Introduction to Proteogenomics

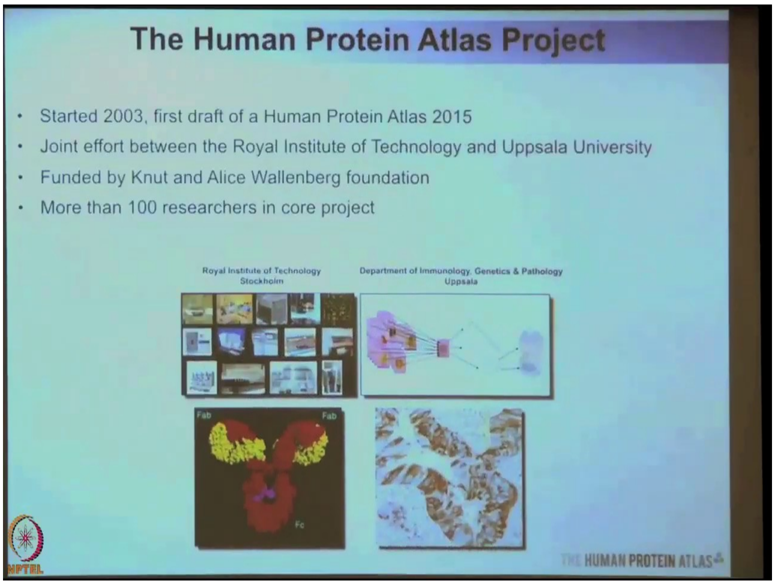
Prof. Sanjeeva Srivastava
Dr. Fredrik Ponten
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 36
Human Protein Atlas – I

Welcome to MOOC course on Introduction to Proteogenomics. Today's invited speaker is Professor Fredrik Ponten, who is currently a Professor at Uppsala University in Sweden. Dr. Ponten will talk to us about the Human Protein Atlas or HPA, which is a Swedish based program. It started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies like; antibody based imaging, mass spectrometry based proteomics, transcriptomics and systems biology. He will tell us about this mega project, how it succeeded despite having multiple challenges.

He will also tell us about how Indian pathologists and research collaborators have played a great role to make everything possible for success of this project. In today's lecture he will mainly focus on the tissue Atlas of Human Protein Atlas. Further, he will tell us about how RNA and protein expression throughout different tissue follows a trend and how this correlation need to be considered for research, if we want to obtain the bigger picture. Dr. Ponten will also talk to us about the sub-proteome, organ based proteome, secretome present in HPA, which will provide you an idea, how to use this useful resource for your own research. So, let us welcome Professor Fredrick Ponten.

(Refer Slide Time: 02:05)



The Human Protein Atlas Project

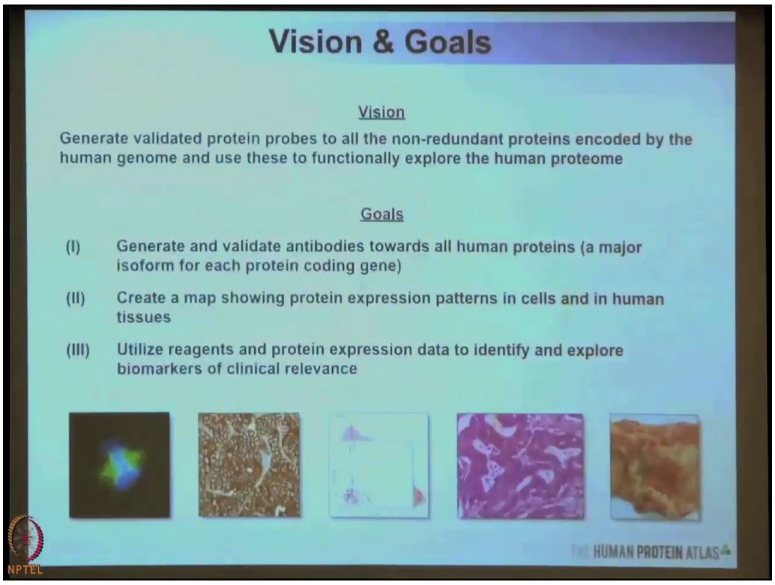
- Started 2003, first draft of a Human Protein Atlas 2015
- Joint effort between the Royal Institute of Technology and Uppsala University
- Funded by Knut and Alice Wallenberg foundation
- More than 100 researchers in core project

Royal Institute of Technology
Stockholm
Department of Immunology, Genetics & Pathology
Uppsala
Fab
Fab
Fc
THE HUMAN PROTEIN ATLAS

What I will talk about today is, is the human protein Atlas and I will give you first just a brief, background about the project. I will give you a little bit of our results and data and where we are right now and in the end, I will give you some perspectives of where we are heading the next couple of years. So, this project started 15 years ago, we received a very generous funding from private nonprofit or research foundation, the Wallenberg foundations and that has kept us alive for these 15 years. And, we had the goal then to have a first draft of a human protein Atlas in 2015 and we fulfilled that goal.

Now, come back to that. The project is a joint effort from the Royal Institute of Technology in Uppsala University and is a head of the whole project is and directors professor Mathias Uhlén, who is a very old friend of mine and I am heading the Uppsala efforts of the project.

(Refer Slide Time: 03:01)



Vision & Goals

Vision

Generate validated protein probes to all the non-redundant proteins encoded by the human genome and use these to functionally explore the human proteome

Goals

(I) Generate and validate antibodies towards all human proteins (a major isoform for each protein coding gene)

(II) Create a map showing protein expression patterns in cells and in human tissues

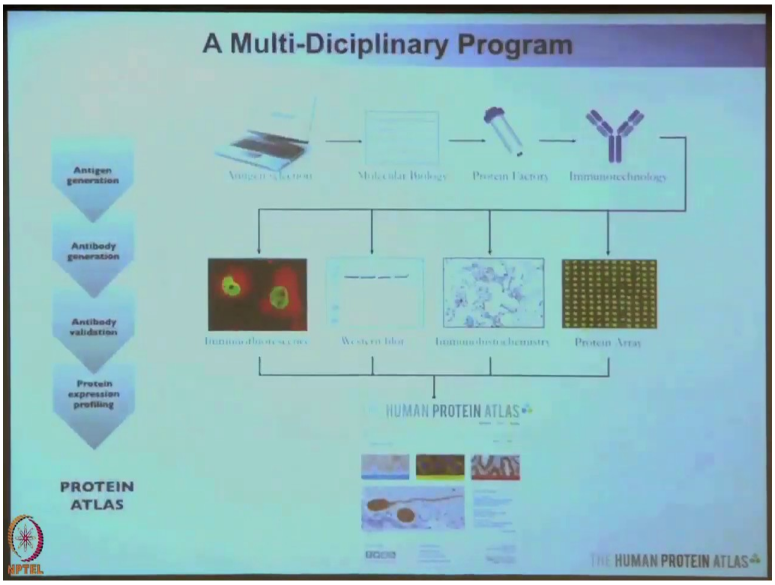
(III) Utilize reagents and protein expression data to identify and explore biomarkers of clinical relevance

THE HUMAN PROTEIN ATLAS

So, our vision then and this is timely this was, if you think back this was started to be planned on during 2002 and if you remember 2001 the human genetic code was published in science and nature by HUPO, not HUPO, HUGO initiative and by Craig Venter and of course, having all the blueprint, having all the ACTs and Gs very logical next step would be to try to add an information layer of what to then all the proteins do that the our genes encode for.

So, that was our kind of vision and the goals, then came down to let us try to make affinity probes, antibodies, let us use these antibodies to characterize the human proteome. And then at last emerging after a couple of years was well, if we have all the data and if we have the reagents, let us try to put this into some clinical perspective and try to make some use into discovery medicine and also try to make some biomarkers and diagnostics, future treatments etc.

(Refer Slide Time: 04:05)



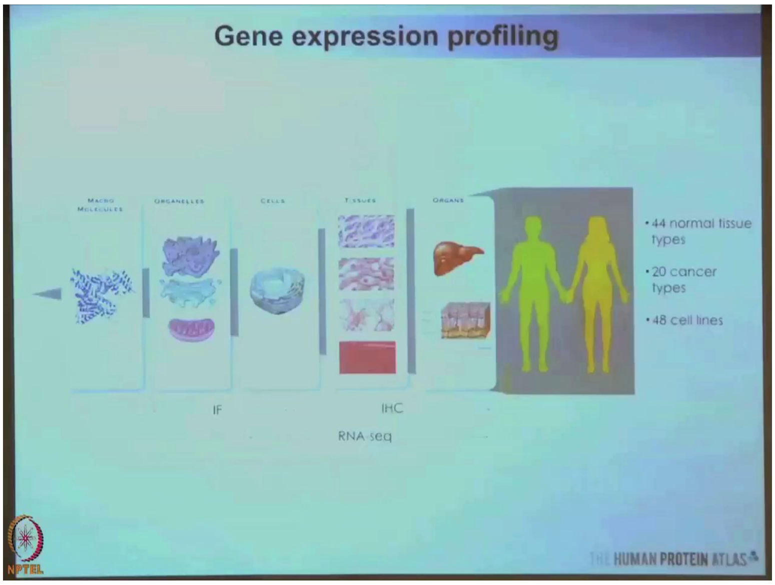
A Multi-Diciplinary Program

So, we set up a multidisciplinary team, a kind of ford factory like research project where we had we defined the different modules. Each module had its own monthly goals and had deliveries to the next goal and so on. And what we did? We started with an upstream bioinformatics part, where we then had the code for all the protein coding genes. We selected a code that was that we blasted against the, I will not go into any details by the way, I think you all know this and you heard about this.

Anyway, this is where we started to make our recombinant proteins and the idea behind it, all is that we blasted the different amino acids against all the rest of the proteome to get as unique sequences, as possible to get as unique protein fragments, as possible to get as unique antibodies, as possible in the end. Outsourcing the antibody production and then we have the immunity technology and we ran everything on protein arrays and the all the antibodies that bound specifically to the right protein fragment.

They were then tested further in immunity chemistry, immunofluorescence and western blots and what was very nice about this whole project was that all the data that we produced was put out in the open space for the scientific community to use. And, that was a requirement from the Wallenberg foundation for the beginning and that has felt very good that there were no restrictions, all data we produce out in the open space.

(Refer Slide Time: 05:27)

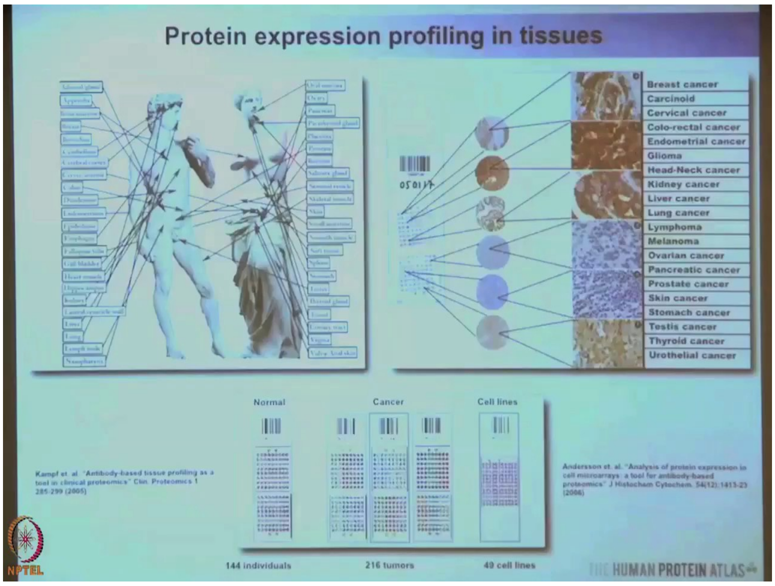


Gene expression profiling

- 44 normal tissue types
- 20 cancer types
- 48 cell lines

So, what we do and what I will focus on is then gene expression profiling and for gene expression profiling, we use an immuno-fluorescence for looking at cells and organelles immune is the chemistry for looking at cells tissues organs that level and then we do RNA sequencing to get quantitative data for our gene expression profiles. Now, briefly just give you the background for this. I am sure Dr. Navani has told you all about this before, but what we use them for protein profiling or then affinity purified antibodies against all the different unique proteins that genome encodes for. And what we do then?

(Refer Slide Time: 06:05)



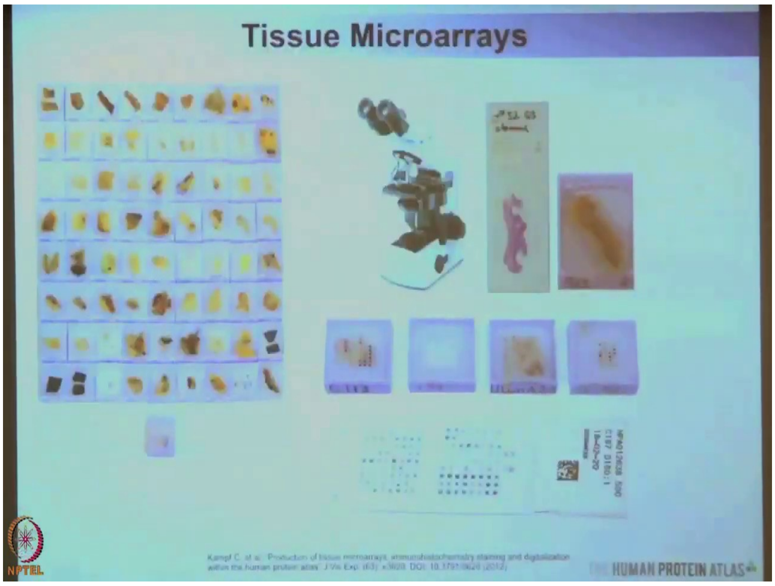
Protein expression profiling in tissues

Kampf et. al. "Antibody-based tissue profiling as a tool in clinical proteomics" Clin. Proteomics 1 285-299 (2005)

Andersson et. al. "Analysis of protein expression in cell microarrays: a tool for antibody-based proteomics" J Histochem Cytochem. 54(12):1413-23 (2006)

We look at how proteins are distributed in all our different organs and tissues and the way we can do this to get a comprehensive look at that without wasting too much tissue and too much reagents, is that we use tissue microarrays and we have them focused on normal tissues, cancer tissues. And, also cell lines and for normal tissues, we have 46 different normal tissue types in triplicates from three different individuals. We they make tissue microarrays by selecting representative pieces of tissues.

(Refer Slide Time: 06:35)



Tissue Microarrays

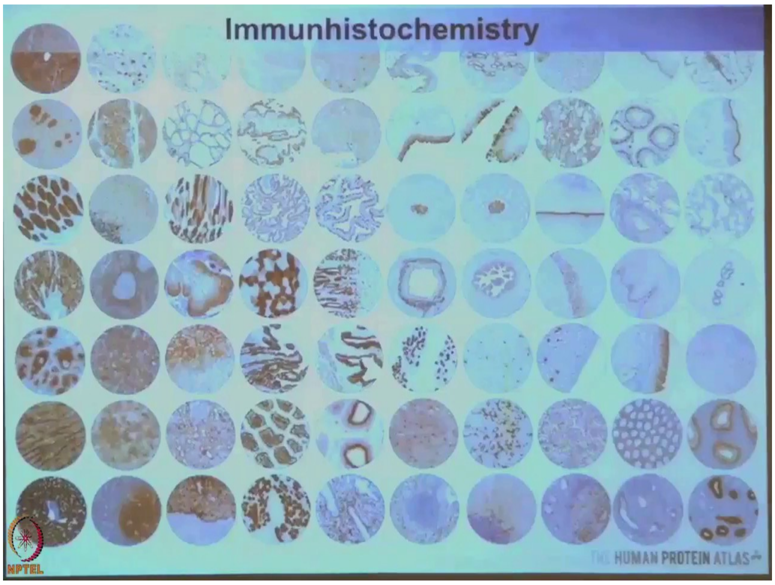
Kampf C. et al. "Production of tissue microarrays, immunohistochemistry staining and digitalization within the human protein atlas" J Vis Exp. (63) e3620. DOI: 10.3791/3620 (2012)

HUMAN PROTEIN ATLAS

We look under a microscope, you find representative areas, drill out a core and then put it in a recipient block to produce tissue microarrays and one of these can we can make about 300-350 sequential sections thus, used for about 300-350 different antibodies and be able then to protein profile. A large part of the human body by using tissue microarrays and this was also very timely, because it was at the end of the nineties, when Kononen coined the term tissue microarrays.

And, the first instruments were made for this and this was also something that made this whole project possible was that we had the possibility to use tissue microarrays. And, I think this slide tells you everything about tissue microarrays, but handling 700 of these blocks for each antibody that just would have been impossible while, handling 4 blocks here is absolutely possible.

(Refer Slide Time: 07:47)



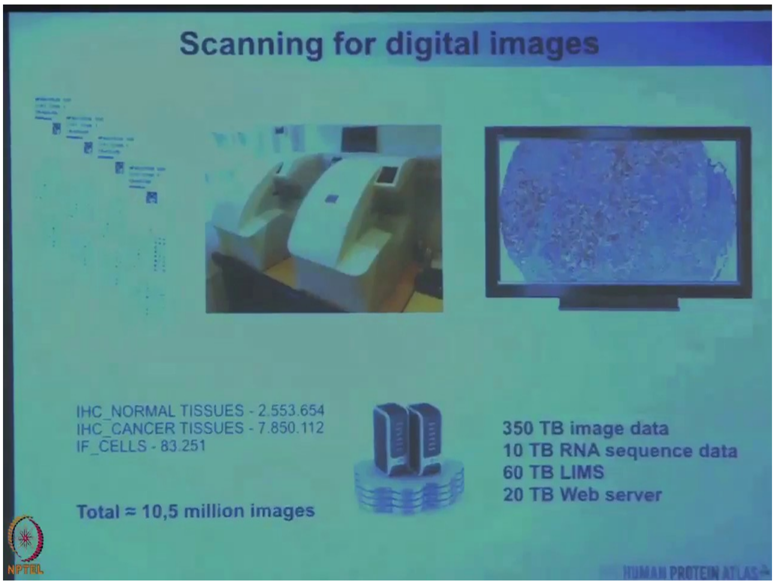
Immunhistochemistry

THE HUMAN PROTEIN ATLAS

Immunohistochemistry is our basic method for, when it comes to tissues when it comes to getting a protein expression profiles and as you know immunohistochemistry is a great method, when it comes to spatial data, but it is a poor method, it is not a method to get any quantitative data. But, there is nothing like immuno-histochemistry that can actually give you what structures, what subtypes of cells do express a certain protein and it gives you a little bit feeling of quantity in the sense that, if you have a complex tissue.

You have one population here that is strongly positive, another one that is weakly positive. At least you know that this population expresses a higher level of the protein then than the other one, but it does not give you any quantification at all besides from that.

(Refer Slide Time: 08:35)



Scanning for digital images

IHC_NORMAL TISSUES - 2.553.654
IHC_CANCER TISSUES - 7.850.112
IF_CELLS - 83.251

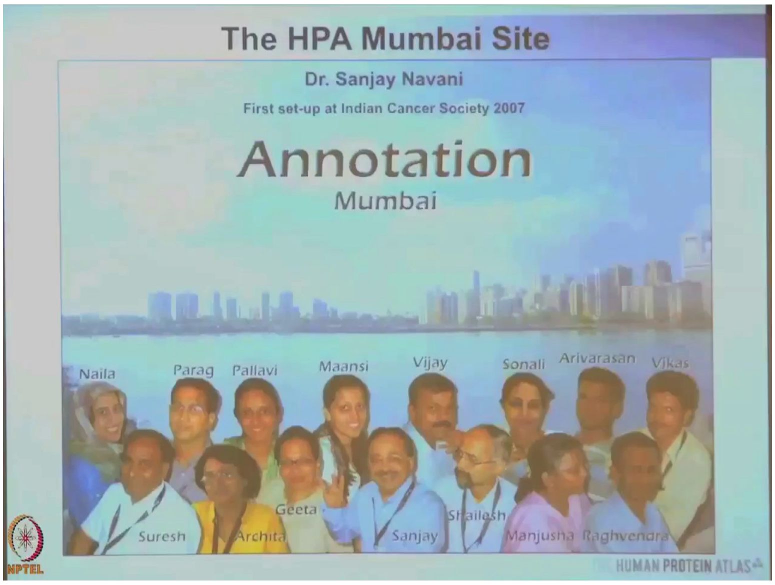
Total ≈ 10,5 million images

350 TB image data
10 TB RNA sequence data
60 TB LIMS
20 TB Web server

And of course, to do this project, we had also to transform the glass slides into digital images and that was also at the time. Than when we started in 2003, a challenge absolutely to handle all the enormous amounts of image data and to store the data and to be able to pick up the data and so on and of course, the magic of the whole project at this time was not just putting out images in a big library, but also making some data from those images and that is where our collaboration with India and with Dr. Navani started.

We realized that you know the scientific community would not have been helped by just having images stained with immunohistochemistry. And the people who can interpret immunohistochemistry and evaluate tissues is a cancer cell, is it a normal cell, is it strongly expressed here or weakly those are the pathologists.

(Refer Slide Time: 09:33)



The HPA Mumbai Site

Dr. Sanjay Navani

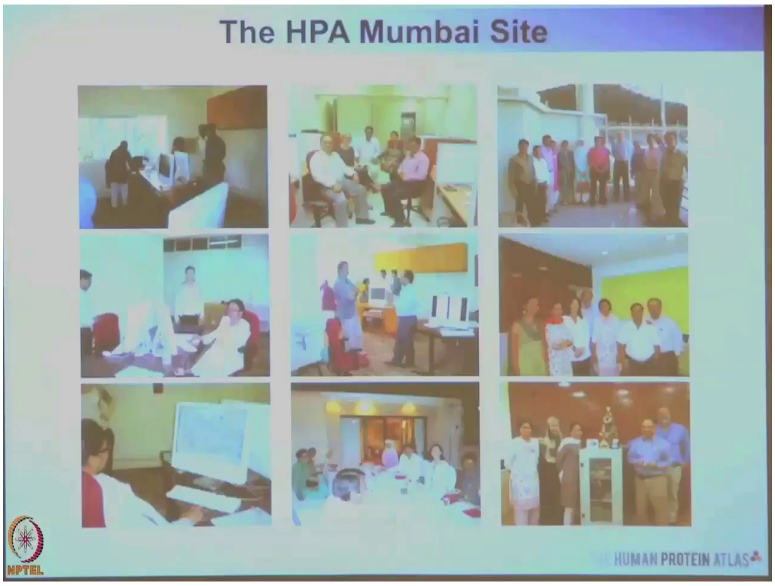
First set-up at Indian Cancer Society 2007

Annotation

Mumbai

And, meeting up with Dr. Navani and his team of pathologists back in 2006 and we started and set the first site was set up at the Indian cancer society, in 2007 by the all these talented pathologists who started looking at images and we have to solve all the internet IT structure challenges and so on, but everything worked out very well.

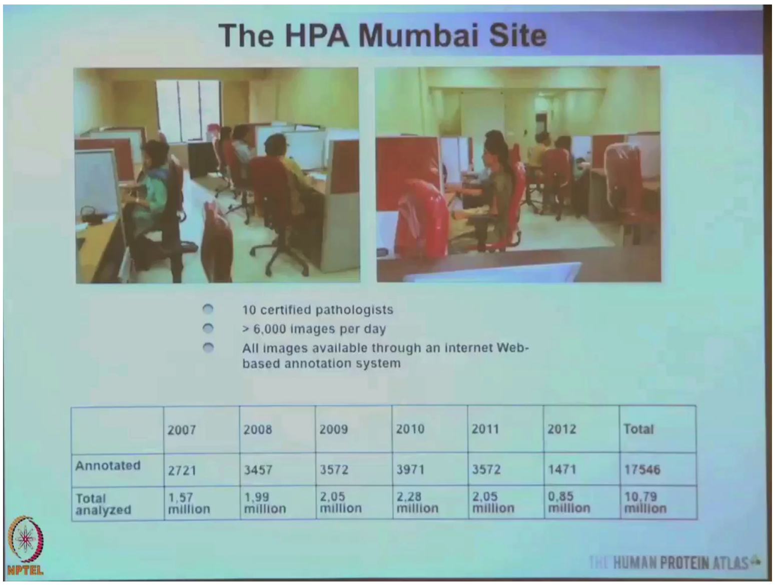
(Refer Slide Time: 10:01)



The HPA Mumbai Site

So, we continued to collaborate and we were down here, many from my team were here for months and worked together with our Indian colleagues and we changed the site to another venue. And we have had just great collaborations with India, Indian pathologists in this project and they have produced all the data which I will show you on the next slide and I have summarized that as being 10 certified pathologists sitting looking at these images.

(Refer Slide Time: 10:23)



The HPA Mumbai Site

- 10 certified pathologists
- > 6,000 images per day
- All images available through an internet Web-based annotation system

	2007	2008	2009	2010	2011	2012	Total
Annotated	2721	3457	3572	3971	3572	1471	17546
Total analyzed	1,57 million	1,99 million	2,05 million	2,28 million	2,05 million	0,85 million	10,79 million

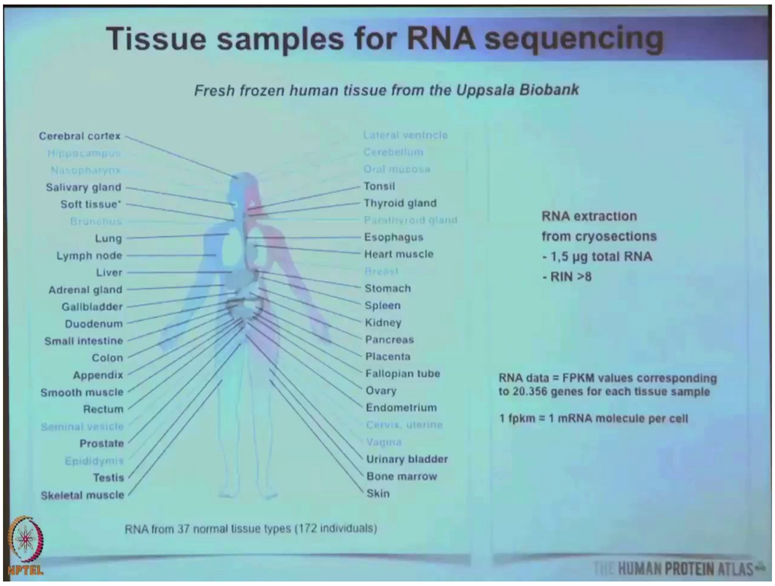
NPTEL
THE HUMAN PROTEIN ATLAS

Evaluating them, putting out annotations is it weakly expressed as a strongly expressed. It is within 25 percent of the cell population or more and you can see here this is not the full figure, but it goes to beginning of 2012. You can see that, they then go through 2 million images per year, which I think is extremely impressive and all together over 12 million images have been annotated by Indian pathologists, but not only the workflow and the volume is impressive. It is also been an impressive time too for the research collaborations and I just did this morning checked out our me and Dr. Navani's, we are co-authors on those papers and they are highly cited papers in science and many good journals.

So, it is not only been production of data, but it is also been a very fruitful scientific collaboration, which I am very grateful for. So, with that is the protein part of the tissue Atlas and also of the of the pathology Atlas and I will come back to the pathology atlas in a while what we realized, a couple of years ago was that spatial data is great, but it, but unique quantification and I know that all of you know this is you work with proteomics with which is a quantitative method to a large extent.

So, what we did was we went back to the Uppsala Biobank and looked for frozen tissue samples and we went through these by microscope to see that we had normal tissue, we selected cases that were representative and where we had high quality RNA.

(Refer Slide Time: 11:59)



Tissue samples for RNA sequencing

Fresh frozen human tissue from the Uppsala Biobank

RNA extraction
from cryosections
- 1,5 µg total RNA
- RIN >8

RNA data = FPKM values corresponding to 20.356 genes for each tissue sample

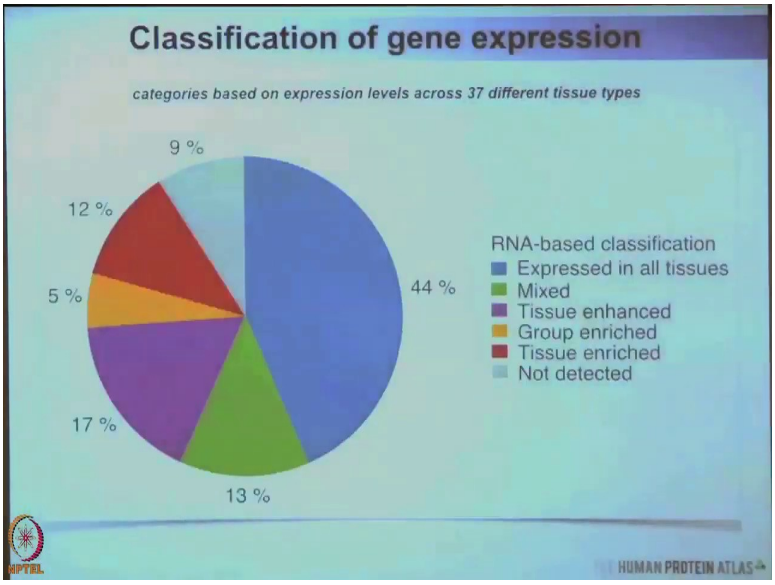
1 fpkm = 1 mRNA molecule per cell

RNA from 37 normal tissue types (172 individuals)

THE HUMAN PROTEIN ATLAS

And, we extracted RNA and then we did RNA sequencing to get then transcriptomics data from normal tissues and we had at least three different individuals for each tissue types. And in the end we had or now, we have 37 normal tissue types in over 200 individuals where we have all the transcriptomics data that has then empowered the human protein Atlas database.

(Refer Slide Time: 12:23)



Classification of gene expression

categories based on expression levels across 37 different tissue types

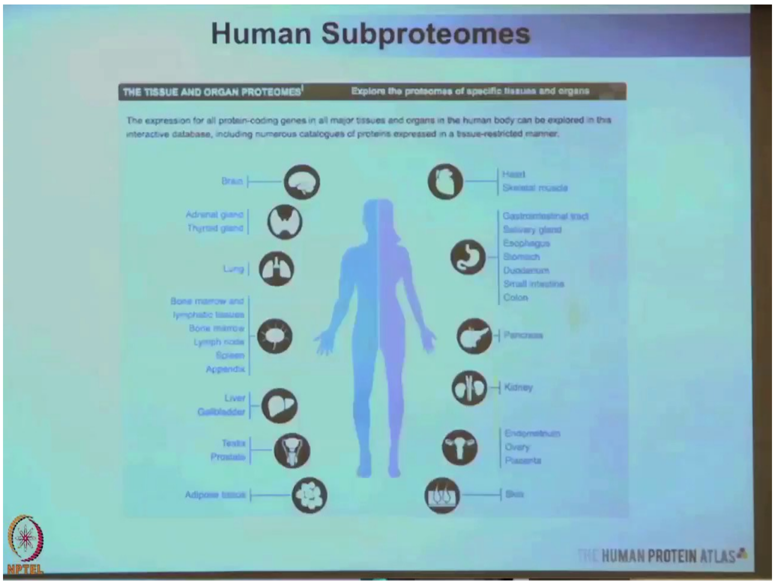
So, this now, we started to learn a little bit more about the proteome and about the human proteome and how are our genes actually expressed on the protein level, because and I want to come back to this more specifically, but it has been shown and this has been a debate and it depends a little bit on definitions, but what about the correlation between RNA and protein. And, I say that for almost all genes there is an extremely high correlation between RNA and protein and when I say that I mean across tissues or cell lines if you have a high level of RNA in one cell line or one tissue type and low level of RNA in another cell line of tissue type the protein levels will follow the RNA levels.

However, for each gene there is a different RTP RNA to protein ratio and that can differ by many magnitudes, but if you go across tissues, the correlation is very high between RNA and protein and; that means, that you can use RNA quantitative, RNA sequencing data as a proxy for protein levels. So, what we learned here was that about half of our protein coding genes, encode for proteins, which are housekeeping proteins 44 percent are expressed in all tissues.

They through the proteins that you know build structure and cell division or all cellular integrity and everything, then there is a mixed bag and then we have these proteins, which are the most interesting proteins, the tissue type specific proteins. The proteins are only expressed in one tissue or in very few tissues or much higher expressed in a certain tissue type than compared to other types.

These are the ones of course, that that are responsible for the special functions of different tissues and these are the ones which will be interesting when it comes to diseases and disease biomarkers. And, about 9 percent at the time we could not find any RNA in our 37 different tissues and these could of course, be pseudogenes, they could be genes that are permanently turned off after development or there could be genes that are in tissues that we did not have like inner ear or olfactory plate or other more remote types of tissues.

(Refer Slide Time: 14:39)



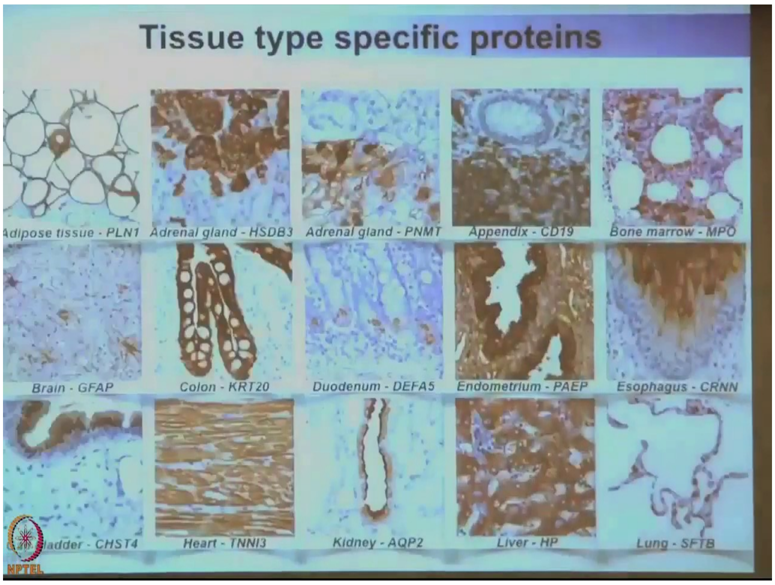
Human Subproteomes

THE TISSUE AND ORGAN PROTEOMES   Explore the proteomes of specific tissues and organs

The expression for all protein-coding genes in all major tissues and organs in the human body can be explored in this interactive database, including numerous catalogues of proteins expressed in a tissue-restricted manner.

With this data at hand, we started them to define the different human sub-proteomes the different organ proteomes and we put this out on the protein Atlas. And, this is a part of the protein Atlas where we built the knowledge based chapters and I will show you just one example after this.

(Refer Slide Time: 14:53)

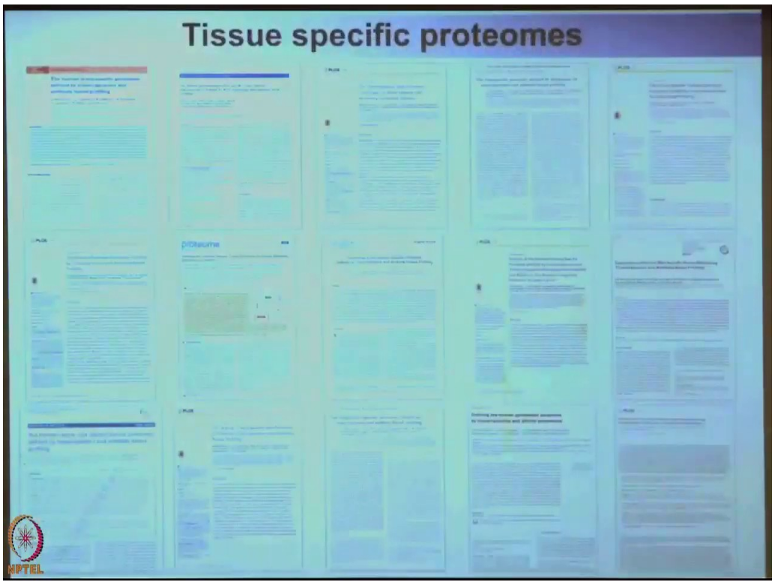


Tissue type specific proteins

What was nice now was that we had the quantitative data from RNA sequencing and we could combine it them with our spatial data from or antibodies. So, we could look aware of the adipose tissue specific proteins how are they expressed; what about the adrenal gland are they expressed in the adrenal medulla or are they in the cortex are they special subtypes of cells etcetera.

And of course, the spatial information together with this quantitative information does not give you function per say, but it gives you a very good hint of function where you see a protein expressed in a certain cell type and it is been in a certain organ and these are just examples of such cell type or tissue type specific proteins expressed in either here exocrine pancreas or endocrine pancreas etcetera.

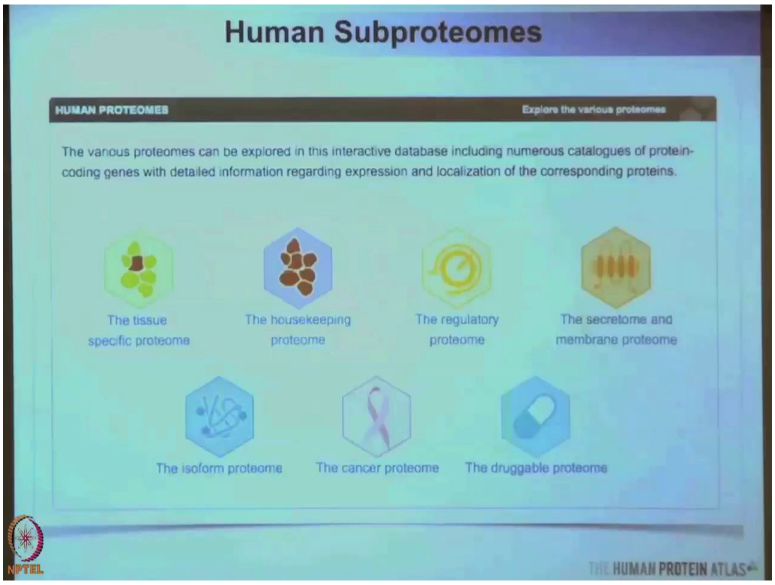
(Refer Slide Time: 15:45)



Tissue specific proteomes

So, we spent a couple of years writing papers. So, if any of you are interested in any specific type of tissue or tissue proteome we have probably published a paper about it, because we thought it was very interesting to go a little bit more into depth; what is what makes up the brain or what makes up the pancreas or whatever. Another way of also transacting through the proteome is to do it not by organ, but expression mode or and I talked about the tissue specific proteome.

(Refer Slide Time: 16:17)



Human Subproteomes

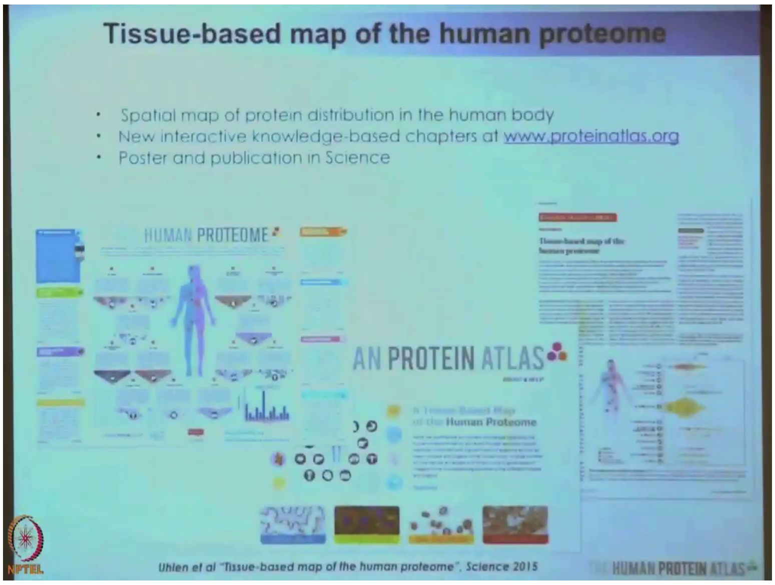
HUMAN PROTEOMES

Explore the various proteomes

The various proteomes can be explored in this interactive database including numerous catalogues of protein-coding genes with detailed information regarding expression and localization of the corresponding proteins.

Of course, there is a housekeeping proteome. What about those proteomes or the regulatory proteomes, what about all the transcription factors where are they expressed are there differences, in different tissue types, cell types, etcetera. Secretome and membrane proteome extremely important for the communication between cells and also as biomarkers of course, isoform proteome the very complex isoform proteome, which kind of empowers the whole, biology with a lot of complexity, cancer proteome obvious and druggable proteome very interesting for the drug industry of course.

(Refer Slide Time: 16:51)



Tissue-based map of the human proteome

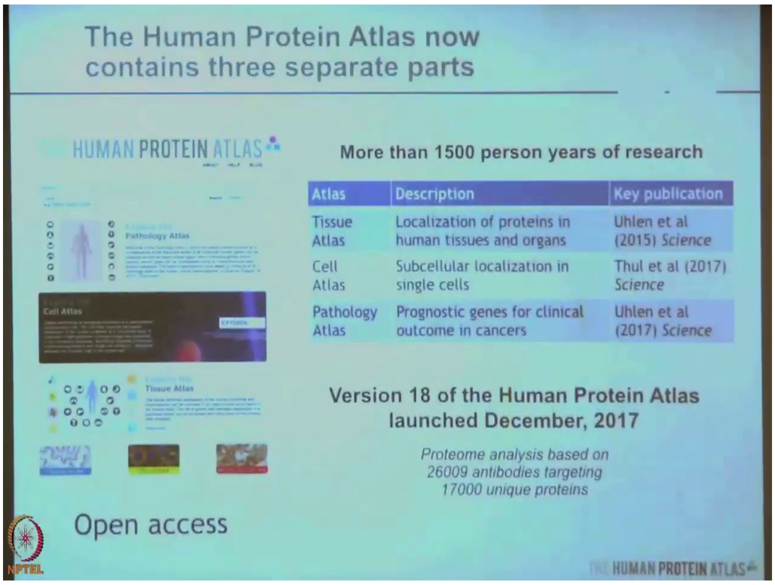
- Spatial map of protein distribution in the human body
- New interactive knowledge-based chapters at www.proteinatlas.org
- Poster and publication in Science

HUMAN PROTEOME
AN PROTEIN ATLAS
Tissue-based map of the human proteome
A Tissue-Based Map of the Human Proteome
HUMAN PROTEIN ATLAS

Uhlen et al "Tissue-based map of the human proteome", Science 2015

And all these pages knowledge based paces, they are then in place at in the protein a plus. So, you can go there and I will show you one example from organ proteome in just a second. So, 2015 we said that now, we have a first draft of the human proteome and we were very successful to publish a paper in science which has been very highly cited we had a poster in science and we rebuilt the whole protein atlas web portal to then integrate the transcriptomics data and the proteomics data.

(Refer Slide Time: 17:25)



The Human Protein Atlas now contains three separate parts

More than 1500 person years of research

Atlas	Description	Key publication
Tissue Atlas	Localization of proteins in human tissues and organs	Uhlen et al (2015) *Science*
Cell Atlas	Subcellular localization in single cells	Thul et al (2017) *Science*
Pathology Atlas	Prognostic genes for clinical outcome in cancers	Uhlen et al (2017) *Science*

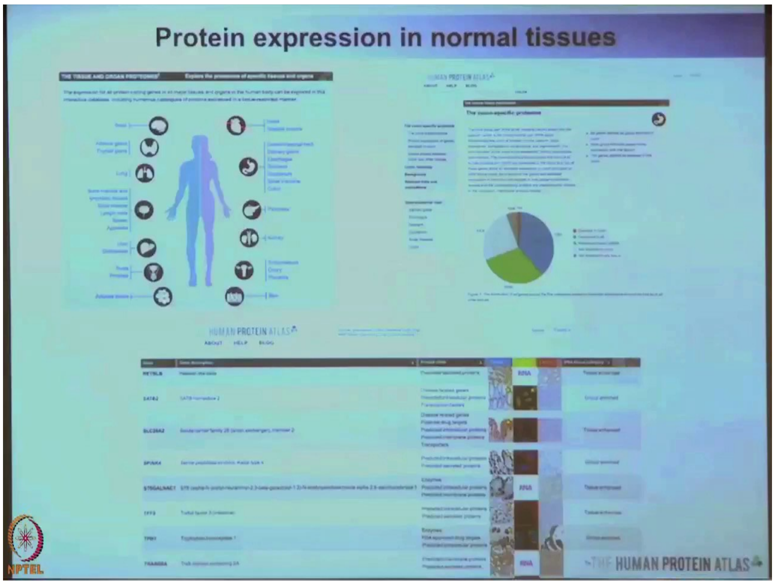
Version 18 of the Human Protein Atlas launched December, 2017

Proteome analysis based on 26009 antibodies targeting 17000 unique proteins

Open access

So, today the human protein atlas has three pillars, it has the tissue atlas normal tissue Atlas, which shows you in which organs and cell types our genes are expressed, it has the cell Atlas, which shows you in what organelles are our proteins expressed in the cell. And then we have the pathology Atlas, which I will come back to which shows you where the how does gene expression correlate to survival for patients that have cancer and I will show you a very short just a couple of slides from each of from the web portal.

(Refer Slide Time: 18:01)

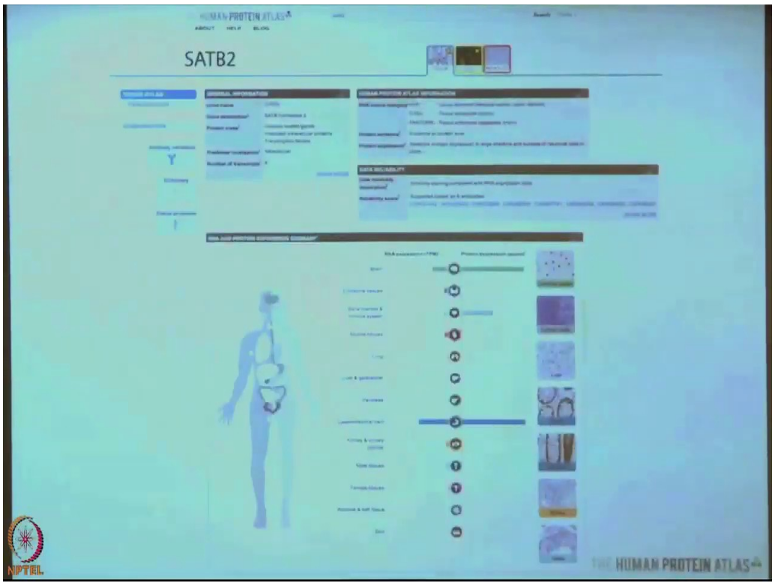


Protein expression in normal tissues

And I will start with the human tissue Atlas and here you can go into and look at these if you want to go through the organ proteomes or the other sub proteomes. And then you can just click on any of these tissue types say here, I click on colon that brings me to a couple of pages that summarizes the gene expression profile in colon.

And if say I am interested in than these colon specific proteins I can then click on that and that brings me into the hit list of the of the protein Atlas and here I get the 165 proteins, which are specifically expressed in the in the colon. I can choose one of these, I can click on that oops.

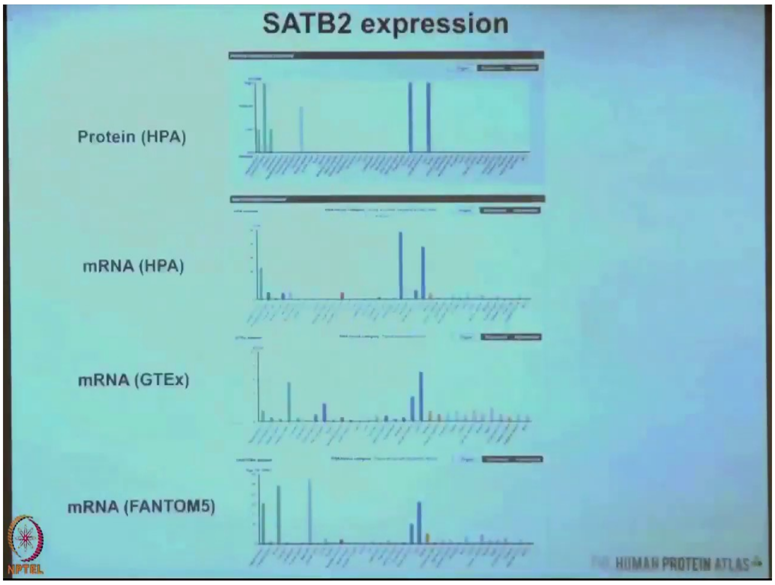
(Refer Slide Time: 18:49)



SATB2

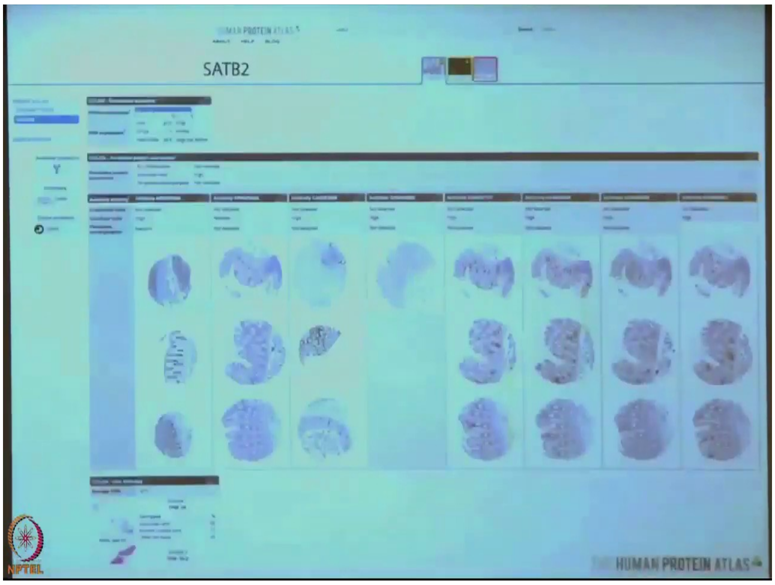
And then I can click on that that yes and then I get to the summary page and in this case this is a gene called SATB2 encoding for protein that is more less specifically, expressed in the colon in the epithelial cells of the colon and rectum. It is also expressed in the brain, we give our a little summary about the every gene all 20000 genes and then the expression levels on the RNA level, which is an FPKM. And then on the protein level, which is an how they are how the Indian pathologists have evaluated the expression level the protein expression levels.

(Refer Slide Time: 19:23)



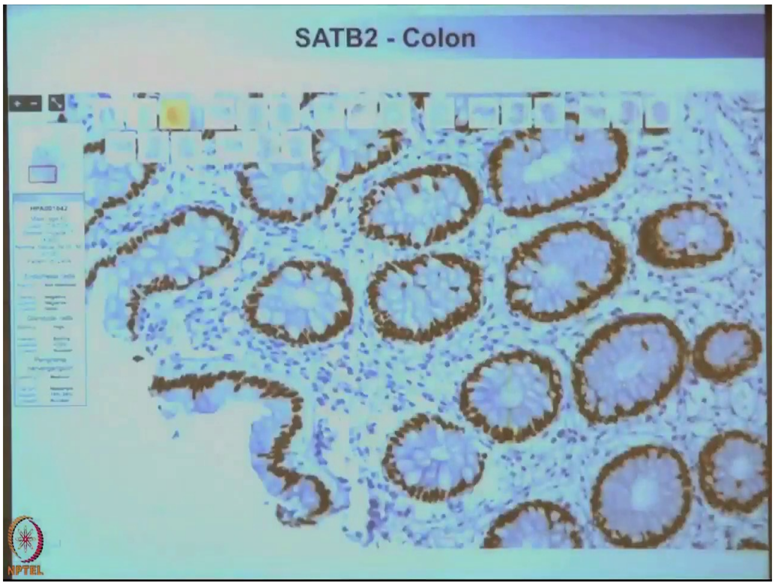
And then one can look at the data in more detail the protein data is about bar diagrams or RNA sequence data, but we also have imported for all genes the data from the broad institute GTEx project and also from the FANTOM file project. So, and as you can see there is a very good consistency from the different platforms and the different specimens that have been used and I think this gives a lot of validity to the expression data that we show on the protein Atlas.

(Refer Slide Time: 19:55)



And then of course, one can go and look at the primary data, the protein data, where we then have three individuals for each antibody and for this SATB2, we had many antibodies.

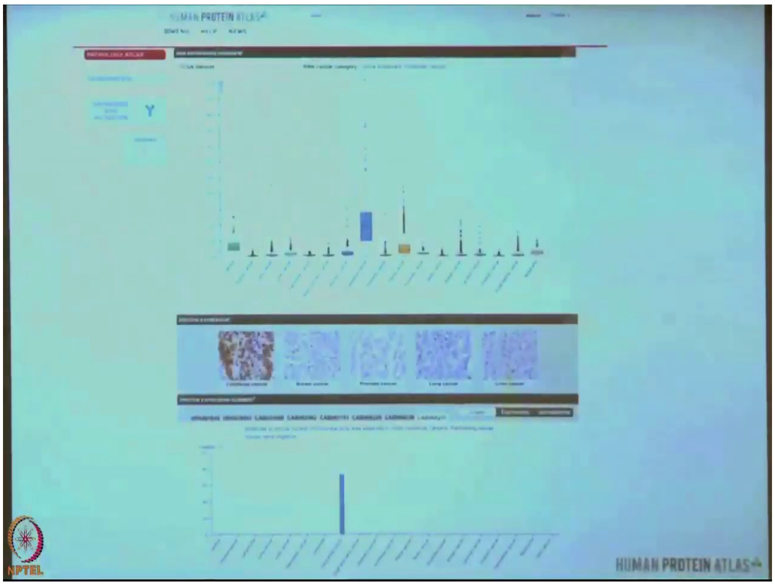
(Refer Slide Time: 20:05)



SATB2 - Colon

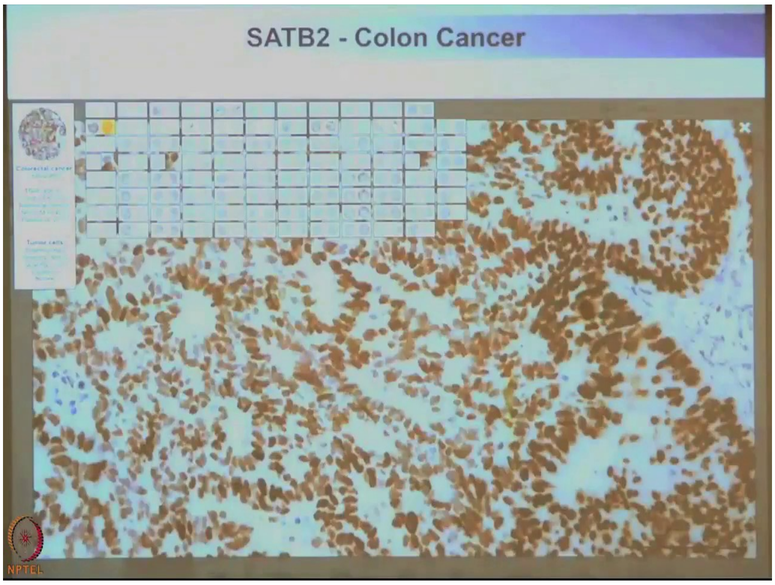
And then at the deepest level you can then go into the high resolution image and look for yourself where is SATB2 protein expressed. Well it is expressed in the nucleus of the glandular cells in colon, etcetera and just as a little parentheses since this was a very highly specific.

(Refer Slide Time: 20:19)



Colon protein we thought maybe this could be a biomarker for colon cancer patients. So, we looked in colon cancer and you can see, it is highly expressed in colon cancer on the protein level the only tissue that expresses the you can see high expression of ZAPPY 2 was colon cancer.

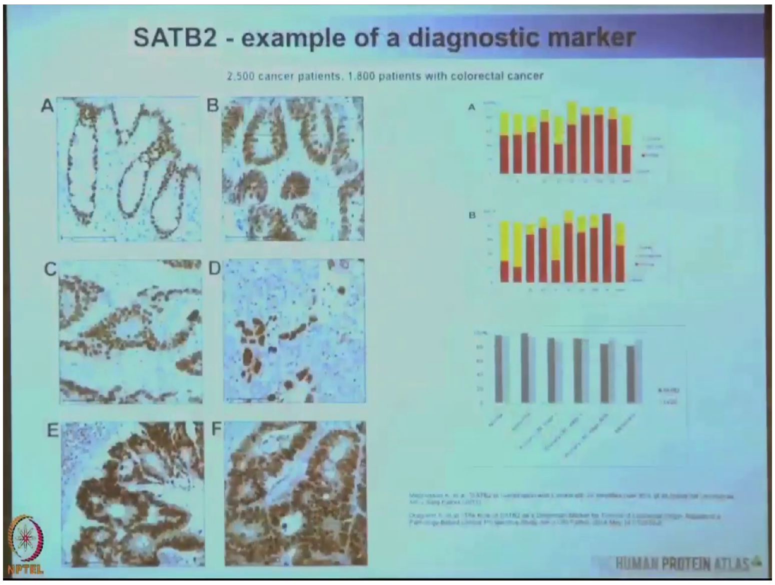
(Refer Slide Time: 20:39)



SATB2 - Colon Cancer

So, here we did and you can look at the high full blown resolution also for cancers of course.

(Refer Slide Time: 20:43)

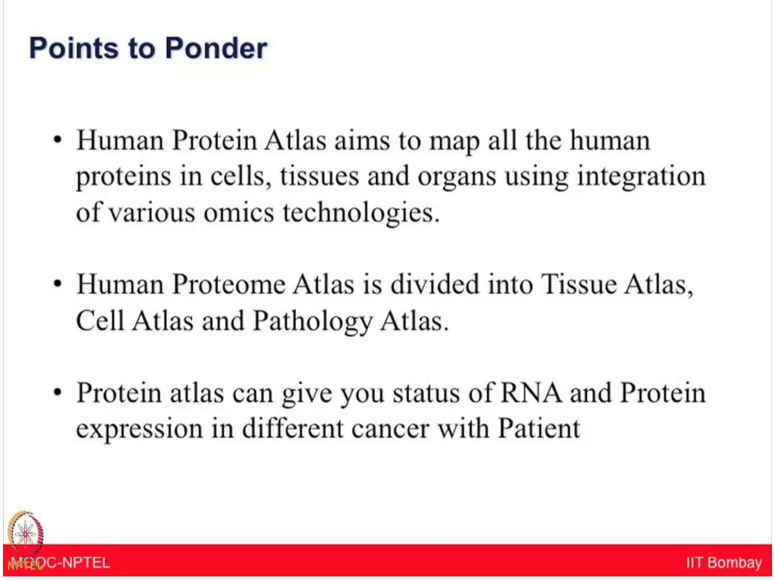


SATB2 - example of a diagnostic marker

2.500 cancer patients, 1.800 patients with colorectal cancer

But here, we then extended the study and did a clinical study including over 2500 patients and actually could establish that this is a good cancer biomarker for colorectal cancer.

(Refer Slide Time: 20:55)



Points to Ponder

- Human Protein Atlas aims to map all the human proteins in cells, tissues and organs using integration of various omics technologies.
- Human Proteome Atlas is divided into Tissue Atlas, Cell Atlas and Pathology Atlas.
- Protein atlas can give you status of RNA and Protein expression in different cancer with Patient

MOOC-NPTEL IIT Bombay

In today's lecture you have learnt about HPA and found that Human Proteome Atlas could be divided into Tissue Atlas, Cell Atlas and Pathology Atlas. Dr. Ponten demonstrated expression level of different genes in 37 different types of tissue and how this information is important to understand diseases and identify candidate biomarkers.

He also talked to us about; how the protein Atlas can provide you the status of RNA and protein expression in different cancer with patient follow up data and I will highly recommend you to visit HPA website and explore it further. It will definitely be helpful resource for your own research in the next lecture; Dr. Ponten will talk about the shell atlas and pathology atlas in more detail.

Thank you.