

**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. Jochen M Schwenk**  
**Department of Biosciences and Bioengineering**  
**KTH Royal Institute of Technology**  
**Indian Institute of Technology, Bombay**

**Lecture – 38**  
**Affinity based proteomics & HPA**

Welcome to MOOC course on introduction to Proteogenomics. Today we have a guest speaker, Dr. Jochen M Schwenk from KTH Royal Institute of Technology. Doctor Jochen will talk to us about affinity proteomics which is a field of proteome analysis based on use of antibodies and other binding reagents has protein a specific detection probes. He will also a talk about the study of human plasma proteome using affinity based methods, which could enhance biomarker discoveries validation and integration from basic research towards the clinical usage.

He will then talk about the resources like Biobank Sweden and Atlas antibody. Dr. Jochen will also talk about mass spectrometry technique and how it can be used to study post translational modifications, PTM peptides in a digested sample. He will then talk about PTM scan technology which allows identification and quantification of hundreds to thousands of, even the lowest abundant peptides and provides a more focused approach to peptide enrichment than the other available strategies. So, let us welcome doctor Jochen for his lecture.

What I would like to talk about today is to give you a bit of a different perspective on what, what we do and what we understand by doing plasma proteomics and maybe there are some aspects of it that could be helpful for you and that sort of provides some ideas of either collaboration or you know for you just to get the new perspective on in your projects and how to move forward and so this is my team.

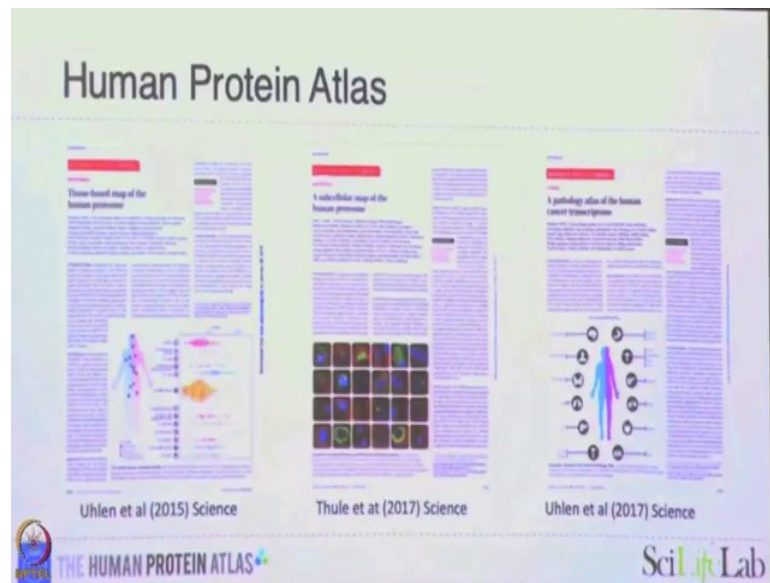
(Refer Slide Time: 02:06)



So, we are currently about yeah 10 people and it is actually Kimmy, the person she half Japanese to the left who made this painting. She is very sort of skilled in arts. And, but I think it is also a nice way to sort of you know illustrating us instead of one group of people with sort of the same phenotype, right.

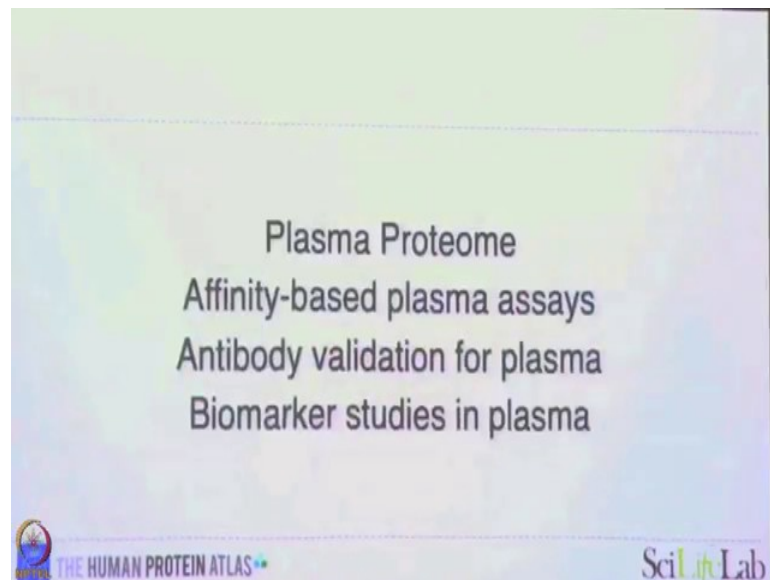
Even though you know we have Philippa, she is from the UK. We have Mun Gwan, he is from the he is from Korea. We have Ragna, she is the postdoc from Germany. We are very international group and now actually we have a new person from Denmark. So, it is really sort of you know the mix of cultures and mix of backgrounds that I think is really sort of important.

(Refer Slide Time: 02:46)



So, I guess Fredrik has talked to you last week about these different aspects of the human proteome atlas. So, the tissue based atlas, the subcellular atlas as well as the pathology atlas. So, I am going to leave you with that and I hope you still remember some, some elements of it.

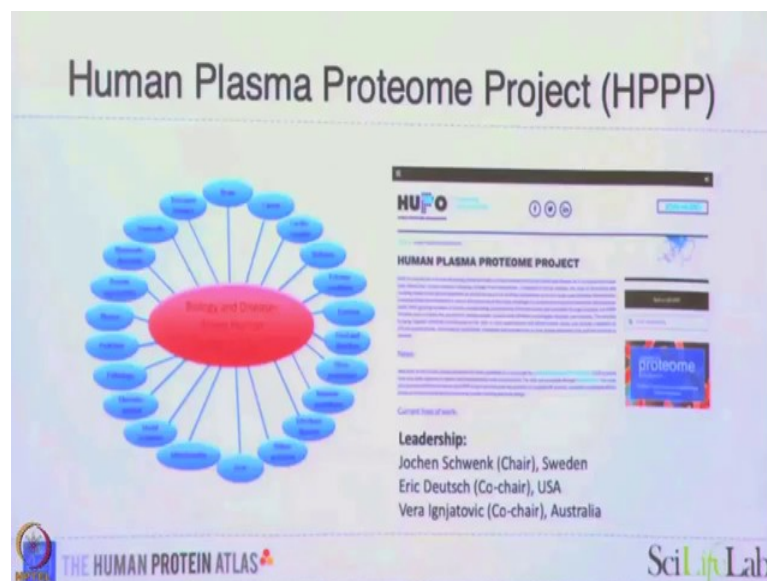
(Refer Slide Time: 03:03)



So, I will talk a bit more about sort of what is actually outside of the cells. So, talking a bit about the plasma proteome as we see it and then how to use affinity based methods for studying the plasma proteome.

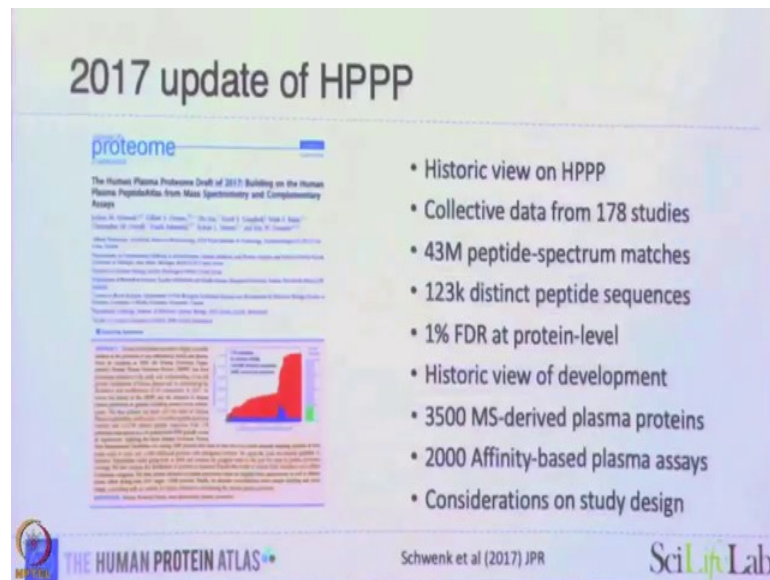
I will give you some examples of how we use mass spectrometry, but the predominantly part will be actually on looking at a plasma proteins with affinity reagents. So, as a we have allotted to I am in the current chair of the human plasma proteome project and whatever; that means, is sort of you know to be defined, but I think what it sort of it is meant to be a sort of an organization that tries to given a global understanding of what are the initiatives that people are working on, in the different areas of the world and with a common feature of studying the plasma proteome.

(Refer Slide Time: 03:48)



And I am doing this together with Eric Deutsch, who is famous by a mathematician from Seattle as well as Vera Ignjatovic from Australia, so really trying to you know have this as a global initiative to.

(Refer Slide Time: 04:02)



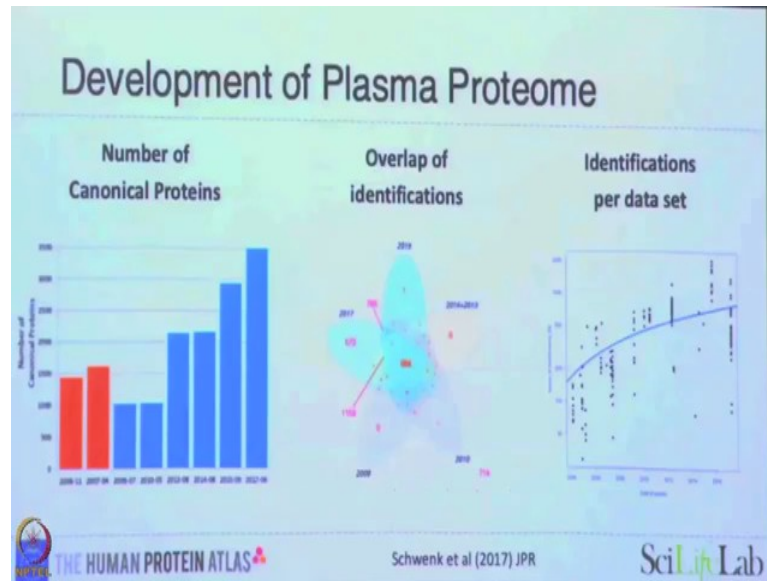
The slide is titled "2017 update of HPPP" and features a screenshot of the Proteome website on the left. The Proteome website header includes the word "proteome" in a blue font. Below the header, there is a title "The Human Plasma Proteome Draft of 2017: Building on the Human Plasma Proteome from Mass Spectrometry and Complementary Assays" and a list of authors. A red bar chart is visible in the lower right corner of the website screenshot. To the right of the website screenshot is a bulleted list of key findings. At the bottom of the slide, there are logos for "THE HUMAN PROTEIN ATLAS" and "SciLifeLab", along with the text "Schwenk et al (2017) JPR".

- Historic view on HPPP
- Collective data from 178 studies
- 43M peptide-spectrum matches
- 123k distinct peptide sequences
- 1% FDR at protein-level
- Historic view of development
- 3500 MS-derived plasma proteins
- 2000 Affinity-based plasma assays
- Considerations on study design

So, about 2 years ago, we published this paper in the annual special issue of JPR and we are actually in the process preparing a new sort of review for the coming issue where we basically concluded that about 5000 proteins that we can detect using proteomics methods in plasma, which is probably you know 25 percent of what the genome actually tells us there is. Of course, this is predominantly driven by the fact that these are the things we can measure. It does not mean that these are the things that actually are useful, ok.

So, given that you know you have new technologies such as Somalogic, who claim that they can measure 5000 proteins. This is within the ballpark of what we see at the moment, mass spectrometry in combination with other affinity based assays can measure. And as of course, one intrinsic challenge to using for instance mass spectrometry and that is, one is of course, you need to have a good detection system and protocols to increase the coverage.

(Refer Slide Time: 04:58)



So, from around a 1000 proteins which we could identify it some 10 years ago, you know I think we have made a big step forward in detecting more. And this is also shown here by the change this Venn diagram showing the progress. There are also interesting numbers here highlighted in red, which are those proteins which actually sort of got introduced over the years over the recent years.

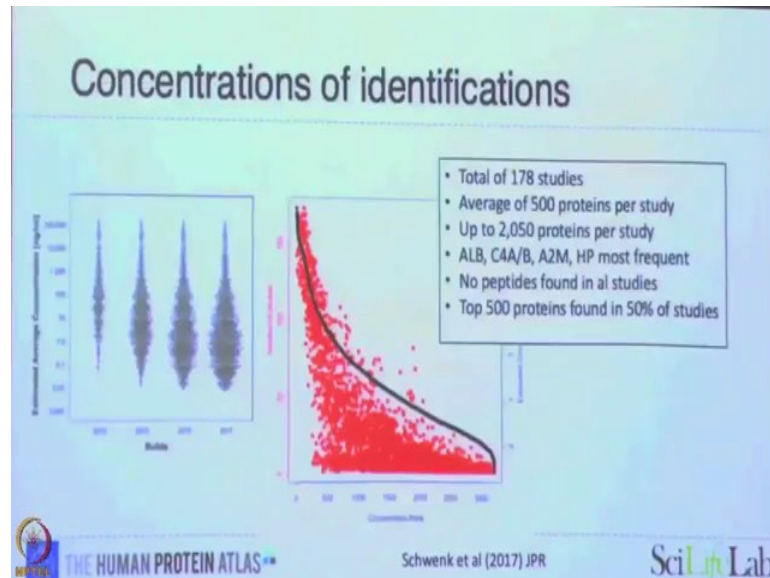
And it is particularly, I think you have these two gaps between 2010 and 2013, but there also about 700 proteins which disappeared from this list and meaning that these are proteins that have probably been missed, annotated and probably glycosylation forms that have been sort of led to the false identification. So, I guess you know we know that the end of sort of having the perfect system together, but I think we have a much better understanding of what the system looks like.

Another challenge in mass spectrometry is of course, the coverage, meaning how many proteins do you in your single experiment actually can measure and this is shown here again they have a time chart on the x axis and this is the number of proteins identified on the y axis. You see there is of course, a progress being made over the years that you can measure.

Nowadays, let us say route about 500 proteins in every experiment, but you can also see there as a quite substantial span in some studies people claim to have identified 2000, by most reason you know some studies have only measured about a 100. So, it is really a

matter of defining what do you call a protein to be and whether it should be identified in only one or in every sample of your measurement.

(Refer Slide Time: 06:31)



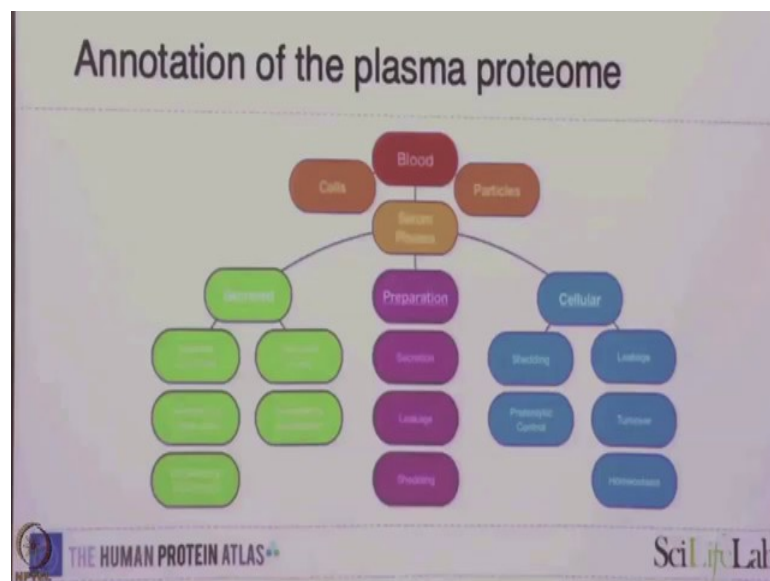
And that brings me to the next sort of point is that, if you look at the concentration distribution. So, this is just basically here the rank based on abundance. You see that you know predominantly sorry this is the rank of abundance, that this is the number of times we actually observe this protein in one of the 150 studies, we looked at.

Is there is a clear correlation between those proteins which a high abundant are seen more frequently than those proteins that are low abundant. Which also comes to the point that yeah now we can measure about to 500 proteins in every study, but really the question is how many times do we actually see this in every sample. I mean this is one thing to do with the concentration, but it also may have to do with the variance or the isoforms of particular proteins.

It was quite interesting for me, you know to see that for the foremost most common or most frequent proteins, albumin I guess that is to be expected. Complement factor 4, A2M or haptoglobin, you did not have all the peptides seen in all studies; so which means that, some peptides for albumin are in some studies so unique, that they are not common and concordant with other studies.

So, again you know here we come to the point that there is much more information in the peptides that we actually currently, as I think using. And next point, we need to make is about quality and I talked to some of you about, you know the challenges of you know, the information that you actually observe. And, and Sanjeeva and I discussed you know, the really having the important connection with the clinician obtaining a sample. So, you have actually control or at least the better understanding what happened to the sample when it was taken. Because if you just think about blood.

(Refer Slide Time: 08:17)



I mean basically, you can sort of dissect it into three elements; you have of course, the cells, you have some micro particles or also lipid vesicles. And then, you have something which we call sort of the cell free component, which is sort of serum plasma. And given that there are lipids that of course, are also important to be considered, you know.

If you just look at the proteins, the reason why you have the proteins in plasma could either be that they are actually actively secreted into blood because of the process that is related to it or they can be cellular a sort of cellular origin, meaning I that they should have been leaking out when the sample was being prepared. Or they have been shed from the surfaces because of the certain protease has basically took care of it, but there is also an important element which means that samples could be introduced into blood because of the preparation.



So, meaning you even have intracellular proteins which by changing the temperature from 37 to let us say 23 or 28 in India. You may just you know, introduce some of the inflammatory cells to secretes cytokines because of the change of environment. And that may have nothing to do with how the system sort of has been before. And, so, what we really advocate importantly is in a particular when you do this large scale projects.

(Refer Slide Time: 09:40)

**Sample Biobanks**

- Sample related data
  - site, date of sample collection. ...
- Donor related data
  - age, gender, diagnose at collection, ...
- Standardized sample collection and processing

1 analyte	→	N=77
1,000 analytes	→	N=235
10,000 analytes	→	N=286

80% power /  $\alpha=0.05$  level  
Andreas Garin, KI

**BBMRI.se**  
Biobanking and  
Molecular Resource  
Infrastructure of  
Sweden

**SciUp Lab**

**THE HUMAN PROTEIN ATLAS**

Is that of course, you want to know about the patient. We call it a donor, it is more of general. Of course, you want to know how old, what gender, what diagnosed and at the sample collection was, but you also need to think about what the sample comes from.

So, when was he collected, the location, how old was, how long was the sample being frozen, has it been freeze thaw, hence forth. So, really about the standardization of the procedures and obtaining information about the samples that you are analysing. So, that I think is really a key. And there are a lot of initiatives in Europe that you know trying to understand and trying to develop a pipeline for sample processing and handing samples.

Because if you just do play in statistics, without even sort of been thinking about proteomics. If you want to claim a sort of a significant finding and this is here just doing a simulation, where we I think took one of the major risk factors of cardiovascular disease with a power of 0.05, 80 percent and an alpha of 0.05. In order to just measure that one analyse, you need to have about you know 80 samples per group. So, it is a 160

samples in total. If you think about a proteomics experiment, let us say 1000 or 10000 you need to have up to 250, 300 sample per group.

So, meaning you are starting as your measurement to be sort of relevant, if I may use that word, when you 600 or more samples. So, this is not always possible. Some diseases you know, they are not that frequent. So, it is going to be extremely challenging to get up to that number, but of course, you know to get really understanding about the diseases that is one way of moving forward.

(Refer Slide Time: 11:18)

The slide is titled "Pre-analytical variables" and features a central horizontal strip of six small images illustrating laboratory processes: a person in a lab coat, hands using pipettes, a person in a lab coat, a person in a lab coat, a person in a lab coat, and a DNA double helix. Below this strip is a horizontal line of text: "Acquisition - Processing - Aliquotation - Storage - Withdrawal - Shipment - Storage - Design - Analysis".

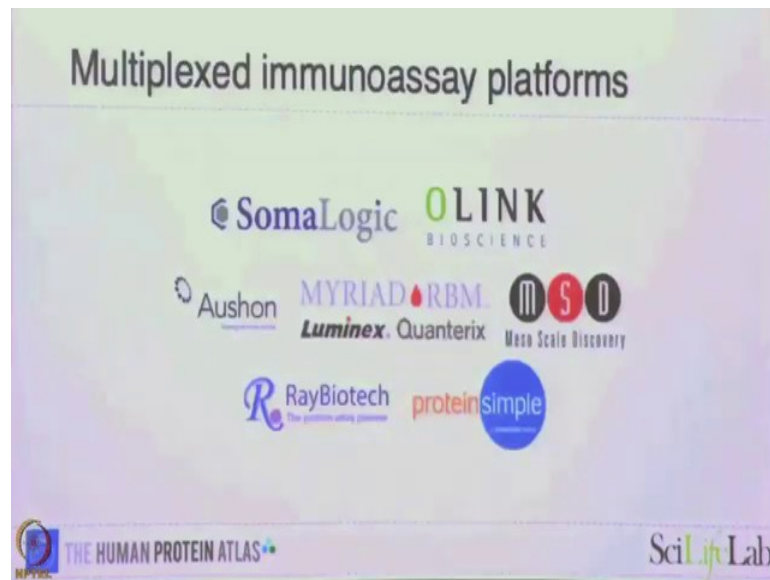
Below the text is a table with three columns and two rows of text, which appears to be a summary of variables and their impact on proteomics. The text is small and difficult to read, but the structure is as follows:

Pre-analytical variables	Impact on Proteomics	Reference
Proteinase K activity and protein quality	Proteinase K activity and protein quality	Proteinase K activity and protein quality
Proteinase K activity and protein quality	Proteinase K activity and protein quality	Proteinase K activity and protein quality

At the bottom of the slide, there are logos for "THE HUMAN PROTEIN ATLAS" and "SciLifeLab". A small note at the bottom center reads "Illustrations provided by Gunnar Tybring (KI)".

So, there is a whole science behind sample preparation variables or pre analytical variables, as we usually call it where people you know really try to understand what is the quality of sample and I think Matthias Mann's group has recent paper on bio archives, where they sort of you know looked at, where they basically try to you know separates plasma and did sort of different centrifugation segments and removed cells and respite them what is the contribution of cellular contamination in blood. So, I think that is important aspect, because cellular contamination is a factor that can hardly be controlled.

(Refer Slide Time: 11:58)

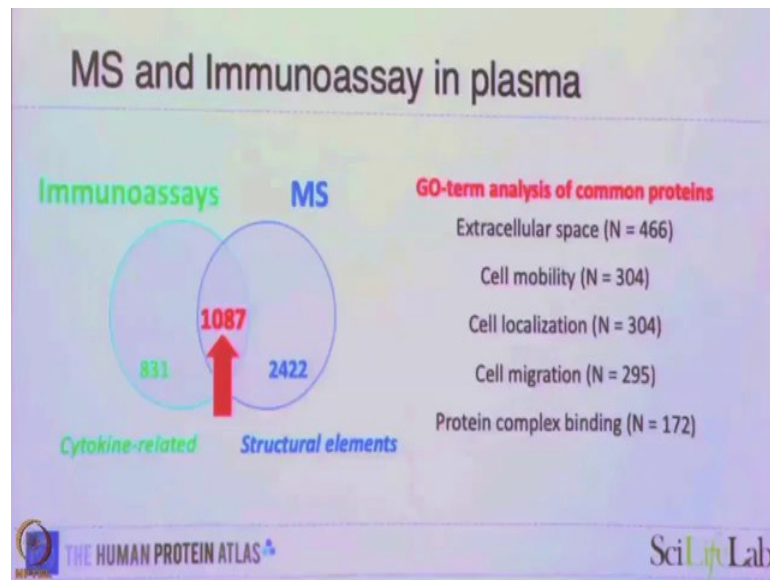


So, what we mostly work with and this is what going to talk in the second part of my talk, but I have some more slides in between is, is how we going to use different affinity based methods to study the plasma proteome.

And I think this is predominantly driven by companies nowadays selling kits. It is a bit different to what the mass spectrometry field is doing, where companies selling instruments and then this is academic environment they has to take care of them. So, of course, you have you know I think biogenesis or some other companies that sell or MRM proteomics that you know sell you kits, but I think there is no company that sells a kit for doing shotgun proteomics.

So, I guess you know it is a bit of a different ballgame because you have, you have a dependency on these companies. But the interesting thing about you know using affinity reagents in comparison to mass spec is that.

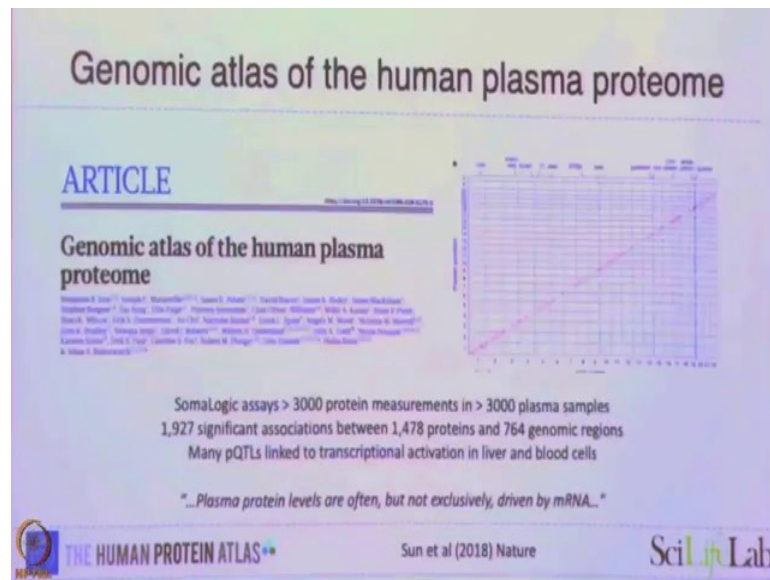
(Refer Slide Time: 12:54)



And here is a comparison about the proteins you can identify in mass spec as whereas, in amino acids is that you have a lot of the low abundant or annotated low abundant proteins that are actually measurable in immune assays. Compared to many structural elements which predominantly may originate from actually cells that are in your plasma, that which you can measure in mass spec.

Of course, a mass spec most oftenly and this is done purely on shotgun data is you of course take all information you get whereas, in an affinity based assays you pre select what you want to look for, all right. So, I mean these are really sort of conceptually different, different approaches.

(Refer Slide Time: 13:33)



And another aspect I mentioned before is really sort of the use of genetic data in combination with protein data. So, how much information about your proteins is already given in your genome. Well, of course, we know that this where the basic information lies.

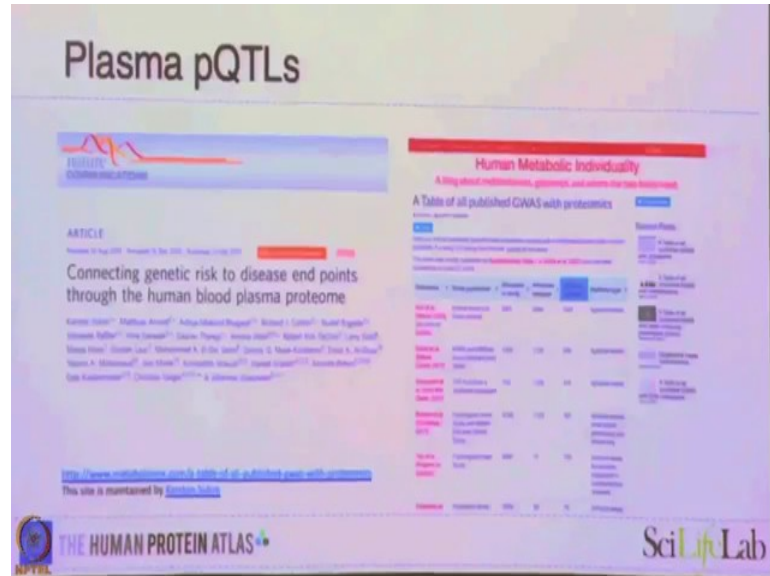
But how much of that information is actually been connected to what the proteins do at the end. I mean we have a whole machinery between the genetic and the proteomic information, but surprisingly and this is a study that for instance analogic has been done you know, surprisingly there are a lot of indications that that your genes over the variants of your genes tell a lot about the proteins that you measure in, in your sample in blood.

So, you know if you know somebody's genotype and if you know that genotype would be linked to a higher or lower risk of a certain disease and you know that genotype is also linked to a so called pQTL. So, quantitative trait loci, then you can say well that person always had a high risk of that particular disease, always the protein level was low which was maybe you know.

And then a slight increase of a low protein level may actually mean much more than if you would have the inverse case, where pro a person has low risk, but has intrinsically high level of a certain protein, right. So, it is really important to include much more data and I mean proteogenomics as one of the approaches, but others I mean you know

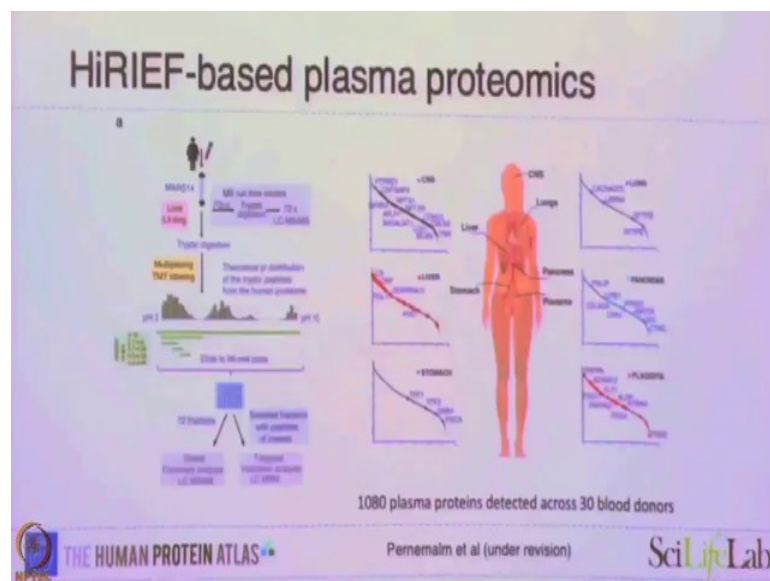
following the way. So, it is really and this is actually one of the resources built by Caster and Sewer, who is of Coops collecting all these informations.

(Refer Slide Time: 15:18)



So, it will be a I think a growing part in many of the proteomics study. Because if you do not know why a person has a higher level you know that might be one of the reasons for it.

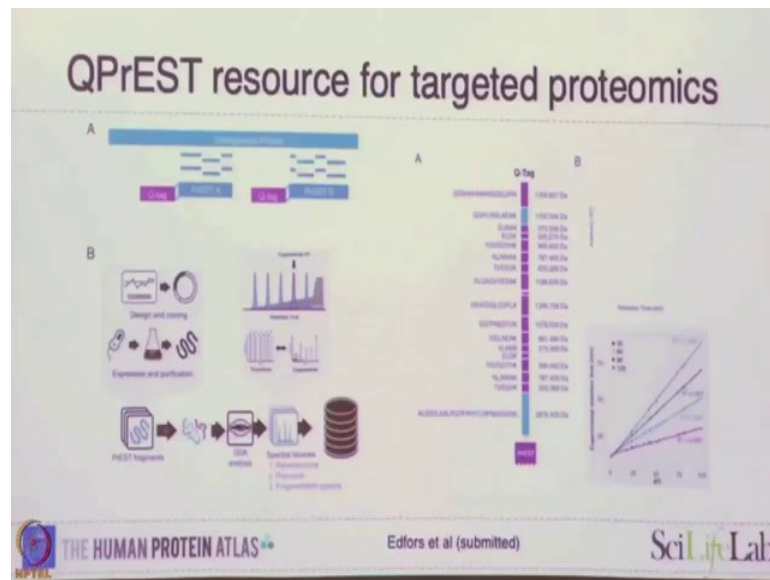
(Refer Slide Time: 15:30)



So, I have been involved in a couple of mass spectrometry related a project. This is one that is led by an electro scoop. So, they have used this high reef system. So, they have



(Refer Slide Time: 16:52)



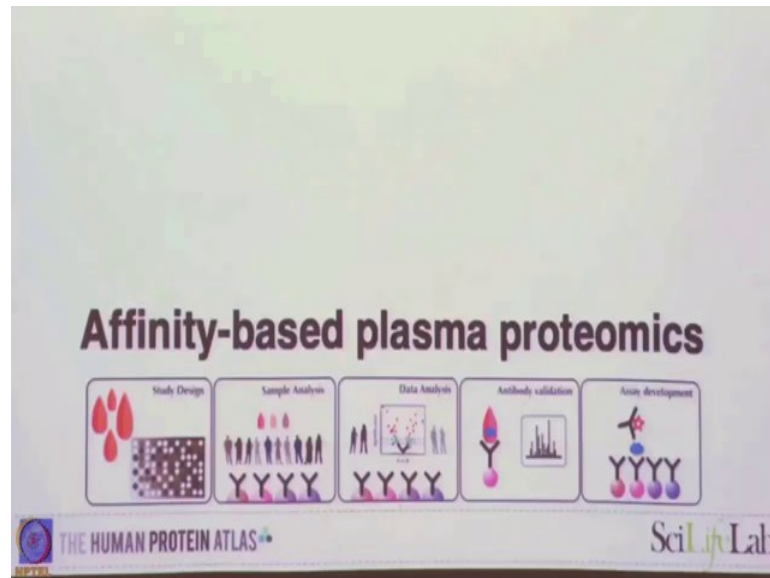
So, the protein atlas has of course, produced these atlases, it has produced antibodies, but it has also produced a lot of antigens and I guess Peter Nielsen, maybe some of you have met, talks a lot about using the antigens for quality assurance of antibodies or using these antigens for autoimmunity profiling. What we are nowadays using is, using these antigens as heavy standards for targeted mass spectrometry. And the reason is because we have these constructs.

So, here you have the endogenous protein and then we select these unique regions, we call PrESTs and all these PrESTs, by default carry attached a tag which we initially used for protein purification. But nowadays and this is basically the representation, this is a fantastic tag to do quantification of that specific sort of standard, right. Because this is a common tag for all the standards we use in our system and you can use this of course, for all your mass spec retention, time adjustments and so forth or what have you.

And now we have done this for about 25000 of these protein for these QPrEST, as we call it, the paper is all sewn by archives and hopefully will be coming out in a couple of weeks time. But, I mean you know using this as a pipeline that we can actually you know use that information to specifically build off the share of targeted proteomics assays for the proteins were interested in and we have shown this for a couple of examples also in plasma now in this study, that Frederick has been heading.



(Refer Slide Time: 18:20)

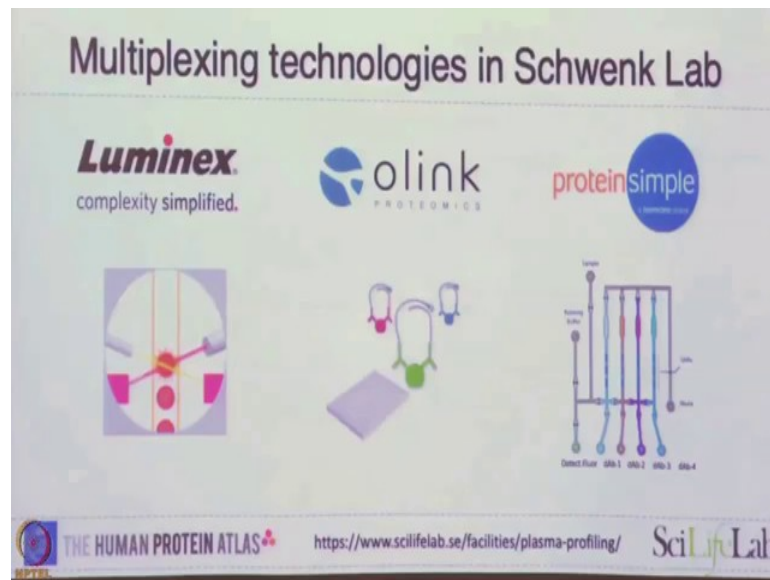


So, the main part of my talk is about sort of what would what do we do, sort of as sort of our core business, if that would be sort of a pitch to do for investors. So, the core is actually used, we do affinity based plasma proteomics and; that means, that you know we use antibodies or different types of affinity agents if they are available for doing protein profiling.

We care a lot about the study design, I think this is something I touched upon before. We care a lot about antibody validation, this has been a sort of a hot topic for us because antibodies have been criticized massively. There has always as in these concentrations been some truth to it, but I think we believe that there are opportunities to change, the perception.

And in part it has to do is redefining or explaining more or less what the antibody is actually capable of, alright. So, an antibody is not an off the shelf universal tool that will solve all your problems, right. An antibody is something you need to know where to use it and for what to use it, right. So, an antibody good for western blot may not work in amuse the chemistry or ELISA. That is not understanding that not many people have, unfortunately.

(Refer Slide Time: 19:40)

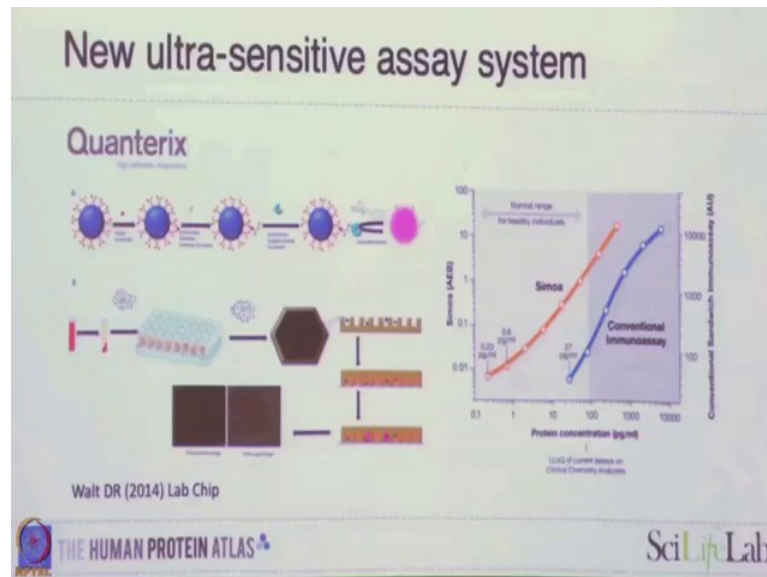


And so, my lab currently runs three different sort of technologies. So, we use Luminex, as sort of our go to platform because it is open access for us. We have all the equipment, we have 10 years of experience of using it in various aspects for protein as well as out antibody profiling. It is for us sort of really easy to use.

But it has some limitations in terms of sort of, in the way we use it quantification and sensitivity. We also for about 2 years now, I have Olink as the technology that we run and we do this for different types of service projects or our own research projects and we also have an interesting technology offered by proteins simple, which uses micro fluidics and the nice thing with this system it is basically almost fully automated.

So, you do not have any user interference that gives you a really excellent batch to batch precision and it actually is a system that we think could be useful for clinicians because they do not want to think about how to run the experiment, they just want to get the data, right.

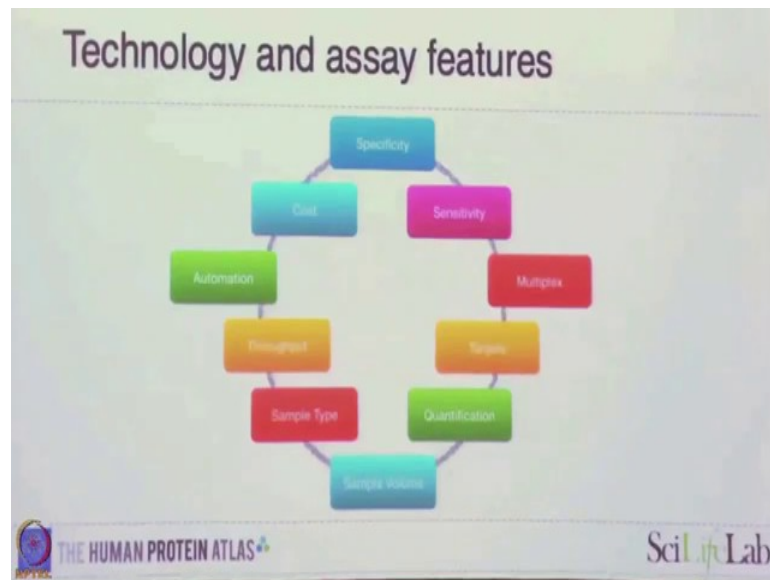
(Refer Slide Time: 20:45)



We are also going to include Quanterix as a new technology probably someone doing this year and Quanterix, is basically also a beep based ELISA, beep based you know as a system just like Luminex. But they have a different way of readout in terms of that they use an enzyme to create criminalists flow fluorescence and they also have a different mode of detecting or counting their detection, meaning that instead of you know measuring the sum of all signals that are sort of obtained.

They actually do they call it digital counting. So, they are not they count the number of particles which actually emit a light at a certain rate level and there that is what they usually call, sort of they are sort of digital amplification range which is sort of then giving them a hundred fold improved sensitivity. At the cost of using more samples and more antibodies, which it is not as well communicated.

(Refer Slide Time: 21:46)



So, of course, and there is a whole portfolio of sort of options, which technology to use and this is usually when we have sort of meetings with users of the facility that I am directing you know, what do you want to do, what is the specificity the cost a number of targets you want quantification, how many samples do you have available, what are the volumes and so forth.

So, it is really sort of a ballpark of different features that you need to consider when you choose a certain method for your application. Which again you know, is a bit of a different concept to mass spec given that you know you can probably choose many systems for many applications, right. Of course, you need to tweet them and some may be less suitable than others, but in theory I guess you from all mass spec you would get some data, all right.

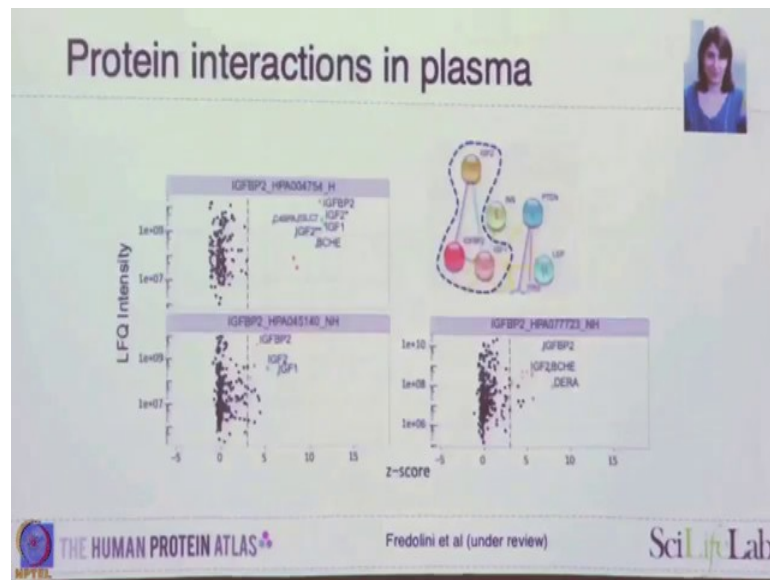


And we have been working on using actually mass spec as a read out for immune capture data and this is I guess most of most oftenly done, you know using cellular systems. So, where we have actually tagged, an antibody towards a tag and the tag is fishing out a protein which has been introduced and people you know calling sort of the crap home, you know the all the proteins you identified even though they I do not have anything to say. So, we have been putting all of this into a plasma.

And this is a just study led by Claudia. So, here we have done more than 400 IPs and built sort of a library of data to judge whether an antibody specifically enriched a protein in plasma or not and these are sort of, these sort of enrichment plots. So, to the left you have you know this crap home, the part that is commonly a fountain and every enrichment that may be due to, you know proteins sticking to the beads, but then to the right hand side you know we chose the Z score of 3 as a cut off. You see some on target detection, you see some code targets meaning proteins are co enriched either because they have a similar sequence or they actually do interact between find very interesting.

We also see you know off target interactions to proteins that are more abundant than the protein that we presume the antibody would binds to. And we actually also have cases where we have no target meaning there is no specific enrichment. So, which I think in a way is interesting because either this could mean that if there is an enrichment that the target has not been sort of detected in mass spec or the protein that you know it is simply too low abundant to be sort of reaching assays core that is of relevance.

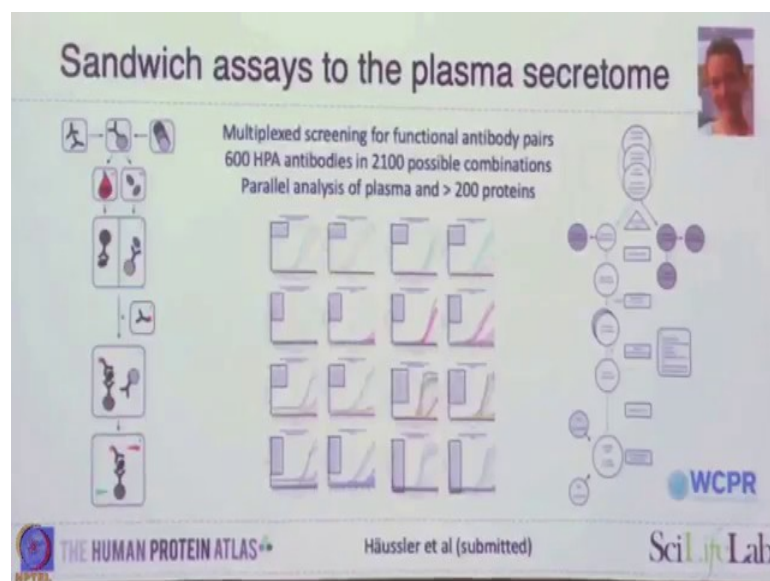
(Refer Slide Time: 24:50)



Yeah as I mentioned, proteins directions we find interesting. Here we know for this insulin growth factor binding family, they sort of interact with another and as shown here and using the string database you have IGFBP2 interacting with IGFBP1 and 2.

And as you can see here using three different antibodies, we could see here is IGFBP, that they actually interact. But we also could to claim new interactors with this BCHE as well as DERA as proteins that are relevant for us, for these complexes.

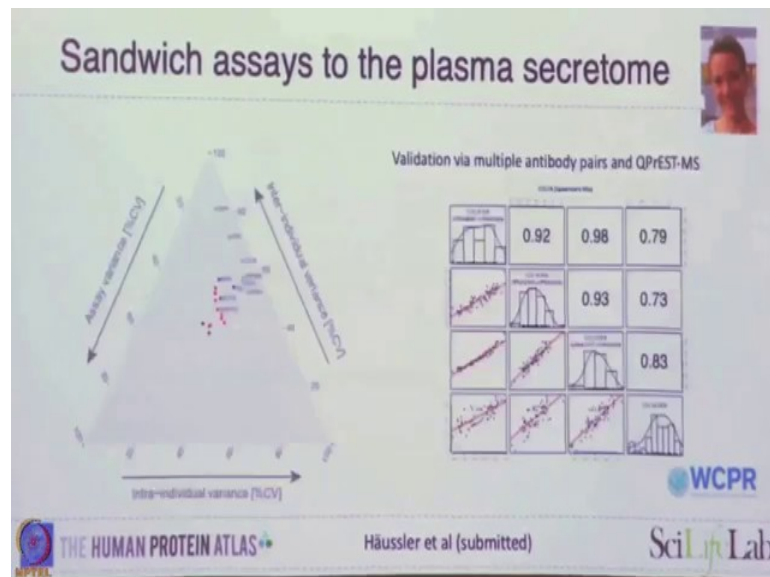
(Refer Slide Time: 25:20)



What we also do you know and sort of going back to our sort of most accessible technologies using Luminex. And here, Rakhna has you know screen more than 200 antibodies, sorry 200 proteins using more than 600 antibodies to find which are actually suitable sandwich pairs.

So, using both for capture and detection, this also now paper which you can find on bio archives and hopefully the reviews will like it. So, we have done sort of you know at the sort of a long term procedure, two screening rounds and meaning this a substantial amount of work with a couple of people involved.

(Refer Slide Time: 26:01)



But at the end of the day sort of it led us to this triangular chart here, where we actually looked at longitudinal samples and the precision the assay provides in this context. So, it is a bit of difficult to read, but basically we looked at sort of what is the variance of the assay in terms of the technical position, what is the difference between the individuals that we observe over time and what is the difference between the individuals themselves, right. And as you can see we have a couple of nice proteins here those ones, that are high and red, highlighted in green.

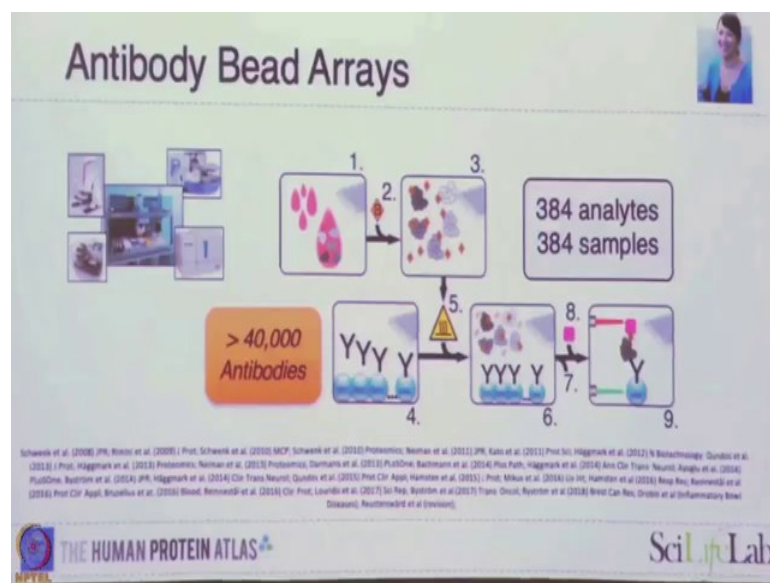
Which are those proteins that are, we can measure precisely, that are stable over time, but they vary a lots between the individuals. Meaning that there is probably a genetic component or some sort of personalised component to it that makes these proteins more interesting than others. Again, we do took a lot of effort to do validation and here is to



some sort of correlation chart, where we compare the different sandwich as a data here in this case for a protein called, I think it is CCL 16.

We have three different assays, we developed in house and this is the assay offered by Olink. So, we have a pretty good precision using completely different. So, this is Olink is a solution phase protein proximity extension assay whereas, you know we have a classical Eliza, where you capture on a bead, you wash and then you add your detection antibody.

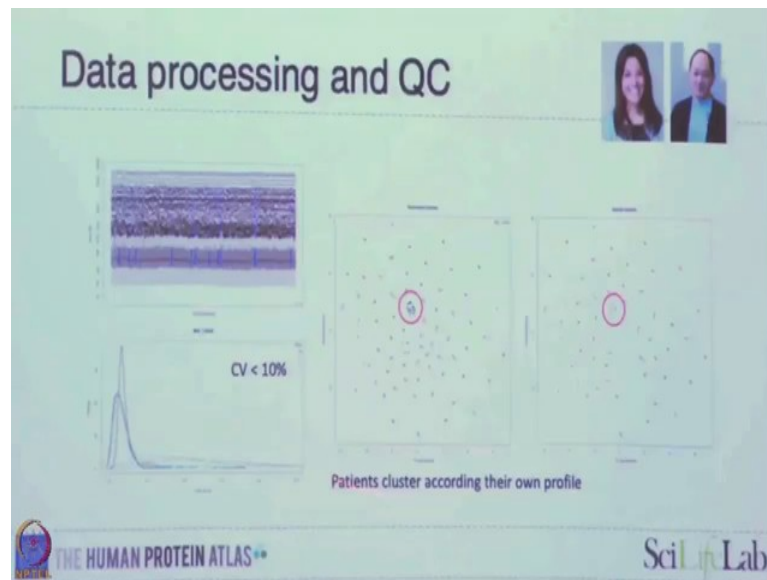
(Refer Slide Time: 27:20)



So, our main workhouse has for many years been these antibody bead race. So, we really use sort of the high multiplexing capacity of the Luminex system, we have a couple of liquid handling devices to do upfront sort of sample prep. So, here the idea is that you instead of using two antibodies, you basically you know label your sample are you doing in different types of EMT assays and then you sort of have a bead array which has 384 different antibodies.

You fish out the proteins that you can find in the solution and then use biotin to detect whether the antibody has enriched, the protein of interest. We have been working a lot on sort of you know, getting the data analysis and processing right.

(Refer Slide Time: 28:07)



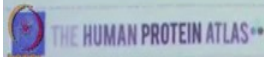
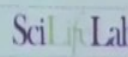
And it is done by mainly Mun Gwan, who is research at my group and we have a pretty good idea about the data, the precision and sort of the accuracy. And this is, these are sort of two teeny plots showing basically the same data, but here basic business these are the replicates, where all the other data points are samples taken from the same individuals every third month over a period of one year.

And you can also maybe see that you know all these individuals actually clustered together. So, show it using our all data and you know we know the phenotype or the plasma proteome phenotype that we measure is constant over time using our data.

(Refer Slide Time: 28:44)

### Examples of Projects (N > 200)

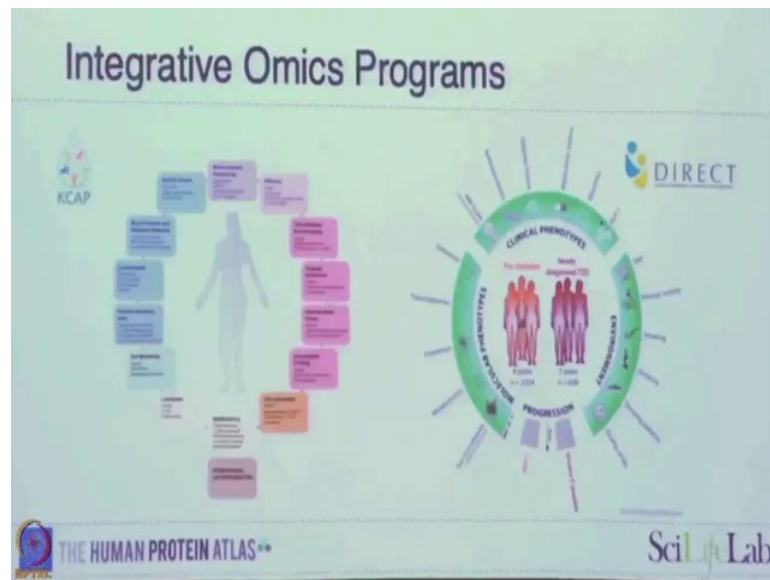
Projects in Verification Phases	N	Study Sets	Ab	SA	ICM	WB	FA	pQTL	Publication Status
Ageing	5,000	X		X	X	X	X	X	in progress
Diabetes (incl. longitudinal)	4,200	X	X	X	X		X	X	in progress
Venous thrombosis	3,000	X	X	X	X		X	X	Blood 2016
Breast density	3,000	X	X		X		X	X	Breast Can Res 2018
Pancreatic cancer	3,000	X	X	X	X		X		in progress
Prostate cancer	3,200	X	X	X	X	X	X		MAF 2016, Transl Prot 2014
Drug induced liver injury	3,200	X	X	X	X		X		Clin Res 2016
Renal impairment	3,200	X	X	X	X	X			JPR 2011
Cholesterolemia	3,000	X	X	X	X				Sci Rep 2017
Multiple Sclerosis (CSF, plasma)	1,120	X	X	X					Proteomics 2012, JPR 2014
Childhood Myopia	800	X	X						PLoS Pathogens 2014
Wellness (longitudinal)	400	X	X	X	X		X		in progress
Sarcoidosis (BAL & serum)	500	X	X	X	X				Resp Res 2016
Atrophic Lateral Sclerosis	500	X	X	X	X				Annul Clin Transl Neurol 2014
Duchenne muscular dystrophin	400	X	X						EMBO MolMed 2014
Pre-analytical sample processing	800			X					J Prot 2013
Osteoporosis	225	X	X	X	X		X		Clin Prev 2011
Neuroendocrine Tumors	200	X	X	X	X				PLoS One 2013
Immunoaffinity (Wid blood spots)	200	X	X	X	X				J Prot 2015

So, we have been involved in a couple of larger and I think actually growingly larger projects which you know we try to have multiple study sets, meaning samples coming from more than one location, using more antibodies actually the building our own sandwich immunoassays. For those candidates we identify to use immune capture mass spectrometry as a way to validate. We still have western blots sometimes as a go to option, but it is actually less relevant nowadays for our approaches.

We do validation of antibodies using peptide or protein arrays. Sometimes this is helpful to certify the selectivity between different off target candidates and more I think interestingly for us in the future will be to do this pQTL, sort of the g west analysis to understand what is the genetic component behind this, these studies that we have, that we will be performing.

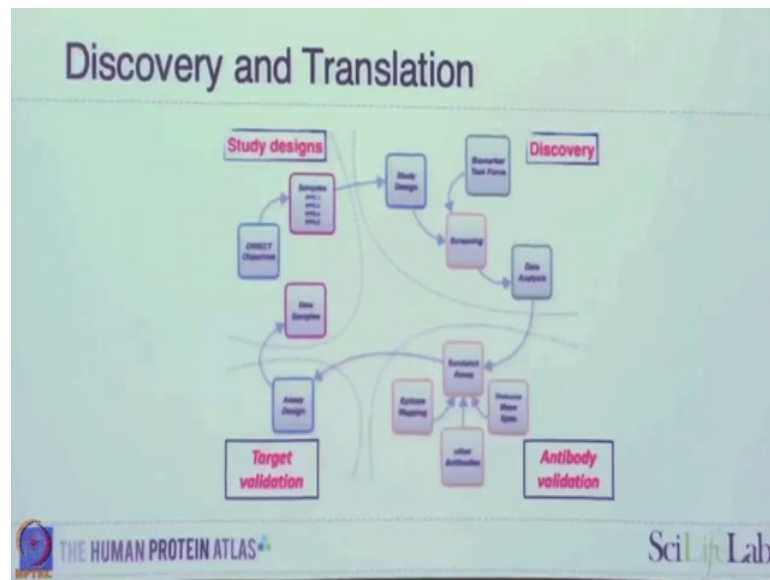
(Refer Slide Time: 29:40)



So, these are two of these initiatives, one is a wellness project which is headed by Matthias and Backstrom. Where, we have taken those 100 subjects, did all the different omics and clinical measurements where we looked at them every third month over the course of one year.

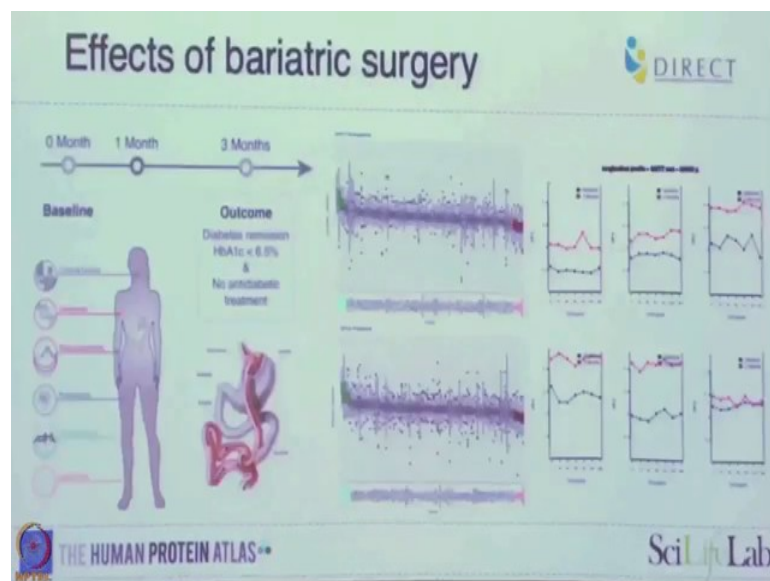
I am also part of a very large new projects with different Pharma companies and clinicians from all over Europe to do basically, the same or a similar type of molecular clinical and an environmental phenotyping in the context of pre diabetes and diabetes progression.

(Refer Slide Time: 30:17)



And again, an important aspect for us are these four elements. So, it is a study design how, do we sort of proceed in terms of you know randomization, how do we get the number of samples right, how do we do the discovery; I mean we have to choose which are the interesting candidates because we make that pre selection, how do we do antibody validation or actually building new assays for target validation and then to you know, go back and sort of study new samples again to prove that our hypothesis is actually valid.

(Refer Slide Time: 30:48)



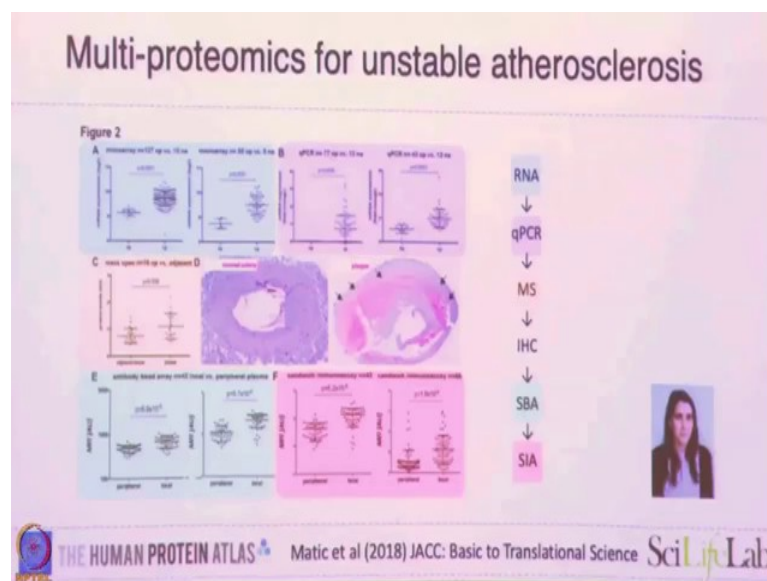
So, one of the studies we have done here on biotech surgeries, a biotech surgery is a major type of intervention for people that are very obese and that is usually at that high risk of the diabetes. So, the idea is that the, this surgery induces weight loss and along with weight loss comes that patient is not longer defined as being diabetic.

And, and we want to understand where the proteins in plasma can give us an indication about either will a patient you know be benefiting from that surgery, that is sort of we are looking at remission which is done using a multi omics approach by one of the postdocs in my group, but also, how do how do proteins change over time; pre and post-surgery.

So, because we have looked at the patients at baseline and as well as following surgery. And we actually could see that there are a couple of proteins or that are consistently increasing knowing that there is an interview individual variance, consistently increasing post-surgery and we looked at 3 months as a time window because between 0 and the 3 months there is a lot of processes that are sort of overruling, sort of the phenotypes we are interested in.

In particular those, that are related to wound healing, right. So, if you measure a patient after day after the surgery, a lot of the things you measure is actually the patient responding to the surgery, not responding sort of on a metabolic level. And we could interesting interestingly see you know some proteins actually sort of do change.

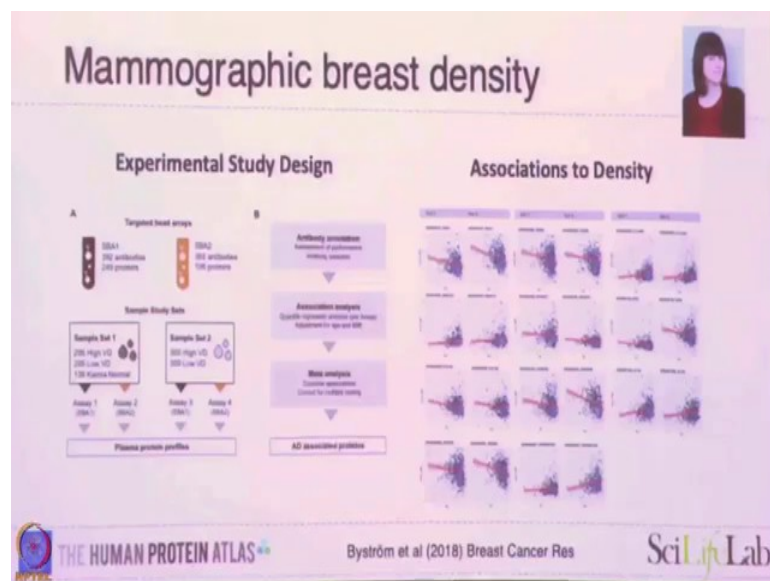
(Refer Slide Time: 32:30)



Also, in sort of the opposite direction, which means that they are actually decreased in abundance. Another type of multi omics approach we have done in the context of unstable atherosclerosis. So, here you have basically the coronary plugs that you are, some people develop and of course, there is a risk that some stable, some are unstable which means that you might actually you have a higher risk of stroke and heart attack. So, with a group of clinicians who have done sort of microarray and QPCR and identified a couple of candidates which they could validate in using mass spec and tissue.

We then took on this type of target and actually could validate the same sort of observation using either this suspension bead array, the screening approach as well as we build a sandwich assay to measure that same difference in plasma samples. So, we really sort of brought from sort of early DNA, RNA detection down to sort of applications in plasma.

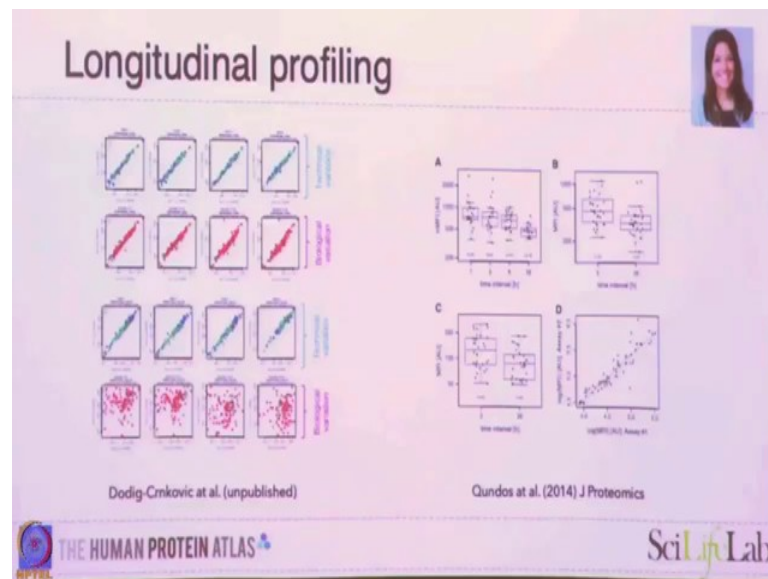
(Refer Slide Time: 33:31)



We also did a large scale study on mammographic density. So, this is a study again we were sort of switching a bit sort of disease areas here, is related to cancer and in particular, is a risk factor for women in the western world. So, if you lose density in your breast, post menopause is actually a very good protective indication. But if the stiffness of the breast stays after menopause, there is a high risk of breast developing breast cancer. But nobody understands what is this density.

And we try to find using association study on a cohort of about 1200 women, whether we could identify features that are consistent and we found a couple of interesting protein related proteins related to the extracellular matrix as well as to proliferation levels that could indicate you know that there is actually a loss and increase intensity visible in the plasma proteome.

(Refer Slide Time: 34:38)



Again, sort of one aspect that we have been working on frequently is this longitudinal profiling and here again, want to bring up something I mentioned earlier which is sort of how consistent can you actually measure a protein.

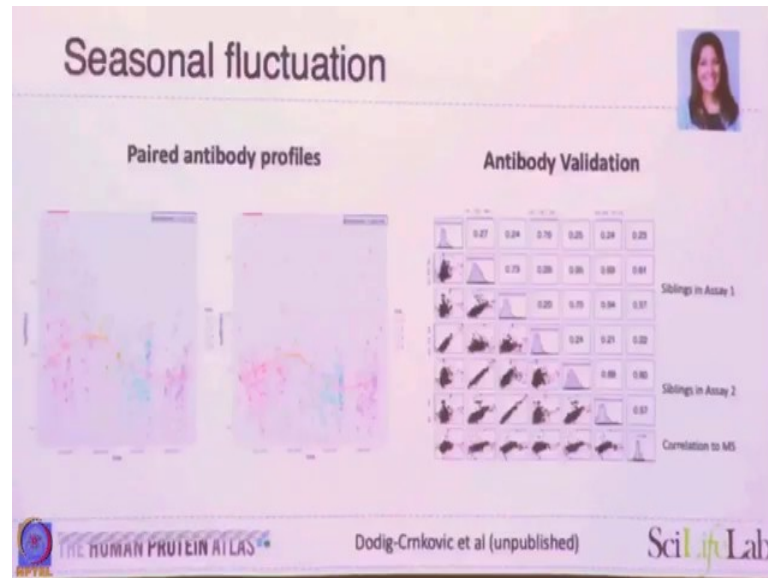
So, here we have looked at basis Bs. We correlated the data we generated for this protein across these four visits and we see that you know the protein is the measurement is pretty stable over time. But then if we look, if we compare the data between the visits, you can see here we have extremely high precision as well. So, meaning that protein can be accurately measured, and it is very stable over time.

The second protein here is again, we sort of replicates this as a couple of times and you can see the precision of the measurement is very good. But if you look at the correlation of the biological variation as we call it, where you compare those data measured at visit 2 versus visit 1 and so forth there is basis 0 correlation which means, each blood collection introduces a factor which cannot be replicated, right.



And of course, if you have a biomarker you know which looks like this on a technical scale, but if it is impossible to replicate because it is and this protein we know is part of the skeleton, a part of the smooth muscle system. So, we know it is actually coming from puncturing your skin and the vein. But again, you know it is the protein you would not have sort of considered, but you can actually measure it.

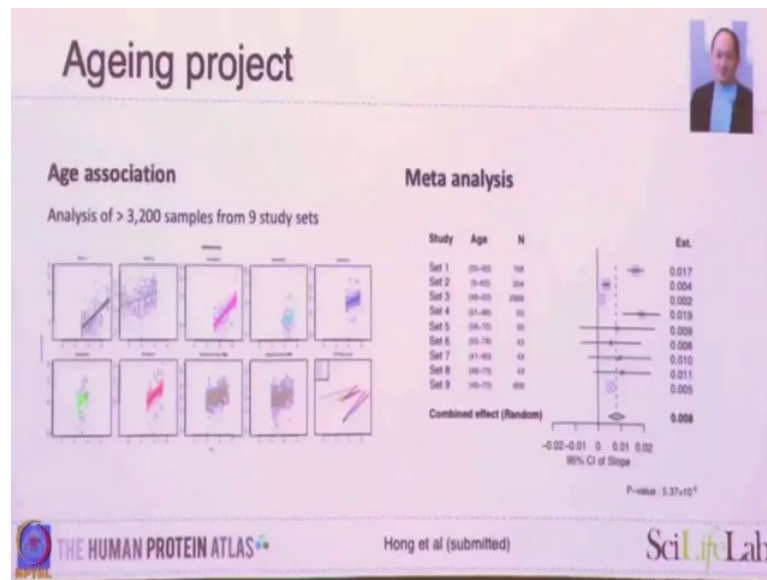
(Refer Slide Time: 36:06)



We also looked at seasonal fluctuations. So, this is of course, interesting you know. In a sense that what are the differences if you measure your protein during winter compared to during summer. And may just be the seasonal have an effect on your protein levels and just assuming let us say this would be sort of a cut off level for this protein here. You know, here you would be actually above cut off and the doctor may say oh you. We may need to check up on you a bit more.

While as during summer, you know you actually have a much lower level. So, of course, these parameters which also I mean relate back to the time point and the age of a sample are important things you need to consider when you do your measurements.

(Refer Slide Time: 36:47)

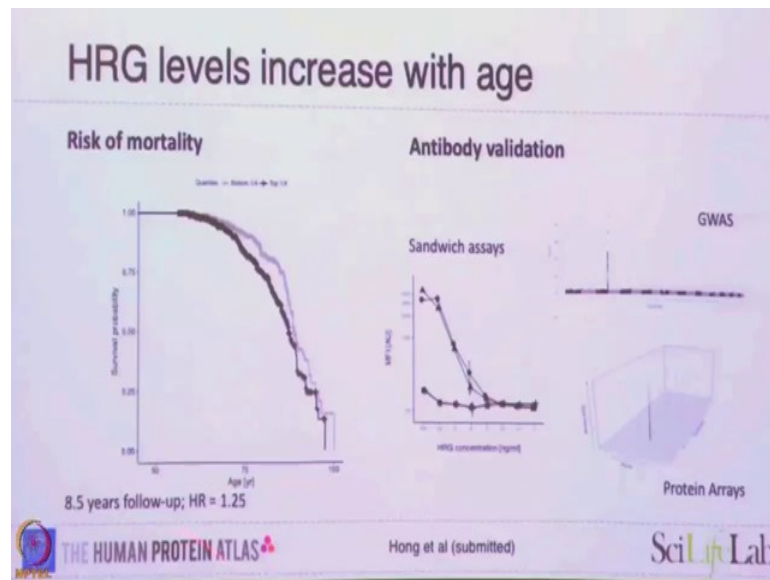


I am going to finish with this projects. So, this is also something which hopefully will be coming out in a couple of weeks time. We have received some very good comments from the first round of review that will be certainly able to manage to handle this. So, what we have done here is, basically this is a study is a have been a bit a bit of a hobby study, actually. Because in most of the projects, we have we know the age of the person which donates the samples.

So, we just started to collect you know, a couple of studies and actually now it is it ended up to be, it is actually 4000 in total, where we actually looked at the same or the same protein over and over again. And as you can see here, in all these studies the slopes may be different, but in all these studies we could see a consistent increase in trend over time. So, basically sort of using this as a additional passenger in the different studies.

We sort of you know as a by product, more or less found a protein which is associated with age. And we have been validating this and this is the meta analysis that we did. So, the p value is, I think far better than most studies you have you have you seen, because it is really consistent across time and. And what we also looked at is, what I mean it is one thing that this protein HRG tells us about your age, but it also tells you much more about the risk of dying.

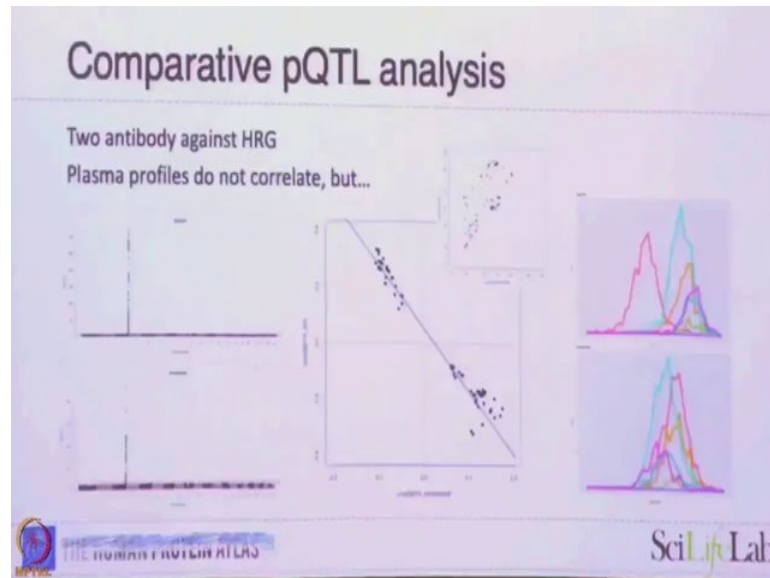
(Refer Slide Time: 38:20)



So, it seems that elevated levels of this protein increases your risk of dying compared to low levels. Specifically, it up to 8 and a half years prior to death and this is sort of on a all cause mortality. So, it is not linked to any particular cause of death like cancer or cardiovascular disease.

Of course, we did a lot of validation of the antibodies. We actually also, to retrieve us and this is actually the protein array that we have run with Peters group. So, these are 20000 spots and you can only see a single peak of this antibody, which was quite surprising. But you know, we know it is we know we can measure HRG and, but then what we actually also did.

(Refer Slide Time: 39:12)



And this is fairly new for us is, again we took this genetic information we had about these individuals and there was another antibody against the same target. And when we correlated the slope. So, meaning what you do in this peak as a base they have a boxplot with three different groups.

So, it is the AA, TA or AT genotype and then just and then you just superimpose sort of a trend. And what we could see that these two antibodies, they have the exactly same list of pQTLs. They have an opposite trend in their association and that is also seen here, by these distribution plots meaning, sort of the red genotype is lower for this antibody whereas, the red genotype is higher for this antibody. So, it is completely new data and nobody has done something like this before. But what it says, we have not really fully understood, but what we believe right now is that every person has a particular variant of that protein.

And that the antibodies have a particular affinity to that protein variant. So, we think and it is likely that many proteins we nowadays study, we do not actually, they do not actually in reality differ in concentration, they just differ in the variant they are. And that the different methodologies may be mass spec or affinity based assays.

Just think or it just reports different signals because it is a different variant and I guess it is a particular challenge for both the assay types in mass spec because when you look at the libraries that you use to match you data, this is this is done on canonical sequences.

Of course, you can do protein genomics approaches, but that is not always possible. But it will be in the future, because you need to have that understanding to know what to what to look after right.

And it is the same thing for affinity assays. If a small variant if you have an exchange, let us say you would change a hydro fill hydrophilic amino acid and you have a non synonymous mutation, meaning that that suddenly becomes from serine you change to a proline. You know, you will change the behaviour of that protein either in the way it is been recognized in your test or how it actually interacts with other proteins and thereby, may be more accessible for let us say you know different types of measurements.

(Refer Slide Time: 41:54)




I like to thank you for your attention and yeah of course, all these people.

(Refer Slide Time: 42:01)

**Points to Ponder**

- Affinity proteomics is a field of proteome analysis based on the use of antibodies and other binding reagents as protein-specific detection probes.
- Affinity-based methods can be use to enhance biomarker discoveries, validation and integration from basic research towards clinical use.
- Environmental condition, race, sex, age of a patient and even sample storage condition need to be taken into account in analysis.



NPTEL IIT Bombay

In summary, today you have studied the human plasma proteome using affinity based methods which could enhance biomarker discovery, validation and integration from basic research towards clinical usage. How atlas antibody from HPA project can help us in getting detailed understanding and background of the affinity based methods.

Doctor Jochen also provided a brief understanding about GWAS and how patient information is important to understand data set variations. In the next lecture, we will listen a clinician doctor Sachin Jadhav, who will talk to us about clinical considerations for omics studies.

Thank you.