

An Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. Henry Rodriguez
Department of Biosciences and Bioengineering
Director, officer of Cancer Clinical
Indian Institute of Technology, Bombay
Proteomics Research, National Cancer Institution

Lecture - 01
Proteogenomics overview-I

Welcome to MOOC course on Introduction to Proteogenomics. What is more powerful, genomics or proteomics? I think this has been a long debate, what is more robust, more powerful; proteomics based investigations or genomics based information? However, today's distinguished scientist Dr. Henry Rodriguez is going to provide you a new answer, that a field of proteogenomics which is now emerging can provide us much more meaningful and more powerful information. Dr. Henry Rodriguez, is a director of Office of Cancer Clinical Proteomics Research at National Cancer Institute: NCI, National Institutes of Health in USA.

Dr. Rodriguez research has focused on understanding mechanisms of cancer and age related diseases, including development of molecular based technologies in basic, translational and clinical sciences. Dr. Henry Rodriguez has led to the development of NCI's clinical proteomic and proteogenomic research programs which today includes the world's largest public repository of proteomic sequence data and targeted fit for purpose assess. His efforts has led to the formation of cancer moonshot initiatives, the International Cancer Proteogenome Consortium: ICPC and the Applied Proteogenomics Organizational Learning and Outcomes APOLLO network, which he developed and co-developed.

Dr. Rodriguez has been very supportive to also include India as a part of ICPC initiative and India has now become the 12th country to join this consortium to look at the cancer proteogenomics research for cervical, breast and oral cancer. Dr. Henry Rodriguez will give an overview talk of Clinical Proteomic Tumor Analysis Consortium: CPTAC which is one of the efforts from NCI to accelerate the understanding of molecular basis of cancer through the applications of large scale genome, proteome or proteogenomic

analysis. He will also brief about how NCI is working and taking the translational cancer research to the next step.

Dr. Rodriguez will talk to us about how genomics and proteomics together in the area of proteogenomic could make much more meaningful impact. The importance of proteogenomics and how the robust field can reveal answer to different biological questions will be addressed. He will then bring various facts that laboratories worldwide should follow a standardized workflow to obtain reproducible datasets. Dr. Rodriguez will also talk about how proteogenomics is providing new prospects in recent projects of CPTAC like ovarian cancer. So, let us welcome our distinguished colleague Dr. Henry Rodriguez for his lecture.

So, welcome everyone, my name is Henry Rodriguez. I am the director for the National Cancer Institutes Office of Cancer Clinical Proteomics Research and I have to admit its been extremely exciting and flattering watching over the past several days this idea of looking at the proteomics based information. And, trying to now blend it more and more with the rich history that has come out over the past 15 years in the genomics landscape.

So, what I thought that I would do is to give you sort of an overview of what we have been doing, now at the National Cancer Institute really for about 12 years, where we see it going in the future for about another 10 years. And, at the same time kind of talk about how we have taken what we have developed at in the US through this program called CPTAC. And, now we have sort of expanded it and it is really nice to see how India has become the latest partner within this international effort.

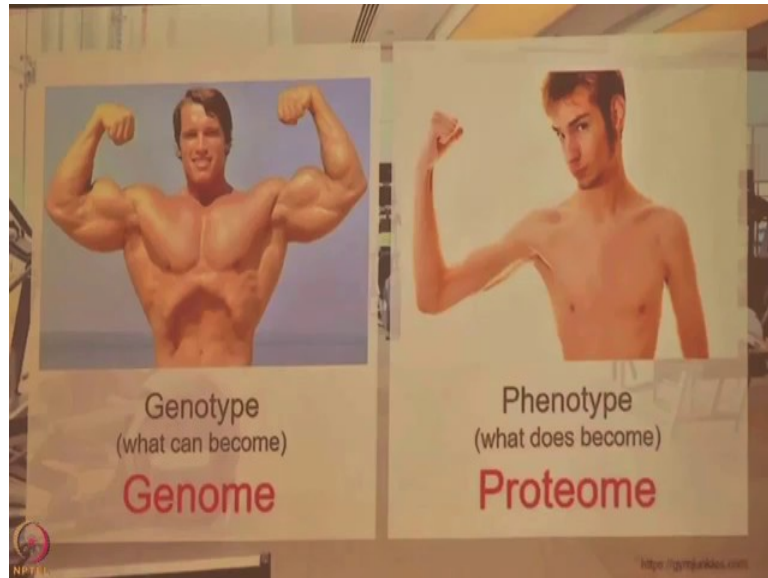
(Refer Slide Time: 04:41)



So, let me do this, the first slide that I am going to do is sort of a cartoon because so, this one presentation that I saw two days ago and people were quite nice and they were very scientific ok. And, they explained to you genotype, they explained to you genes, then they talked about phenotype, but here is my simplistic perspective of trying to understand a genotype. And, how it rolls up ultimately to a phenotype which is what you want to get your hands on.

So, imagine if you are at the gym. So, in a way if you want to look at what genotype is which is going to be representative of your genomes; this in way could actually be your genome which is your genotype, which kind of tries to represent it is your blueprint. And obviously, when you have a blueprint what you are trying to do is to say this is what I could potentially could become, that is your genotype. All the potential is there, but as people know we all aspire to do certain things and sometimes those things actually do not come to fruition.

(Refer Slide Time: 05:37)



So, the reality is your genotype which is what you wish to become as in this individual which is Arnold Schwarzenegger, the phenotype which is actually your functional space and today you can kind of look at it as a proteome; this could actually become your phenotype. So, not always do you get what you want, to put it in a good way; however, though that is actually because, I think ultimately to understand the different states between the genotype in the phenotype; it becomes really important to begin to blend these worlds together. Quite frankly, I think if you study only the genome and then you ignore the proteome or quite if you look at the proteome, you completely ignore the genome; you are going to be missing a tremendous amount of biology.

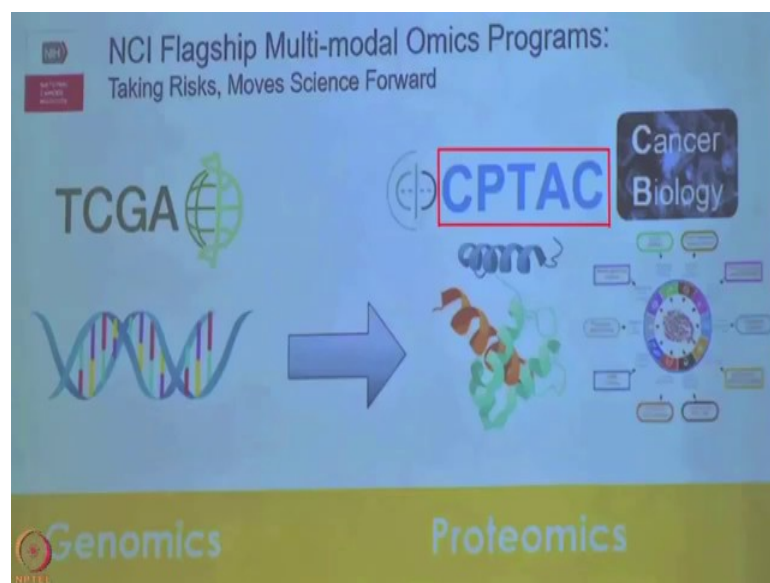
And, I hope in the next 40 minutes, I can give you an example and how now we have seen that in the space of oncology that is the case. So, more and more as technologies are becoming very mature which are going to blend these world's together. So, this is the history sort of the genome that I see it from the perspective of the National Cancer Institute. So, I actually got recruited to NCI just about 12 years ago and one of the things that I kind of liked about it is, is that I love organizations that enjoy taking risk. So, for very conservative organization they kind of did take a risk; politically because it did cost the lot of finances in the space of omics based research.

And, if anyone talks about genomics lot of people will talk about the cancer genome atlas. So, the cancer genome atlas actually gets officially launched in 2006. So, the dates now become very important here and TCGA in a span of 10 years of course, they had a

lot of capital to do this. But, in a span of 10 years, they did an amazing thing; they basically catalogued about 34 different cancer types.

These are all solid tumors and they actually went through about just a little bit over 14,000 individuals to achieve that goal and all the information they placed it in the public domain so, that is good. Here is the part that a lot of people do not know about the history of NCI; actually whenever actually trying to come up with this idea what do we look at the genomics, they all along did not want to do genomics in isolation.

(Refer Slide Time: 07:45)



They actually did want to go after the proteomic space and in that and actually what they ended up doing was at the same time that they launched the cancer genome atlas which was mandated to go after biology; they launched a proteomic based effort. Now, a lot of people knew about it, but that program at the time which is now today kind of known as is referred to as CPTAC.

Now, the reason they want to do it was quite simplistic or in the early 2000, the first draft of the human genome project gets released again it is a draft, but that really raised the interest of a lot of oncologists in the US; especially our cancer center directors. And, they basically did these series of workshops and one of the things that came out of these workshops, they said we need to now begin to explore omic based technologies in cataloguing different cancer types and they made it very clear; we want to go after genomics and proteomics.

Now why proteomics? Who are the first one you connect with actually understand which is what was talked about in the days prior was, you need to get an understanding of the underlying biology of that disease. And, if you talk about biology try to understand the different pathways and not just taking your RNA seq data and computationally predicting from a bioinformatics perspective, what the abundance are of proteins, or more specifically what those modifications would be, it is never a one to one correlation.

So, they knew they had to understand the underlying biology, before any other biology could even move potentially towards patient care. The other reason was is exactly what I said patient care, if you ignore the space of IO, which is Immuno-oncology; the vast majority of all our patients are still being treated with compounds that are typically are chemo based. And, those compounds are actually do not target in fact, the vast majority do not target DNA. There is very few that play this the intercalating the DNA, the vast majority will always go after a protein.

So, they need to understand not just hey my target binds here, but again trying to understand off target sites and all the wiring of the biology and all the off rows that you could get. But, here is now what happens around 2003 using the instrument of a mass spectrometer, a publication gets released into the public. And, that actually looks at early stage ovarian cancer, they actually did not identify the proteins that they were measuring. They basically looked at these pattern recognitions and based upon that they basically argue that simply looking at proteins by a pattern, ignoring the genomics information were able to identify early stage ovarian cancer.

And, I think they talked about like 90 percent specificity with a 99 percent of specificity built in which is incredible, if you think about it because 99 not existed within the DNA diagnostic space. Well, it turned out it was too good to be true; there were errors at all levels of this. So, the NCI decided to do was, when it came to proteomics, they did not move forward in 2006 when they created this to go after biology that was taken off the table. So, they basically wanted to go after the analytics and determine, can you standardize these powerful next generation methodologies; predominately a mass spectrometer.

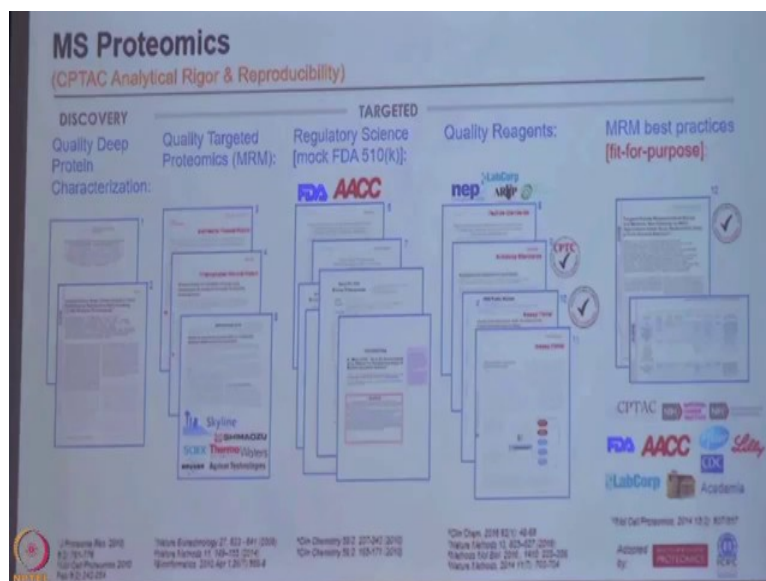
And, if you can, then you would come back to our board, you would give us the confidence that we could trust the measurements; in other words what you are measuring

it is going to be represented the biology trying to go after. It is not going to be attributed to an artifact to tribute it the way the sample is collected or the way you are processing your sale, ok, that is very important. Everybody is going to measure something, but you got to ask yourself is what you are measuring going to be represent of the biology of a disease state. Or, is it an artifact because if it is an artifact it would not go towards patient care most likely.

And, then if you could do that you could go after biology. So, for the very first 5 years we had to try to standardize as much as we can. I am not going to go through all science, but here is what we ended up doing. We basically carved the space of proteomics exactly like you do in genomics and genomics, you first do a comprehensive characterization, once you identify what you want then you basically develop targeted panels. So, targeted panels that today were exactly drives a lot of our patient care especially within the clinical trial space. So, when I came to proteomics, we decided take a very similar based approach. If you do a deep dive that is basically a lot of people refer to a shotgun, I am not a fan of that terminology.

So, basically I tend to call it very deep comprehensive coverage, you are trying to measure as many things as you can.

(Refer Slide Time: 12:01)



And, there we basically showed if it distributed a standard operating procedure amongst multiple laboratories; guess what? You get very good components type CVs typically

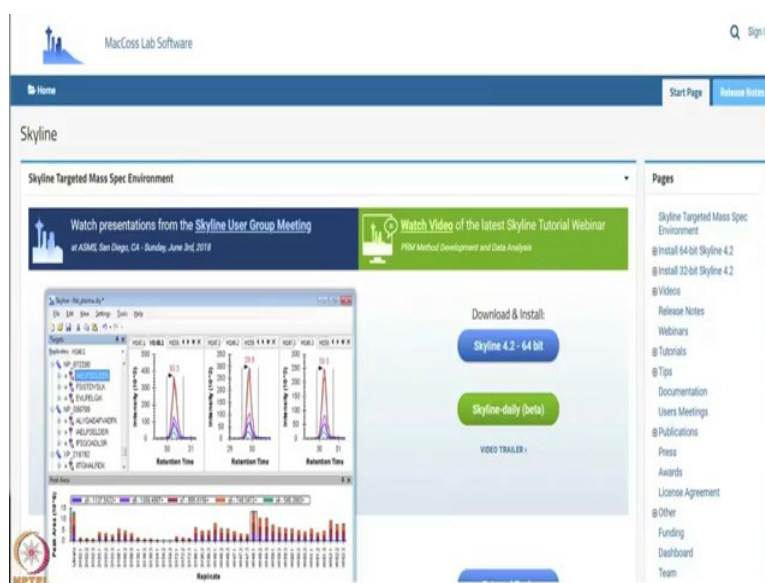
less than 15 percent and sometimes even less than 10 percent which is very good. Then we wanted to explore the space of targeted mass spectrometry because, once you identify the large landscape, you do not want to do this comprehensive based approach all the time.

It is very hot cost, it is a little throughput and it requires a lot of sample. So, you want to get something that is going to be very locked down and for lockdown, you typically targeted based assay.

So, in that space we basically at the time looked at what is now referred to as multiple reaction monitoring and you have different ways of phrasing this. Never invented it, this existed in clinical labs for 30 years; they basically use it for measurement of small molecules. But, basically when I asked a question, if you roll it up to a peptide can you use it in that space and is it reproducible and more importantly can you transfer the technology across laboratories and get very good tight measurements. So, we ended up doing we basically looked at multiple reaction monitoring, we did a series of Round Robin studies.

One involved 8 laboratories in the US, we got very good results; another one that we did an international study labs on the east coast, west coast of the US and we had a lab in Asia; again very good results that we obtained from that.

(Refer Slide Time: 13:16)



People was talking about Skyline a couple of days ago. So, skyline is actually a little product that came from one of our laboratories, when we were actually creating this program. It is a great little tool and it shows you how from basic science, you could get computational product, that is now is being used broadly by the research community. Do you think we started to ask is what if you could take your technology and you could potentially move it a little bit further towards regulatory approval? Because, ultimately that is the goal you want to put it in a clinical laboratory and hopefully use the information to go back towards patient care, in US that is the Food and Drug Administration.

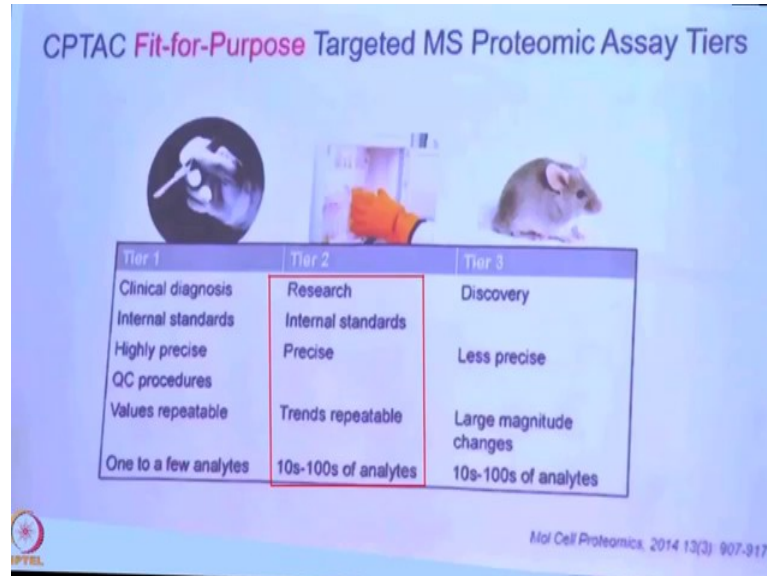
So, typically they get a device cleared as an IVD MIA, you need to go through the FDA and there is two stages behind that. A very first one is what we referred to as a 5-10 k, document what happened in the past was typically a manufacturer will submit it. It gets all marked up by the Food and Drug Administration and then to give it back to the submitter, but the submitter never wants to release it. We were very interested in releasing all information to the public. So, we ended up doing was we held a workshop with the regulatory agency and we basically made up all the data, but we did not make up the analytical workflows.

The beauty that was it allowed us then to submit all our data to the regulatory agency as an official filing. They marked it up like they would for any device manufacturer, but because we submitted it, we made up the data. Then were able to take the document and we published in the public domain. We actually got to published in a clinical chemistry because, we partnered with the American Association for Clinical Chemistry in the United States. The other stuff we realized early on a lot of the commercial grade reagents that are out there in the community were not to the standards, we felt they should have been.

So, we have worked with the commercial sector trying to raise the sort of quality of the products that they release. And, then the other one was a lot of people talk about I have developed a targeted based assay. I will be honest after a while a lot of us did not know what that even meant, because people develop assays and you find out what they mean is that they have either developed a theoretical assay or they develop the assay running it in buffer. The last I checked if you draw blood or a tissue from a patient, it is not theoretical

and there is no buffer flowing in that system. So, we wanted to develop a clinical based way of thinking about it; so, basically it is a fit for purpose based criteria.

(Refer Slide Time: 15:33)



The slide is titled "CPTAC Fit-for-Purpose Targeted MS Proteomic Assay Tiers". It features three icons at the top: a hand holding a pipette, a laboratory flask with orange liquid, and a mouse. Below the icons is a table with three columns labeled Tier 1, Tier 2, and Tier 3. The Tier 2 column is highlighted with a red border. At the bottom left is the CPTAC logo, and at the bottom right is the citation "Mol Cell Proteomics, 2014 13(3): 907-917".

Tier 1	Tier 2	Tier 3
Clinical diagnosis	Research	Discovery
Internal standards	Internal standards	
Highly precise	Precise	Less precise
QC procedures		
Values repeatable	Trends repeatable	Large magnitude changes
One to a few analytes	10s-100s of analytes	10s-100s of analytes

And we actually did that and what is quite nice about it in a very simplistic manner, you could kind of see it as the following. We developed tiers: tier 1, tier 2, tier 3; tier 1 is basically a clinical grade assay, we do not do that within our program. Tier 3 there is less analytical rigor involved in that when you have to submit these sorts of a criteria's. But, tier 2 is a nice little sweet spot that everything within the CPTAC program, we actually adhere which quite nice is that this ultimately not got picked up by the molecular cellular and proteomics as a journal and also by the international community.

So, anytime you know submit to this journal and you say that you have developed a targeted based assay, you to you will have to adhere and describe your assay based on one of these analytical tiers. So, with this in now with this is a 5 year window, at this point we go back to the board of NCI. We could we actually demonstrated that we were able to get very good analytical understanding of these technologies, predominately mass spectrometry. And, now we get approved to move it to the next stage and in next stage was interesting, we wanted to explore as a pilot to go after biology.

The biology we wanted to go after was specifically the cancer genome atlas because, that started biology 5 years before us, we are 5 years behind. And, the way we basically phrased it to the board was we want the exact tumor that just went from a patient to the

cancer genome atlas, and was comprehensively characterized and we will take it. And, then were going to put a comprehensive proteomic characterization right above it. And, ultimately what we are trying to find out is are you able to identify additional biology that is either difficult to obtain or simply not feasible through genomics.

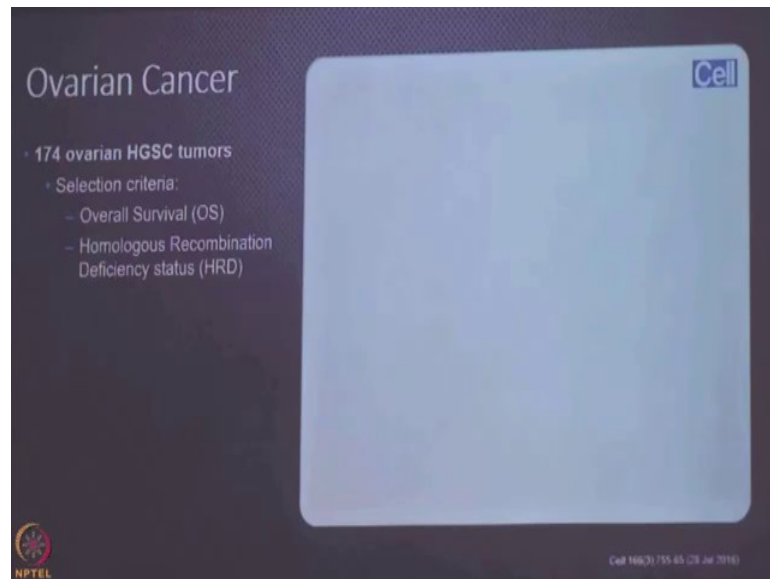
So, think about it because, if what you come out of that sort of a finding is I could confirm what my genomic colleagues just found, it is going be very difficult to convince people proteomic has a role. Because, proteomics cost more and it is lower throughput and does require a higher amount of sample input. So, that was the goal, can you find additional biology pure and simple. So, here is kind of what we ended up doing, we went after three cancer types of TCGA.

(Refer Slide Time: 17:47)

The slide is titled "Proteogenomic complexity of tumors (CPTAC pilot phase)". It features three overlapping article covers from the journal Nature, each representing a different cancer type: Colorectal Cancer, Breast Cancer, and Ovarian Cancer. Each cover includes the title of the article, the journal logo, and a note indicating that more than 100 patients were included in the study. To the right of the article covers is a section titled "Overall Highlights" which lists three key findings: "New biology identified", "New molecular subtypes identified, with outcome clinical associations", and "Possible new targetable antigens". At the bottom right of the slide, there are three references: "Zhang B. Nature 513, 382-387 (Sep 2014)", "Martini P et al. Nature 534, 55-62 (Jun 2016)", and "Zhang H et al. Cell 168(2):755-65 (Jul 2016)".

We went after colorectal cancer, ovarian cancer and breast cancer; on average about 100 individuals for each one of our studies suffice to say here is the overall highlights, in every one of these we found new biology. Now, here is sort of a little example of what I mean by a finding additional biology.

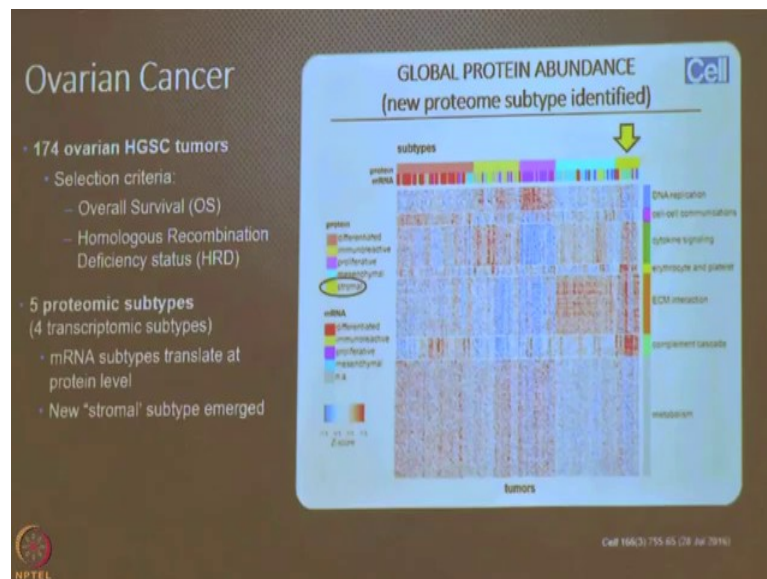
(Refer Slide Time: 18:06)



If you look at the ovarian study, this is just one little slide that comes out of that paper. So, in the cancer genome atlas we actually catalogue just shy of 500 patients to come up with us with the observations that we did for ovarian cancer. And, in that they did a whole series of analytical different ways of looking at the datasets. So, what our investigators had an interest in is, if you look at the proteomics landscape are able to tease out two features, that is associated with ovarian cancer. One is going to be overall survival, typically we wanted to find out if you could separate short versus long term, less than 3 more than 5 years. And, at the same time, they were interested in homologous recombination deficiency or brokenness as it is commonly referred to as.

So, what they ended up doing was the following, out of the 500 we took approximately shy of 200 of samples and we distributed it to two laboratories. They were blinded to what the samples were and they performed a whole series of bioinformatics on the information. One of the things they did was a consensus clustering, kind of analogous to what is done at the RNASeq level. And, the question is if you look at the information at the protein landscape; when you looking at protein abundance what do you get is going to be different or we simply confirm what you did at the transcriptomic level.

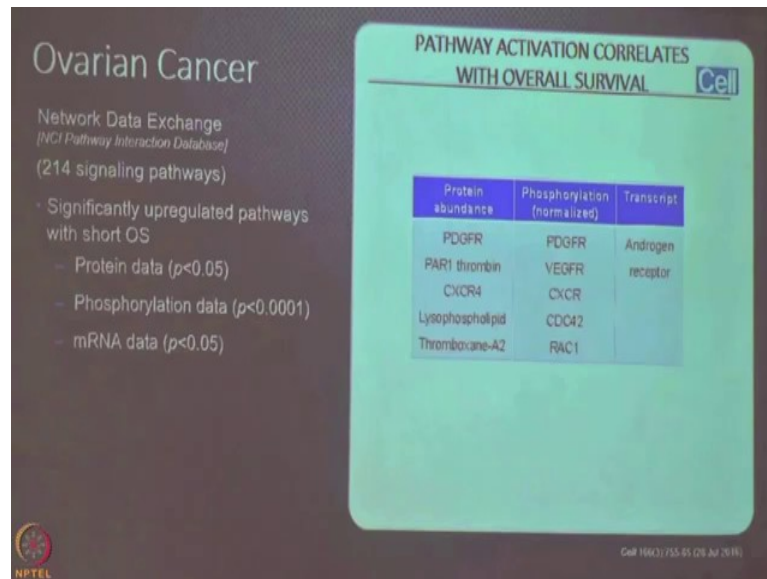
(Refer Slide Time: 19:23)



So, here is what they found out, not only do you confirm, but you are also to infer additional biology. So, out of the four initial subtypes that you get at the transcriptomic level, they nicely roll up to the protein level. But, in addition to that they identified a dip a additional subtype that is identified here. This one they simply refer to a strong role because, a lot of these proteins tend to be associated with things like angiogenesis. But, again the key of this study that they had an interest in when they got these samples, they wanted to identify can the protein information and abundance level separate out for me either overall survival or HRD status; and it actually it turned out the answer was no.

So, protein abundance in itself, in this type of an analysis could not separate overall survival or HRD status, but that actually was not bad and here is why. Because, the same type of analysis was performed by TCGA either in their flagship study or an additional study down the road that TCGA did and they also cannot identified those two criterias. Now, here is why it gets interesting so, these investigators they had an advantage. They had genomics based data from TCGA and we had protein information, they also had modified proteins at the same time. And, it is supposed to be asking these questions and trying to analyze the information from a gene base level way of thinking, they wanted to roll up the information into biological pathways.

(Refer Slide Time: 20:49)

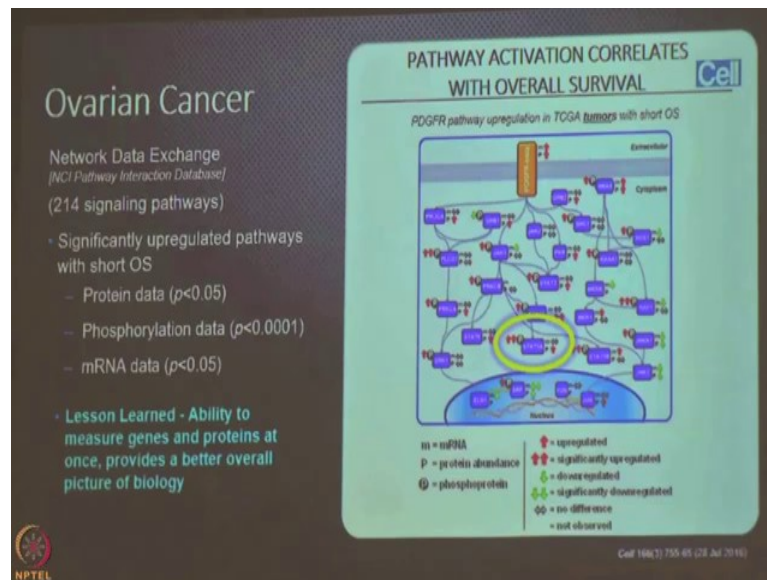


So, what they ended up doing was they took all the data and then they plugged it into the inside pathway interaction database and they identified just over 200 signalling pathways. Now, just focusing on the feature of overall survival, they asked a simple question. Can I use the information now looking at cellular pathways and try to separate short versus long term survival? So, here is what they get; looking at protein abundance, while it turns out 5 pathways all send rise up to the top from those just over 200.

If you normalize against abundance and I look at phosphorylation an additional of 5 pathways became apparent. There is a nice crosstalk PDGFR, one of these growth factor receptor pathways, but because we also had TCGA data from the same tumor; we also analyzed it at the RNA seq level, a different pathway came up. Now, you could begin to see what started happening to our program, in other words if you were to perform an experiment either looking at only protein abundance and you are done. Or, you want to look at phosphorylation and nothing else or you just want to look at genomics, most likely you are going to be looking at an incomplete picture of the underlying biology for this study that we were about that that we were involved in.

So, that became sort of a very turning point for us. So, at the end of the day what we learned from this was, if you have the opportunity as these technologies are now mature and you can begin to actually perform comprehensive genomics with proteomics at the same time; most likely blending these worlds together is going to give you a better understanding. Not only of the underlying biology of the disease, but we hope we hope that the biology could potentially translate towards patient care.

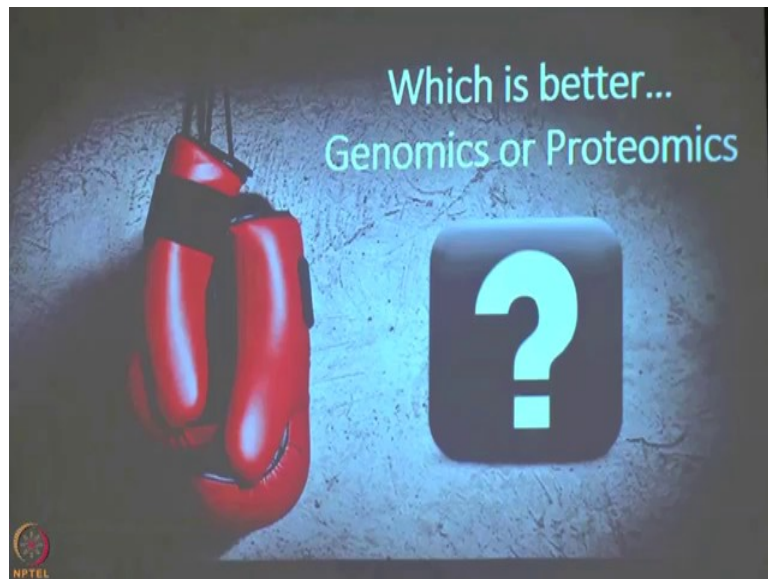
(Refer Slide Time: 22:33)



Now, in addition also to be developing very detailed pathway based maps as they are shown here, you can also begin to tease out those funky features that people like to look at. So, for example, this looks at the what is a further growth factor receptor pathway. A very commonly looked at transcription factor turns out to be STAT5 a. And, actually by rolling both information together, if it is not that obvious here basically what we showed was at the transcriptomic level looking at overall survival not much change.

At the protein abundance level, it kind of resembled at that point what you found at the transcriptomic may be a slight little bump and nothing really this could it be significant, but really saw huge increase at the level of a phosphorylation. So, again three cancer types that we did all similar observations. So, here is not what starts to happen and I have seen this question being asked also in the past couple of days. So now, we have standardization first 5 years, the next 5 years which we which we just wrapped up focus on trying to tease out biology. And, we had to go back to our board and when people kept on asking, it is the same thing that people were asking for the past 2 days and that is the following, wow.

(Refer Slide Time: 23:39)



So, which ones going to be better, should I only do genomics or should I only do proteomics? Should I do proteomics a completely alone genomics which ones better between the two? So, the way that I kind of viewed it was just take yourself back to a book of biochemistry. The first thing you learn is that everything has to relate to one another and if you could get I a good comprehensive systems perspective view of the biology; hopefully it is going to be more representative of the disease state itself.

So, for us the answer became no, I seriously doubt if you do not understand any of biology what you are going after, why do you want to go after one of these omics, now when the technologies have become quite mature. And, here is why which is what is the same argument that I made to the board about 4 years ago.

(Refer Slide Time: 24:23)



If you look at the cancer genome atlas, again the cancer genome atlas I am a huge fan of this program; simply for what it was able to achieve in a 10 year window. They went after 34 cancer types, just over 14,000 individuals and in the process they found a lot of interesting biology. Again you cannot put clinical context behind this because, the samples are never collected with a clinical question in mind, but nevertheless a very good resource that is been given to the public at large.

(Refer Slide Time: 24:50)

The slide has a dark background with a DNA double helix. The title is "Genomics aims to advance Precision Medicine through finding and targeting genetic alterations". The bullet points are:

- TCGA has analyzed samples from 14,000 individuals (34 tumor types)
- Identified actionable mutations, therapies
- Many tumors with actionable mutations do not respond to targeted therapy
- Many responses are temporary

On the right side of the slide, there is a screenshot of a genomic data visualization, likely a heatmap or a plot showing gene expression or mutation levels across different samples. The NPTEL logo is in the bottom left corner.

But, in that they also identified a whole series of actionable mutations, then now some of our small molecules it is actually driving a lot of our precision oncology trials, sites the good news. Now, you can actually look at the other side of your story which is what we

are learning now, 4 years down the road and running a lot of these very precision oncology trials. What we are learning is that a lot of these tumors that they had these actionable mutations, that we develop all these GMP facilities to develop these small molecules. Those individuals actually are really not responding long term to the therapy that they are being administered. If they do respond in short term and a lot of action to develop toxicity, they get a ticket from one treatment arm and then quickly move them into another. Why? We have no idea why that is the bottom line.

(Refer Slide Time: 25:38)

The slide features a dark background with a glowing DNA double helix. The main title is 'Genomics aims to advance Precision Medicine through finding and targeting genetic alterations'. Below the title is a bulleted list: 'TCGA has analyzed samples from ~14 000 individuals (34 tumor types)', 'Identified actionable mutations, therapies', 'Many tumors with actionable mutations do not respond to targeted therapy', and 'Many responses are temporary'. A blue rounded rectangle with the text 'Missing Biology' in yellow is connected to the list by a thin line. The NPTEL logo is in the bottom left corner.

Genomics aims to advance Precision Medicine through finding and targeting genetic alterations

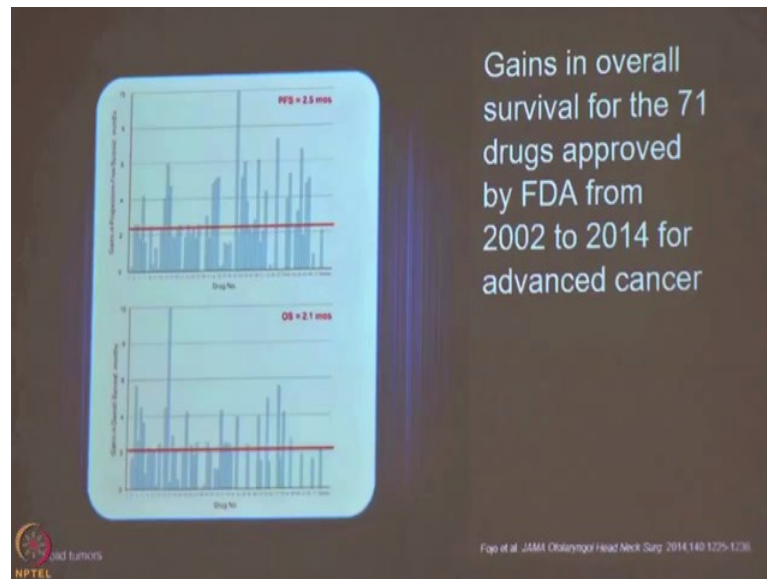
- TCGA has analyzed samples from ~14 000 individuals (34 tumor types)
- Identified actionable mutations, therapies
- Many tumors with actionable mutations do not respond to targeted therapy
- Many responses are temporary

Missing Biology

NPTEL

So, for me what that tells me is that there is still a tremendous amount of missing biology strictly focusing on a one omic based approach. Now, you can actually flip the coin just look at what is going on within a therapeutic perspective.

(Refer Slide Time: 25:48)



So, this is a nice little paper that people could look it up, it is by a colleague named Tito Fojo who used to be at the NCI and now moved to New York City but he did basically did this little analysis where he looked at solid tumors and what Tito did was actually quite savvy. He went in the public domain, he said look if you look at the first main precision oncology drug that came out which is Gleevec along with Herceptin in the early 2000s and what transpires, over the past 15 years.

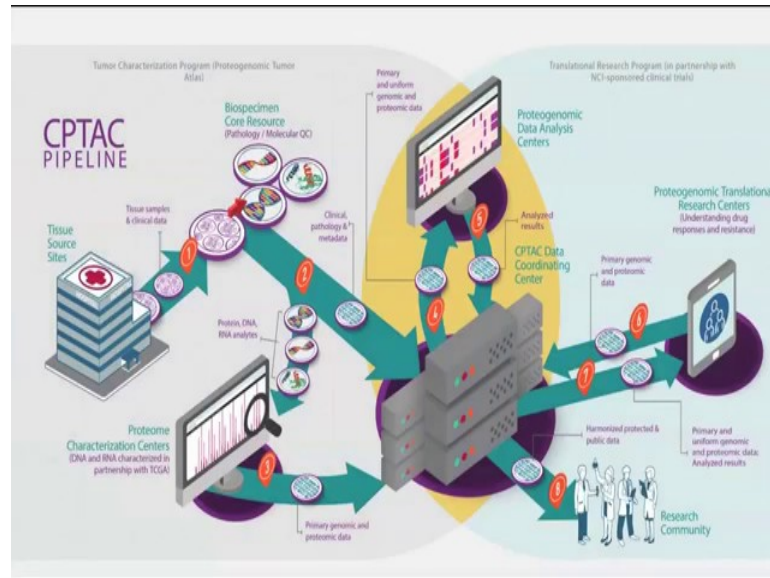
There is about over 70 of these drugs now and if you look at the drugs for all the different cancer types, that the that they are being used either as a single or in combination on average just on the average what is the two main criterias that people look at either overall survival or progression free survival.

Now, you exclude it from this study; obviously, the exceptional responders and what you found out is for all these therapies on average for both of these two different metrics it is typically no more than 3 months. So, this played a big role the way that CPTAC now evolved in its current round. We still go after biology like we did when in the prior program but now are slowly trying to move into that translational space.

So, this is CPTAC today, so, CPTAC is still held responsible to characterize deep comprehensive genomic characterization along with proteomic characterization for five additional cancer types and, all the information we put into the public domain because we see it as pretty competitive.

At the same time for the very first time the National Cancer Institute has now partnered a proteomics laboratory with an on-going precision oncology typically genomically driven NCI sponsored clinical trial.

(Refer Slide Time: 27:28)



Now, what is interesting there is that the information is not going to go back to tumor board to figure out exactly what treatment arm or what therapy to administer to a patient. On the other hand, the information is basically going to be used in a reverse engineering manner. So, based on the study itself you will be able to get samples from these trials which are very well controlled to the amount of clinical inference you are able to pull out is tremendous and you will get pre-door and post.

(Refer Slide Time: 27:58)

Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Proteogenomics Research)

Builds on TCGA

- **Tumor Characterization Program**
– characterize proteins and genes to better understand the molecular basis of cancer
- **Translational Research Program**
– understand [predict] drug response and resistance to therapies in context of a clinical trial (NCI-sponsored)

proteogenomics builds on genomics

Public Resources

And, hopefully what we hope to learn from that program is if the individual did not respond to the way we think they should have responded based on the genomic information, can we identify the biology to the root cause of that by looking at the protein landscape of those subjects and if that turns out to be very revealing my goal is that in the next iteration we want to combine those two worlds fuse them together and actually go directly now toward tumor boards.

(Refer Slide Time: 28:25)

Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Proteogenomics Research)

Builds on TCGA

- **Tumor Characterization Program**
2 years YOUNG
- **Translational Research Program**
1 year YOUNG

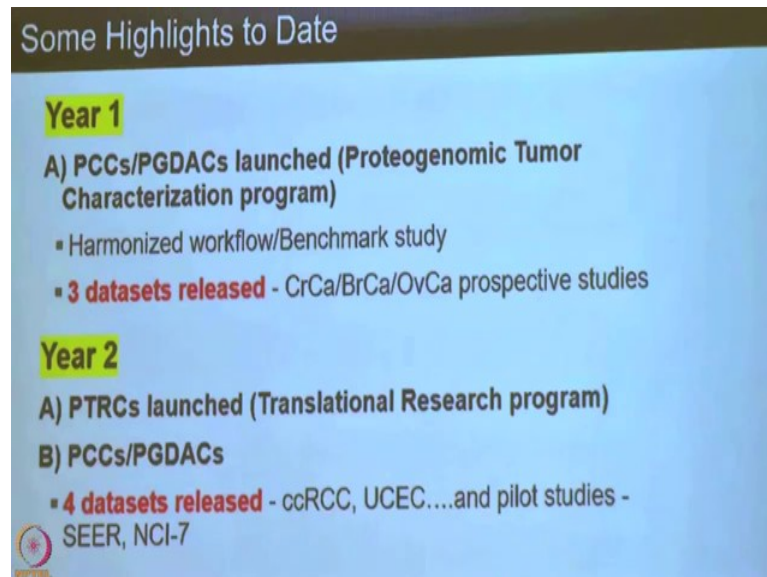
proteogenomics builds on genomics

Public Resources

Now, in terms of how old is the current CPTAC program actually it is not that old. So, in terms of the comprehensive characterization that is now 2 years old, as a program or young as I like to say because, I have reached my middle age crisis. So, I do not like to

use the word old anymore and the one of the translational now is 1 year young. So, what have we done over the past 2 years because that is really a 2 year window that the program has been around.

(Refer Slide Time: 28:46)



So, here is the way we ended up doing, these programs are very complex. The reality is you just cannot get something off the ground and expect it to work, you have to build your infrastructure. So, the first one that we launched was that was the characterization component. The first thing we realized is that we had three main of what we call data productions facilities for sites. Now, we tried then to try to standardize the best we can or harmonize the analytical workflows of the way that they would be producing those data sets.

So, that became very important for us and that pretty much took about a 12 month cycle for us. At the same time they also released an additional 3 data sets to the public which is sort of a continuance of the last program but these are now freshly collected samples that have been optimized for both comprehensive genomic characterization and comprehensive proteomic characterization and that of the way for colorectal cancer, breast cancer and ovarian cancer.

In the second year of our program, we officially launched our translational arms those are partnerships with our clinical trials and at the same time we also then continued in terms of that Brute force characterization arm and we released in additional 4 data sets to

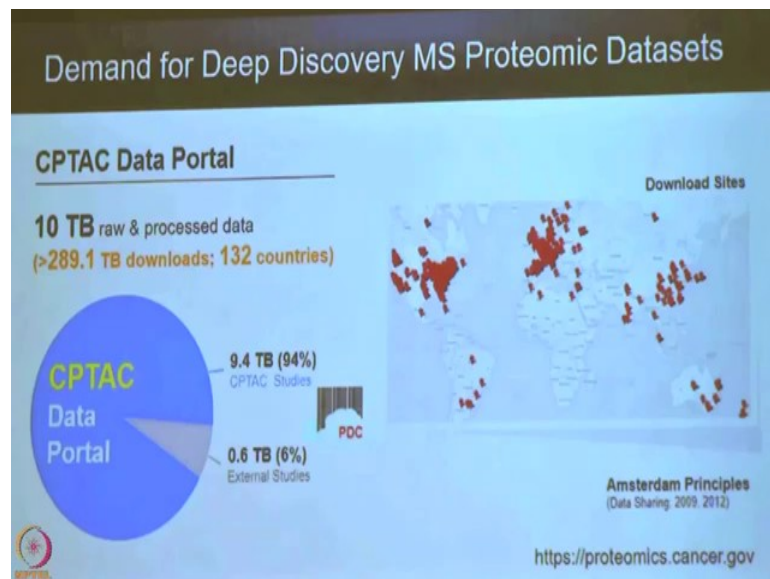
the public. One while two of them actually well in fact, all of them occurred in the fall of this calendar year; colorectal cancer we released and endometrial cancer we released and we also released two additional pilot studies: one focusing on 30 year old samples just trying to understand the stability of these bank materials and the other one was sort of a cell line study and we hope to release another one in the next several weeks.

Now, I talk about a lot, we give all this information to the public. The other question I get all the time is; ok so, you give all this stuff to the public, is it being used? It is like developing a business right, if you guys develop a business and if nobody comes to your store and actually uses your products, your store typically would not stay in business too long. So, I am always paranoid you know are people going to use these materials, I would argue giving away your data and everything you find in a pre-competitive manner is truly advantageous. Not for your own program but at the same time for the globe as a whole and for three basic reasons.

One, if you give away those the raw material, just datasets, reagents your standard operating procedures, it stimulates outside individuals that do not have wet laboratories that are computational scientists, that could reanalyse your datasets and hopefully develop new hypotheses to pursue science in a way that you could not figure that out, a couple of years in the past.

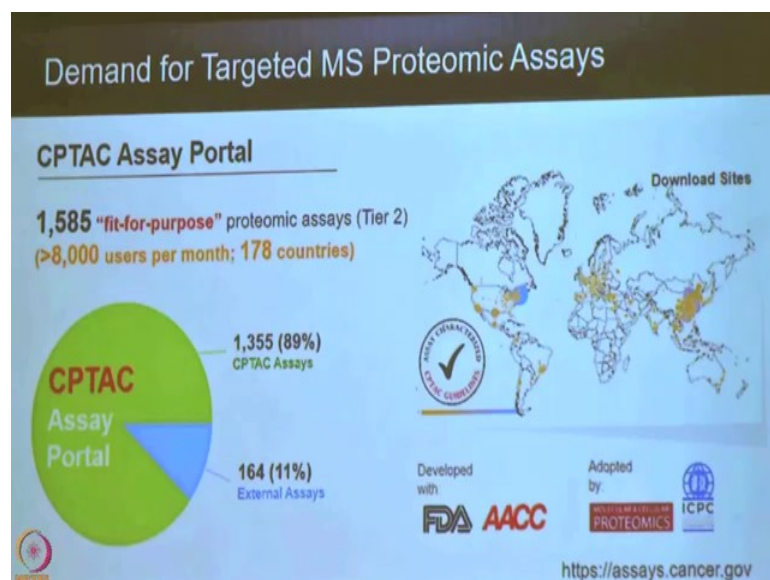
Secondly, if you could take your raw ingredients, work with industry to develop kits, you could further disseminate that to the public and thirdly you hope that some of these kits or the reagents you could put them in a way together that actually could be used in a clinical setting. So, let me give you an example of all three of them.

(Refer Slide Time: 31:34)



So, in terms of our data do people use the data sets of CPTAC, that it turns out it does, it is very simple to get analytic metrics on it. So, our program has about 10 terabytes worth of raw and processed data files available to the public as of today. We know that our data is being downloaded all over the world, specifically just over a 130 countries and actually at the small little 10 terabytes worth of raw files, that those downloads have now exceeded well almost have reached 300 terabytes worth equivalent of our datasets.

(Refer Slide Time: 32:08)

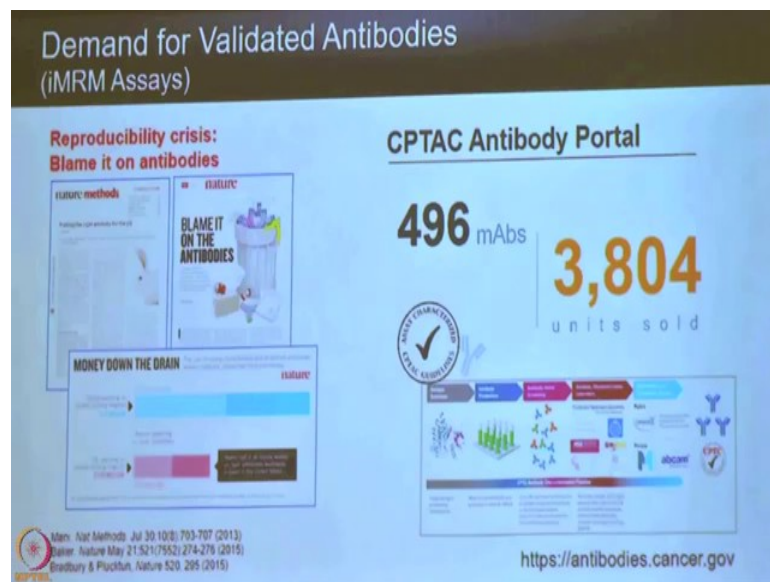


In terms of the other components that I talked about, we also give away a assays for those targeted based assays that we developed. So, we have a portal, we give away all the parameters behind the assays that we develop. We currently have just over 1,500 these fit

for purpose based assays. Do people go to our website? Yes, it turns out on a monthly basis over 8,000 people are now going to our website and they are grabbing whatever information they want, hopefully conducting studies in their own laboratories.

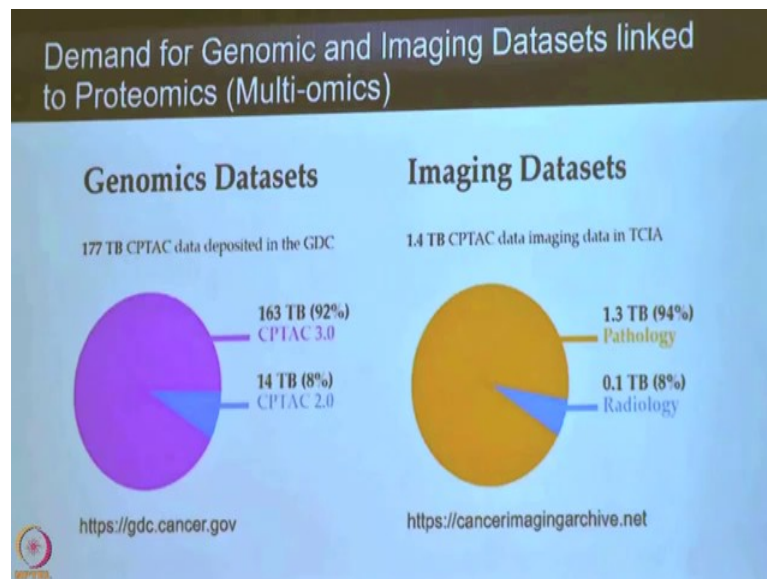
And, actually those download sites come from almost 180 countries, the vast majority assays actually do come from CPTAC but as we developed these analytical criteria for the public were starting to allow outside investigators to deposit their own assays within our own portal. Now, some of these assays do require a higher level of sensitivity, if you want to measure endogenous levels and individuals. So, for that we do develop reagents those are antibodies.

(Refer Slide Time: 32:58)



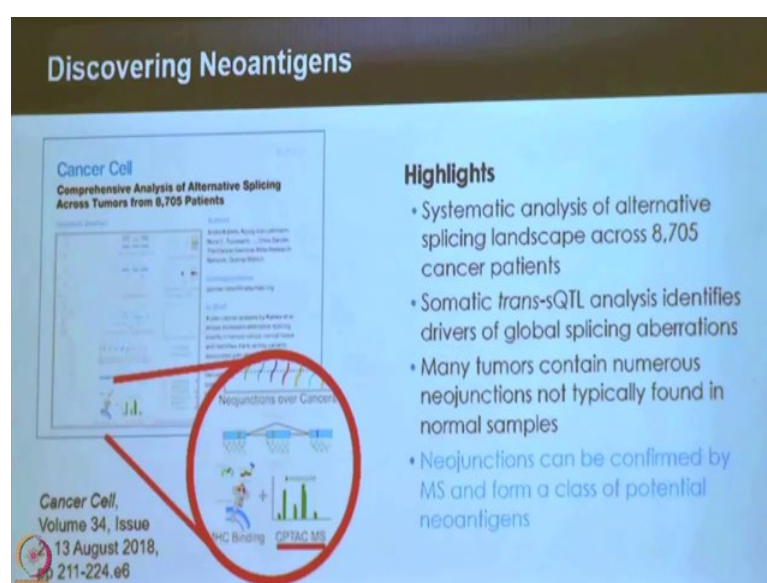
So, in antibodies we have almost 500 monoclonal antibodies that we have developed and fully characterize, we give away all the characterization in the public and we give it away through different distribution arms. One we one distribution arm is a very low cost to the academic model and the other one is through industry and we have been able to sell these units. So, we have now sold almost what 4,000 units of our antibodies which is really good for this little small program out of the National Cancer Institute.

(Refer Slide Time: 33:27)



And of course, we just do not do proteomics isolation, we do genomics and we do imaging. So, all the imaging that comes from the histopathology lab or from the radiology lab we give it away into the public domain. So, it is not just proteomics, we do genomics, transcriptomics, proteomics and imaging everything we put it in the public. Now, another great example is this recent study in fact, I ended up getting this paper from the director of my institute about a couple of weeks ago. He and his comment was have you seen this, it was flattering that somebody else saw it and not me and here is why.

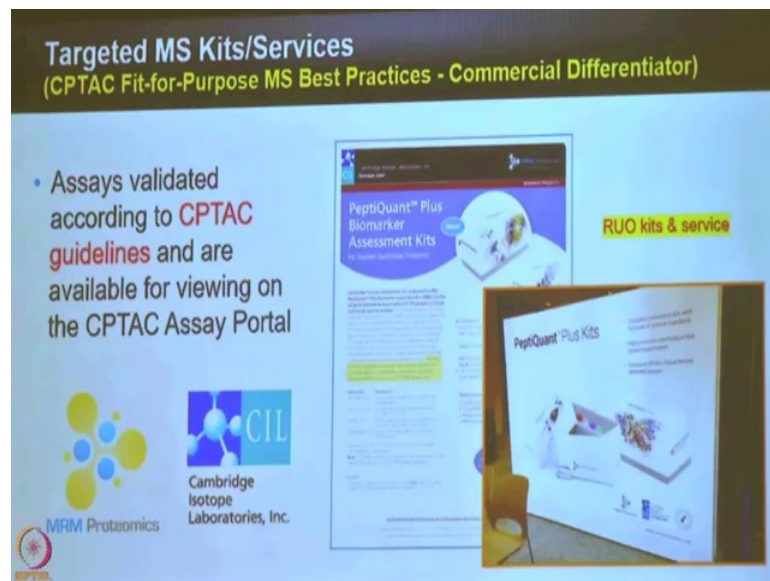
(Refer Slide Time: 34:01)



It is a neat little study. So, our program really was looking at neoantigen, neoantigen is like this hot little terminology that people use now, basically looking at mutated components but at the end of the day this study comes out in cancer cell in a late summer and what they looked at was neoantigens but they looked a publicly accessible data sets it turns out. Obviously the one that typically people will think about is the cancer genome atlas.

But, the part that I liked about it was is on the front cover, if you look at the image that they had the data sets that they pulled out the conduct their analysis actually comes from the US based CPTAC program, why; only because we place it in the public domain. So, this is a great example on how giving up the information, we never explored neoantigens within our datasets another investigator group is able to do it for us and again it further stimulates the science world. Now, these are raw ingredients would even work with industry to develop small kits; so, one of our colleagues is in Canada.

(Refer Slide Time: 34:58)



MRM proteomics so, MRM proteomics develops kits and these are targeted based assays and one of the thing they wanted to know was how did they differentiate themselves from other manufacturers. Our comment was you might want to look at our analytical criterias and we are able to adhere to them, we will host your assays on our own portal, further drives traffic towards your company and at the same time you put out a kit that

has a higher level of standardization than typically what is out there within the research landscape and that is exactly what they did.

So, actually when they now put out kits, that they actually run it as in house health service or the act you can actually purchase their kits and run it within your core facility. These are all researchers only, right there basically tell that they adhere to the CPTAC guidelines for their analytical kits themselves.

Now, this is still the research space, what about the translational space and developing these targeted based assays. Here is a great little study that actually came out a couple of weeks ago and this is a partnership from one of our laboratories on the west coast that actually partnered with AstraZeneca and, here is a great example how proteomics helps the therapeutic side of the landscape.

(Refer Slide Time: 36:05)

Targeted MS in Drug Development
(Novel PD Biomarker to Guide Clinical Trial Advancement)

BJC
British Journal of Cancer

ARTICLE
pRAD50: a novel and clinically applicable pharmacodynamic biomarker of both ATM and ATR inhibition identified using mass spectrometry and immunohistochemistry

- AZD0156 & AZD6738 (in early clinical trials) are ATP competitive inhibitors of ATM and ATR signaling pathways
- IMRM-MS guided the selection of a PD biomarker (pRAD50) of both AZD0156 and AZD6738 to inform Phase 2 dose selection

AstraZeneca
FRED HUTCH

Br J Cancer 2018 Nov 2; doi: 10.1038/s41416-018-0286-4

So, in this one they were looking at is two compounds, these are basically tyrosine kinase inhibitors and attacks two pathways ataxia, that that has a lot of affiliation with DNA damage response but basically this investigator, Amanda Paulovich, she developed a targeted based assays that looks at the DNA damage response, that was a huge advantage for AstraZeneca and in the partnership what they ended up doing was they actually, then identified a marker and this is a pharmacodynamic marker.

That PD marker actually helped AstraZeneca move these two compounds from a phase 1 study, using now this PD marker and they are able to translate it into a phase 2 and it is being used to actually determine the dose that actually is going to be administered to these individuals. Now, this is still the translational space, can you get it in a clinical environment?

(Refer Slide Time: 36:59)

Targeted MS in Clinical Reference Labs
(CPTAC Fir-for-Purpose iMRM to Tg)

iMRM for thyroglobulin:

Clinical LDT

- **Goal:** Quantitate thyroglobulin in human serum in the presence of Tg autoantibodies.
- **Solution (2014):** iMRM circumvents interference of autoantibodies in 20% of the population using conventional ELISA

Logos: LabCorp, MAYO CLINIC, ARFP, Quest Diagnostics

Small text at bottom left: Clin Chem 2009;55(11):1796-804; J Clin Endocrinol Metab 2013;95(4):1343-52

We have actually played in that space, here is one example. Now, here is one where a lot of people try to find new biomarkers but again that is very complicated because you are trying to figure out new biology and believe me new biology towards patient care takes many years but that is ok because biology is complicated. So, we decided to do was to take the analytical techniques, that we have developed and ask clinical laboratories are there existing tests that are problematic, that might be alleviated if you were to bring this orthogonal measurement into your portfolio.

And, in this case they went after thyroglobulin; the reason they went after it is that you find out about 20 percent of the population, individuals with the thyroglobulin they actually suffer from autoantibodies. The autoantibodies, the issue with that is that it is going to interfere with the secondary antibody of analyzer. So, you get a lot of hook effects and basically it you end up with false positives.

So, to circumvent the 20 percent of the population that is missing out on a very good test, we actually our investigators ended up developing a targeted aspect assay dedicated

against thyroglobulin itself. That test today now is being used by every major clinical reference laboratory in the United States.

Now, this is still being used as a laboratory developed tests, what if you wanted to take it to the FDA and get something approved that is a whole regulatory path. Well, that is an interesting space. So, this is what our investigators are now doing within this environment.

(Refer Slide Time: 38:32)



It turns out when you go directly back to the FDA, they will say well mass spectrometry we do not develop the standards or we do not tell people what to do, we look at the community to come up with a consensus document. Once we understood the process, we said so, what is one of the communities you look at. It turns out there is a organization referred to as CLSI which is the Clinical Laboratory Standards Institute aspects and we ended up doing was the following.

In 2016 we worked with the FDA to put on a workshop dedicated toward mass spectrometry, not again not mass spectrometry for metabolites but to move it into the measurement of this, in this case your measurement was going to be a peptide. That ultimately then led in early of 2018 to an existing governing body of CLSI. So, they have always had historically a document referred to as C62-A for using mass spectrometry in a clinical setting for the measurement of metabolites but not for peptides.

We were not working with them to develop one dedicated to the measurement of peptides and the goal is hopefully within the year 2020 it takes a lot of time apparently but in 2020 they are going to be released a document that is dedicated for the measurement of peptides which is a lot which is basically the targeted mass spec, there are lot of people have been referring to.

Now, the other question I get is well are these technologies very specific to cancer? It turns out they are not, technology is ambivalent that is the beauty of it. So, here is a great example of it. So, at the National Institutes of Health, I belong to the National Cancer Institute; one of my sister institutes the National Institute of Diabetes they basically put out a funding solicitation early this calendar year.

(Refer Slide Time: 40:07)

Targeted MS Assays for Type 1 Diabetes Research

FOA released (Feb 2018)

- NIDDK FOA (RFA-DK-17-019)
- Proteomic assay validation to follow **CPTAC Assay Guidelines**
- Assays to be deposited in public domain, such as **CPTAC Assay Portal**

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Department of Health and Human Services
Part 1: Overview Information

Authority/Justification:
Funding Opportunity Announcement:
Funding Opportunity Title:
What Government Agency is the Primary and Responsible Division of Responsibility for this Announcement? Type 1 Diabetes Research (DK-17-019) (RFA) (Public Health Service)
What is the Announcement Number?
What is the Announcement Title?
What is the Announcement Number?
What is the Announcement Title?
What is the Announcement Number?
What is the Announcement Title?
What is the Announcement Number?
What is the Announcement Title?

The reason I loved it was the following when we found it, they talked about is that there this that they are going to be funding laboratories in the US to develop targeted based assays, against I believe type 1 diabetes; yes, but that is not the part that is interesting is taking proteomics and the diabetes, the part that we liked the most was they basically said it. When you develop your target based assays, you have to adhere to the guidelines developed by the National Cancerous to CPTAC program and more importantly you have to deposit the analytical criterias of your assays in the public domain. So, it sets a precedent for other people to be replicating that process.

Now, CPTAC I pretty much do not do anything in the program I have to admit, I have the pleasure of being at the National Cancer Institute and overseen this effort. This is really a team based program and this involves multiple institutions when the United States, just a series of incredibly talented scientists; it is been one of the most privileges I have had over the past 12 years.

(Refer Slide Time: 41:10)



But now this program actually has spawned these other sorts of initiatives of blending these two worlds together.

(Refer Slide Time: 41:19)

Points to Ponder

- Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a national effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis, or proteogenomics.
- Both genomics and proteomics need to be understood thoroughly to understand disease biology. That means all we need to focus is PROTEOGENOMICS.



(Refer Slide Time: 41:31)

Points to Ponder

- CPTAC Data Portal contains huge number of data from 132 country world wide.
- CPTAC is trying to come up with a standardized reproducible workflow which can be followed by laboratories around the globe.



MOOC-NPTEL

IIT Bombay

I hope after listening today lecture, you are convinced that whether to choose genomics or proteomics which one is better. Probably you will not ask this question anymore and you will agree, that both of these technologies are good but probably a good integration of proteogenomics could provide us much more meaningful information. Dr. Rodriguez provided very good example, that if you open a biology book what we find is the correlation which defines the complexity. So, both genomics and proteomics need to be understand thoroughly so that we can understand important questions for disease biology; that means, we all need to focus on the new area which is proteogenomics.

I hope you also heard various pathway, networking correlation in the ovarian cancer project which shows new aspects, new information could be obtained using proteogenomics and phosphoproteomic analysis. He also provided you brief overview of CPTAC data portal which contains large number of data from 130 countries worldwide. Finally, he provided highlights of some of the facts which are related to the targeted proteomics and how CPTAC is coming forward with different guidelines to standardize these assays.

In the next lecture by Dr. Henry Rodriguez, he will talk about other programs and initiatives which are generating and managing multimodal data other than CPTAC. He will also brief about data common framework and cancer research.

Thank you.