country where HPA is part of sponsoring and setting up a couple of meetings at Keystone Symposia in April which is an proteomics. We will have an Affinity meeting in the summer and you really I mean this is a great meeting of course, but June is a great month, if you are going to visit Sweden. And, then I hope you all will come to the HUPO meeting in 2020 which then will be hosted in Stockholm.

(Refer Slide Time: 18:57)



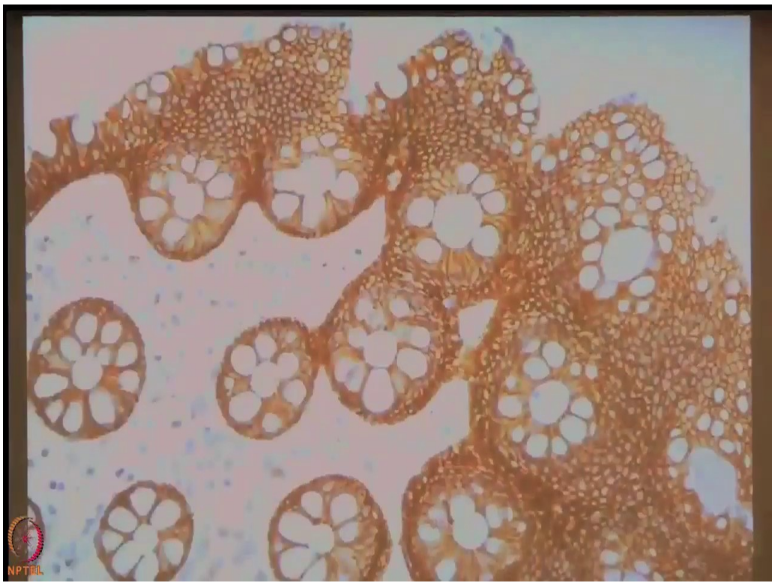
100 antibodies sent every day

More than four publications from external users using the HPA antibodies every day

ATLAS ANTIBODIES

I also want to say that what has been a very nice part of the protein atlas is that we made a spin out company, the atlas antibodies and that we are now distributing antibodies to the whole world and this company is going very well. And more proud of being part of this a company is of course, that our data the science that were being so much sighted and that the data that we are producing is being used all over the world.

(Refer Slide Time: 19:24)



And, for those of you who do not look at immunohistochemistry every day you it is always nice seeing different creatures when looking at different proteins, sometimes you meet somebody who looks happy, sometimes you just get a heartwarming image pack and of course, sometimes it kind of scares you to death, but there is a lot of data.

(Refer Slide Time: 19:51)

Mathias Uhlen

ATLAS ANTIBODIES

Cecilia Lindskog
Caroline Kampf
Emma Lundberg
Peter Nilsson
Jenny Ottosson Takanen
Hanna Tegel

Kalle von Felitzen
Linn Fagerberg
Björn Hallström
Åsa Sivertsson
Per Oksvold
Per-Henrik Edqvist
Jan Mulder

Anna Asplund
Dijana Djurenovic
Julia Bergman
Patrick Micke
Karolina Edlund
Gabriela Gremel
Gillian O'Hurley
Angelika Danielsson

Jochen Schwenk
Evelina Sjöstedt
Max Karlsson
Agata Zieba
Jacob Odeberg
[illegible]
Victor Pontan
.... and more

Sanjay Navani

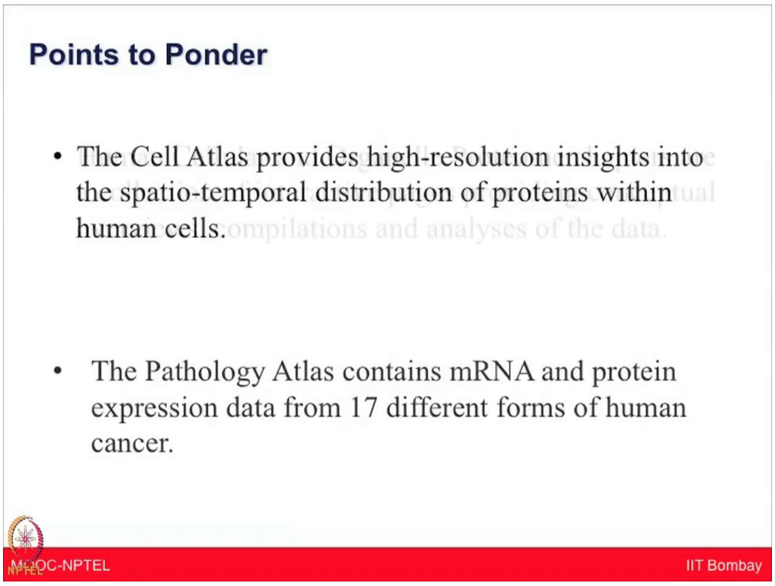
Adil Mardinoglu
Cheng Zhang
Sunjae Lee

HUMAN PROTEIN ATLAS

So, last acknowledgments Mathias Uhlen you know him, he is a great driver takes us from new heights to even higher heights.

So, thank you very much.

(Refer Slide Time: 19:55)



Points to Ponder

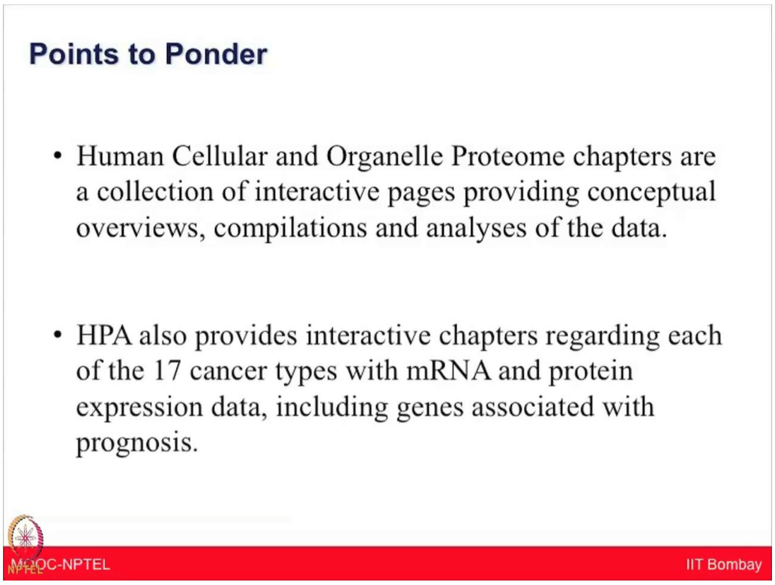
- The Cell Atlas provides high-resolution insights into the spatio-temporal distribution of proteins within human cells.

- The Pathology Atlas contains mRNA and protein expression data from 17 different forms of human cancer.

MOOC-NPTEL

IIT Bombay

(Refer Slide Time: 20:05)



Points to Ponder

- Human Cellular and Organelle Proteome chapters are a collection of interactive pages providing conceptual overviews, compilations and analyses of the data.
- HPA also provides interactive chapters regarding each of the 17 cancer types with mRNA and protein expression data, including genes associated with prognosis.

MOOC-NPTEL

IIT Bombay

In summary, I hope you got good understanding of HPA related projects you have seen that transcriptomics data has been obtained from the cancer genome atlas and proteomic data has been generated in house using the same antibodies as in protein expression profiling in normal human tissues. Dr. Ponten talked to us about the human cellular and organelle proteome chapters, which is a collection of interactive pages providing conceptual overviews, compilations and analysis of the data.

Further, he talked about how survival scatter plot of 17 different cancer types can give you different information of different RNA and protein expression with each patient survival. In the next lecture Dr. Jochen Schwenk will talk about Affinity Based Proteomics.

Thank you.