

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

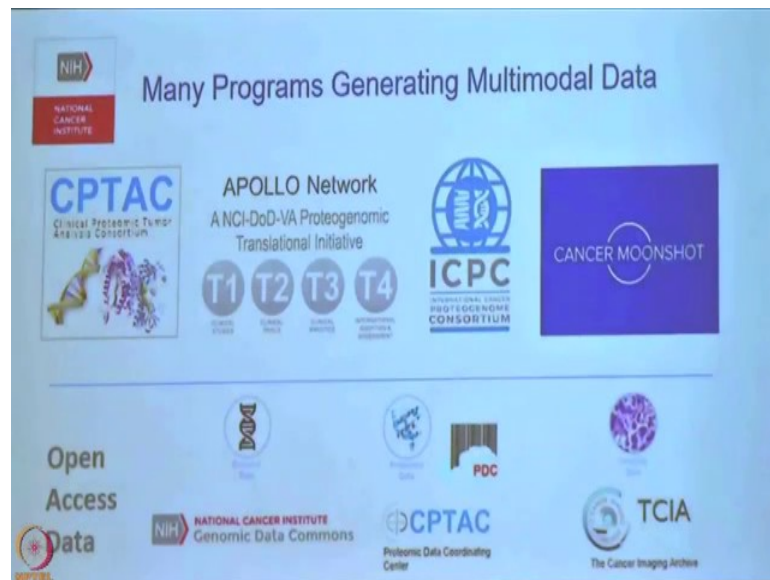
Lecture - 02
Proteogenomics overviews-II

Dr. Henry Rodriguez has provided a very broad and comprehensive overview of need for proteogenomics research. In today's lecture, he will discuss about two more latest initiatives, International Cancer Proteogenome Consortium or ICPC, and APOLLO Network. The Cancer Moonshot program is a very ambitious initiative from US government, and now it is also expanding with various other international countries to build the networks where all the countries could discharge sharing the data which could accelerate cancer research. These initiatives by Dr. Henry Rodriguez are not only accelerating the cancer research, but also helping in worldwide data sharing.

Dr. Rodriguez will talk about APOLLO network which is Applied Proteogenomics Organizational Learning and Outcome. The emerging field of proteogenomics aims to better predict how patient will respond to a given therapy by screening their tumors for both genetic abnormalities and protein information. An approach that has been made possible only in recent years due to advances in proteogenomic analysis

Dr. Henry Rodriguez will demonstrate how Cancer Moonshot can accelerate the cancer research, how it can make more therapies available to more patients, while it will also improve our ability to prevent cancer and detect at an early stage. Lastly, Dr. Rodriguez will talk about NCI genomic data commons. The mission is to provide the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine. So, let us welcome Dr. Rodriguez for today's lecture.

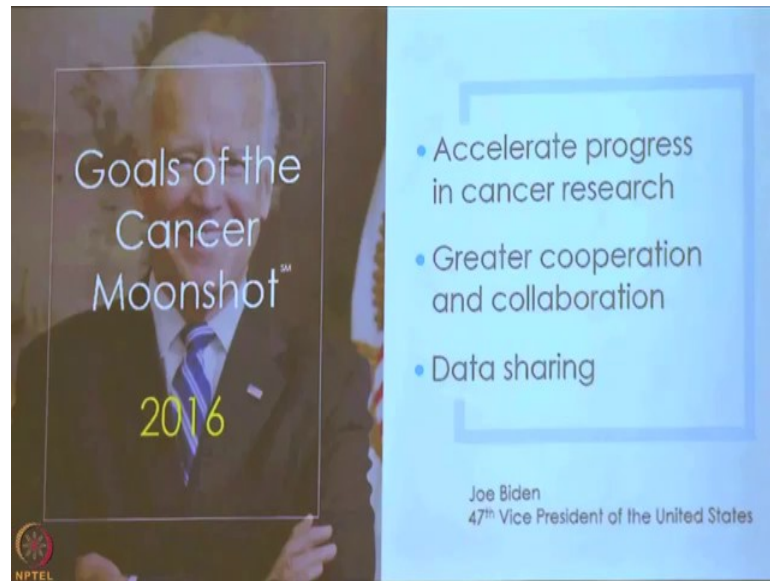
(Refer Slide Time: 02:22)



So, CPTAC today is not the only one that begins to blend the worlds of genomics with proteomics. Two other programs really have come into play, and a lot of its attributed to the Cancer Moonshot. One of them is referred to as APOLLO, and another one is referred to as ICPC or the International Cancer Proteogenome Consortium.

And again the part that is quite nice about all these three is that not to do they blend these worlds together, but everything that we produce genomically, transcriptomically, proteomically, and from an imaging from the pathology suite and from the radiology suite, we place it in the public domain. So, that said how are these other two programs and what is their relevance, APOLLO and ICPC that comes out of the Cancer Moonshot.

(Refer Slide Time: 03:05)



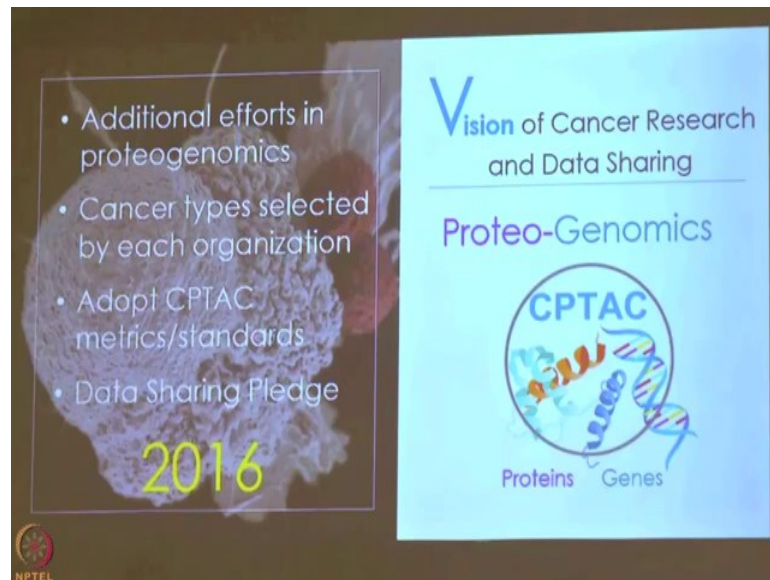
So, the Cancer Moonshot is an interesting program, this is something that was launched in January of 2016 and at the time that was being led, in fact, the inspiration was the former Vice President United States Joe Biden. And, the part that struck me the most was the simplicity of the overarching goals that the Cancer Moonshot wants to achieve. And, that was accelerate progressing cancer, in other words if something would take 10 years, can you condense it down to 5 years. And, there is many ways that you could achieve that to be quite candid.

But the last two are things that the CPTAC program within the US has been doing now for years, and one is greater cooperation and collaboration to be very clear that is not simply collaboration within your own laboratory, with your colleague next door on your lab bench. Or with another person in another lab with a new institution that really was implied on an international scope which is what they were trying to achieve.

And, another one was data sharing. Make all the data that really deemed pre-competitive and place it in the public domain as a way of accelerating progress in cancer research. Now, when this actually came out, then I actually then was asked by the White House Cancer Moonshot taskforce to come up with some ideas behind it. The Cancer Moonshot is much larger than these two programs, but one of the things they asked was individuals, hey, can you come up with ideas that we could see that might be interesting to develop.

So, what started going through my head was basically taking this idea what CPTAC been doing with International Cancer Institute blending proteomics with genomics and begin to expand it. So, one of the things was well can you develop additional efforts, and you know in research groups that are now have an interest to blend these two avenues.

(Refer Slide Time: 04:48)



At the same time, if you look at different organizations, they would be the best to determine what cancers would be most relevant, whether be within their organization in the United States or outside of the United States. Furthermore, because CPTAC has spent a lot of quite frankly time and money developing a lot of these metrics. People will begin to adopt them as appropriate. They are not forced to do it, because I think that is wrong, but if it is appropriate we want to do that.

But the part that was really nice was is that everybody would sign what now affectionately referred to as a data sharing pledge. And the pledge is basically a document and it does say it if you wish to partner with our organization, the information you produce from this research collaboration will be placed in the public domain. We will host it at the NCI, or you could host it anywhere you want, but we want to see it in the public that was a key thing for us. So, the very first one that we decided to do was keep it within the US. So, right across our hospital in Washington DC happens to be the naval hospital and that was the first one that we decided to struck the deal with.

(Refer Slide Time: 05:49)

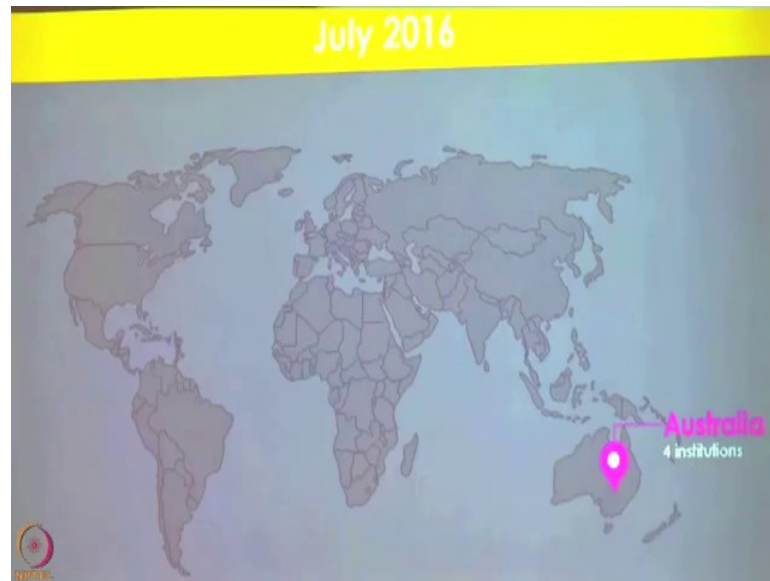
The slide features a light blue background. At the top left is the NIH logo. The main title is 'Three Federal Agencies - APOLLO' in bold blue text, with '(Inspired by the Cancer Moonshot and CPTAC)' in smaller text below it. Underneath is the full name 'Applied Proteogenomics Organizational and Learning (APOLLO) Network'. A central photograph shows a large group of people in professional attire standing in front of a building. Below the photo is the caption 'APOLLO Investigator Retreat | March 9, 2018'. To the right of the photo is a section titled 'Collaborative Partnership among NCI, DoD, and VA' followed by three bullet points: 'From cutting-edge proteogenomics research to wider uptake of standard care', 'Adoption of CPTAC metrics and standards', and 'Data accessibility that is universally usable'. At the bottom left is the URL 'https://apollo.cancer.gov' and the NPTBL logo.

And that program is now is officially referred to as APOLLO. So, APOLLO involves the National Cancer Institute, the Department of Defense and the Veterans administration. And APOLLO basically takes the CPTAC model, and rolling across all the VA hospitals and all the military hospitals within the US. The ultimate goal is to be conducting research that begins to blend the existing genomic based methodologies that is being driving a lot of the patient care with now blending it with the proteomics landscape.

Now, this one when this got completed I thought my job quite frankly was done I could give myself a nice pat on the back, I could go to my wife Haley you would not believe what I just ended up doing, my daughter would be like wow that is amazing what you ended up doing. It turns out it was not that easy, because when this got done then I get another phone call. The phone call is well we love we ended up doing here, but can you bring outside countries now into this mix, I thought it was an interesting call. So, we decided to take the challenge upon us.

So, in summer now of 2016, we decided to take the program on an international level, partly because we also had little collaborations with some institutions across the US. And but we decided to now formalize it across the Cancer Moonshot activity. So, the very first country that signs on to this idea of developing this partnership with the National Cancer Institute becomes Australia.

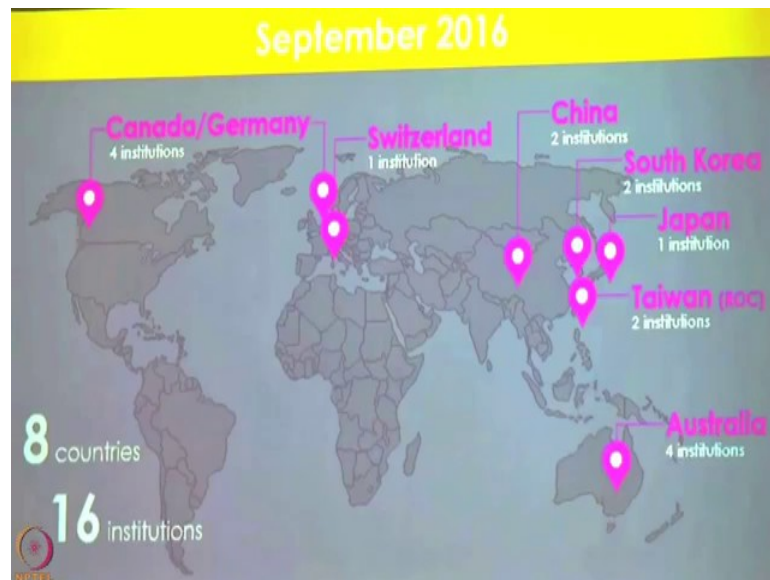
(Refer Slide Time: 07:17)



So, we had Australia on board and we brought in 4 institutions. Now, again I thought my job was done. I satisfied a phone call, nope, it is never that easy it turns out. When you deliver typically people want more. So, rule of thumb you want to under promise and over deliver which is what I have learned, because the many start delivering everybody expects even above that now.

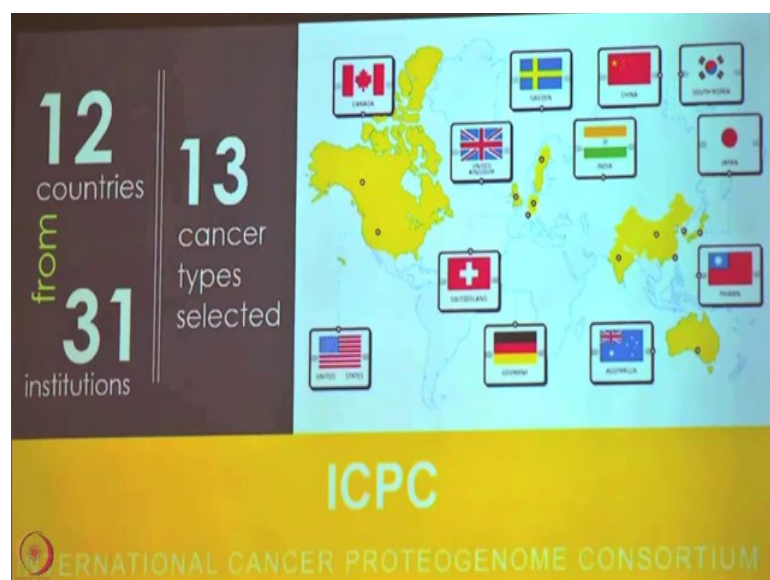
So, I ended up then getting that second phone call, the second phone call becomes ok, we love we ended up doing with Australia. And in fact it was great because the former Vice President Joe Biden flies to Australia and he does this big opening. Another phone call comes down the road and the question is well can you bring other countries? And by the way you have 8 weeks to do this. I had no idea why 8 weeks was relevant I found out later on that it was going to be announced at the United Nations, but nevertheless I decided to take the challenge again.

(Refer Slide Time: 08:09)



So, in July of 2016 we ended up going from 1 country 4 institutions in a span of 8 weeks, we expanded this now to 8 countries and we brought in 16 institutions. Now, this happens to be September of 2016. Obviously, now we are in December of what 2018, so the question is whatever came of this program. So, to my surprise, but to my pleasure I have to admit this that has now taken a life on its own.

(Refer Slide Time: 08:33)



So, this now is officially known as ICPC. This is the International Cancer Proteogenome Consortium. This now involves 12 countries, spanning 31 institutions collectively all

working together on just over a dozen cancer types. Some of these cancer types do overlap, but that is actually fine, because I am the first one that I have wanted to know for years. Why for example, individuals in the United States they are predominately going to be of European descent develop breast cancer women, yet these individuals a lot of its going to be smoke based you find out, but you go to Asia a lot of women really do not smoke, and they are developing breast cancer.

For me the goal is very simple of ICPC ultimately what we want to do is develop a database a resource that now is finally going to be representative of the diversity of individuals along with their cancers across the globe and give all the information back into the public domain. So, what has the program done in the past 12 months? Not, you know since the time this thing was created.

So, here is an example of what we have done. So, the very first dataset because at the time people said, oh, you will never get other countries to make the data public, no one is ever going to do that. I have no idea what people say this because if you sign a paper and that is part of what you sign onto, it is like a marriage contract to me. Yes, you said I will I do, and you expect something to happen, and we are not having difficulty thus far.

(Refer Slide Time: 10:01)



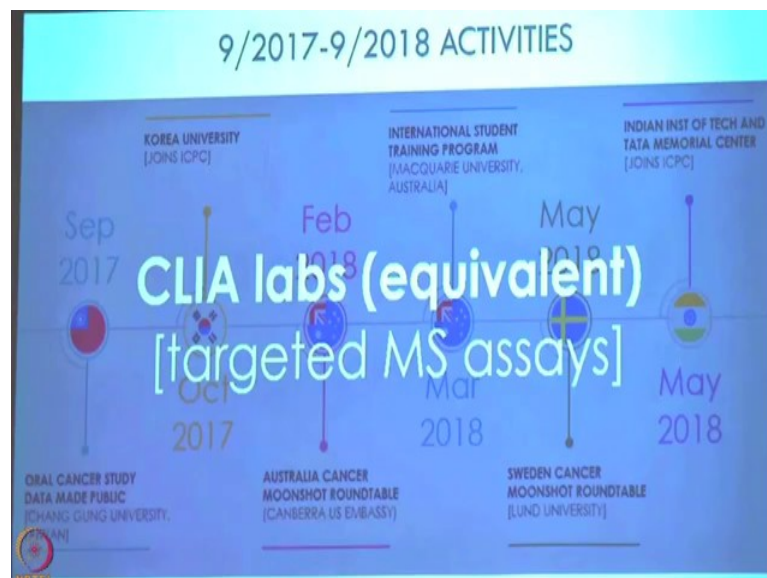
So, the very first data set that got released was in September of 2017 by our colleagues in Taiwan. A very unique study that they ended up doing for oral cancer which is very

dominant there, especially within the rural population because of the betel-nut that they happen to chew along with all the components that they add to that.

Then at the same time what we did last year we actually held local Cancer Moonshot workshops as a way of raising greater awareness within the local municipalities. Same thing that was going to be happening here within the Cancer Moonshot, activities raising the awareness that helps then those individuals, those universities in those countries raise their own capital funding to launch larger initiatives within their own component.

At the same time we actually have, so one was it was actually being held in Australia and the another one was done in Sweden. We also welcomed last calendar year officially to a 3 additional institution spanning 2 countries. The very first one was Korea University which joined us in what October 2017. And of course, India joined in May of 2018. And we also launched an international or we piloted a student exchange program from Australia with one of our laboratories based in the US. The other thing that we are starting to do is because all this is research based pretty much use only is that we are starting to convert some of these laboratories on an international level to become clear certified.

(Refer Slide Time: 11:30)



And these are the targeted based assays. Why do you, why would it be advantageous to be clear certified? Because, that means, you could take the information that comes out of your instrument, and take it directly back to a tumor board to give it back to a patient. So,

we are starting to build the infrastructure more and more on an international skill. This is a hugely fun program, I have to admit we get together now at least once per year.

(Refer Slide Time: 11:47)

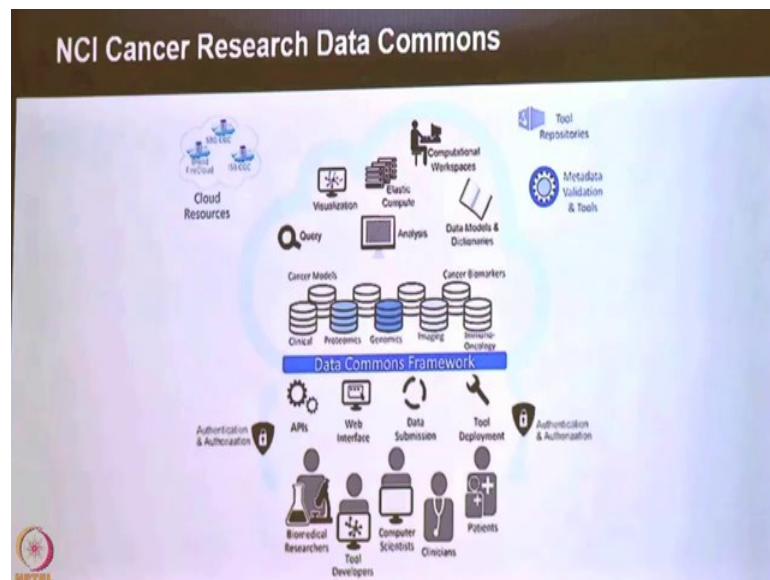


The very last time we got together was in the United States in the State of Florida. As you could see down on the right these are the other times that we have gotten together, but now as a sense of pride, we all get together and at the meetings we all hold our representative country flag, because it is really multiple nations recognizing the cancer is simply not something that is locked to one nation, it is an international effort that we need to resolve.

So, now what we have done within the NCI is we developing data commons. So, in the genomics landscape, we have launched just last calendar year, what is now known as the genomic data commons, this is predominantly the datasets that comes out of the cancer genome atlas. The part that is quite nice is that the ultimate goal is everything is based in the cloud.

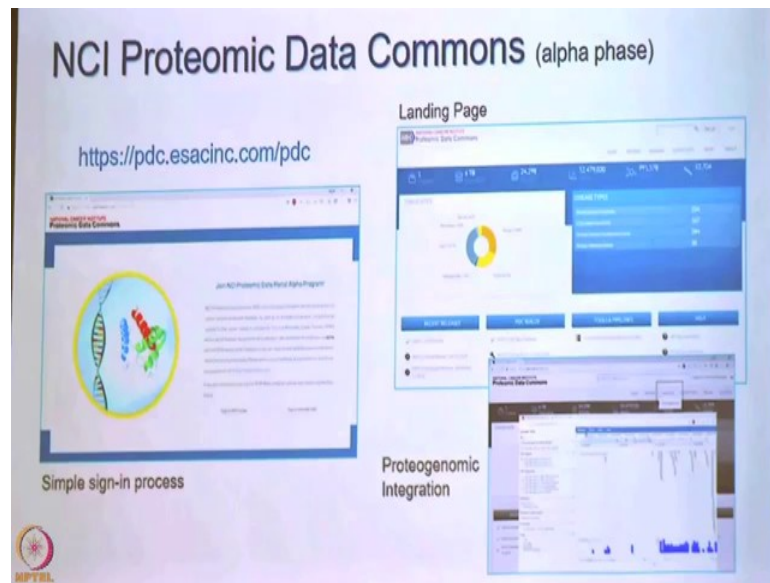
You do not even to download the data packets anymore. And a lot of the software tools are all docker eyes with in the cloud itself, so that is the genomics landscape. So, the question becomes well NCI what else are you doing, this is more than genomics. And I remember Rodriguez you just said that you love the proteomics landscape mixing it with the genomics.

(Refer Slide Time: 13:36)



So, here is what NCI is now doing. The ultimate goal now for the institute is know I am going to have this genomic data commons is basically to have a cancer research data commons, and that is going to involve multiple modalities of the different types of omics. Obviously, the one that I want to point out which is why we are here is we are going to be building a proteomic data commons. And you will be hearing lectures tomorrow exactly how the proteomics information is slowly being rolled into this landscape.

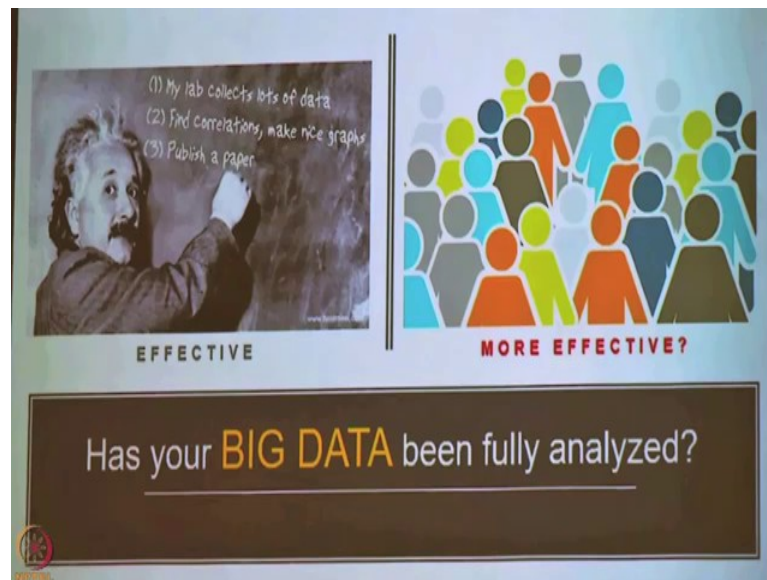
(Refer Slide Time: 13:55)



But if people want to play with it today, they did have a soft launch this calendar year. Here is the website. Please go through it. Look at it; play with it; critique it; provide your comments, because only the things that we are trying to do this is basically as you have your restaurant, it is a soft launch.

So, this is like our alpha launch equivalent, but it is out there now and ultimate a goal is all the data sets is predominantly populated with the CPTAC data sets, but it all directly links back with TCGA. And obviously, the goal for us in the future is to have all everything based in the cloud along with the computational capability that we want to put into this. But ultimately the goal is simply to have a cancer research data commons.

(Refer Slide Time: 14:37)



Now, lastly is this one. These programs produce a lot of datasets, I have done them well. So, what do you do with the data? Well, our investigators we do give grants and they try to analyze the data the best they can. The question becomes is have you really taken or extracted all the knowledge out of the data that you possibly could. Quite frankly I have to admit my answer for many years was, of course, we have looked at everything you can look at the datasets. So, about 5 years ago and at a meaning an individual named Gustavo Sulvezki, who actually came up with something called drink challenges, and he has an academic appointment in New York, but he predominately works for IBM.

And his comment to me was quite frankly he said you are an idiot, I actually liked him when he phrased because my comment was; so, what do you say this. And he explained to me that he created something called challenges, where he basically takes existing datasets that are out there, and he challenges then the community, just like you guys were doing with these questionnaires, can somebody come up with a better way or a better algorithm to go after the information that either you could not extract from it, or you did extract, but it was not as efficient as time moves on with better tools that are not being developed.

So, ultimately the goal that I came up with is this easy cartoon that there is a cool little website, you could actually take a picture of Einstein, and you can literally type in questions that appear that Einstein is writing. So, for me this is exactly what I do when I

used to be within the university setting is typically what you do when you run a laboratory is you do your experiments, you collect a lot of information, you find very nice correlations, you try to develop these fancy graphs like volcano plots.

I saw being talked about yes very attractive, you know they are quite complicated. I think people just say nod their head, yes, I understand I do not know if they do, but at the end of the day what you want to do is basically publish a paper. It is a very effective model, because what you want is these individual laboratories which are very elaborate like artisans that is a lot of the creativity that is out there and I love that landscape space. But the question becomes is well what do you took that information and you put it on in the public and you begin to crowd source a question, so that is what we wanted to explore.

(Refer Slide Time: 16:51)

DREAM-CPTAC Proteogenomic Computational Challenge

DREAM CHALLENGES
powered by Synapse

Global Crowdsourcing
\$25,000 award (NVIDIA Foundation)
Nature Methods (journal partner) - supports submission of overview paper and insights that emerge from it.
Launched Jun 26, 2017; Closed Nov 20, 2017

504 participants from 20 countries

CPTAC STATE LEVEL

| Subchallenge A | Subchallenge B | Subchallenge C |
|--|--|--|
| Protein-protein interactions from mass spectrometry data | Protein-protein interactions from mass spectrometry data | Protein-protein interactions from mass spectrometry data |

CPTAC partners: Synapse, NVIDIA, IBM, etc.

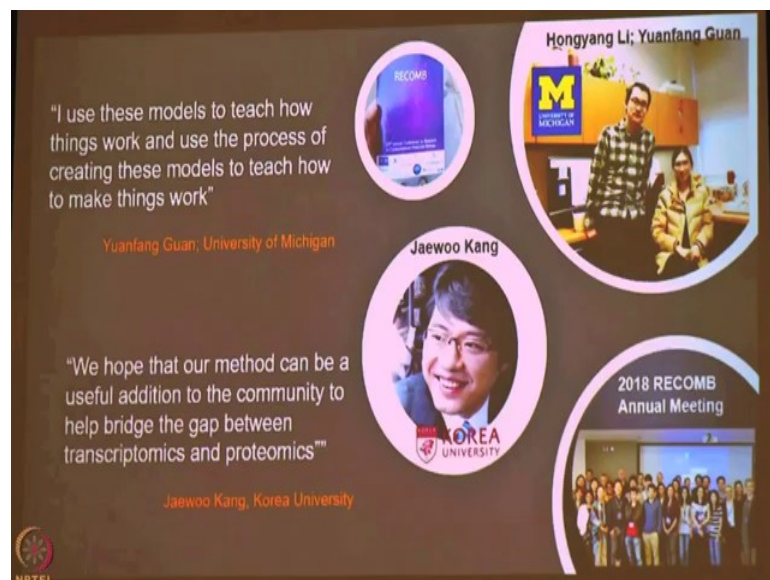
<http://www.synapse.org/ProteogenomicChallenges>

So, the very first one that we ended up doing was about 2 years ago, and it was a very first proteogenomic computational challenge that was crowd source. We teamed up at this point with the dream organization we actually brought in NVIDIA, which is the graphical chip manufacturer. We partnered with them. They gave us some a server frames for what they wanted to do, but ultimately they also contributed a monetary award to this. And we also partner with nature methods, you are not guaranteed to get your findings published, but you are pretty much guaranteed to send it off for review which is quite nice.

But ultimately at the end of the day is we literally thought we would get maybe 50 individuals applying to a challenge to our surprise we got over 500 individuals that applied to the challenge and actually span 20 countries. Now, this challenge predominantly was a biological driven challenge. When we took our data sets we basically asked simplistic questions. So, challenge 1, you can see basically is if we give you DNA and RNA how good are your predictors now determining the abundance. If you give you DNA and RNA in abundance, how could are your predictors at looking at the phosphorylation.

And the way that a lot of us now like the phrase it is the good news is you have winners. The very good news is that it turns out the computational tools are still not as good as a physical based measurement. So, people that are going in proteomics you clearly will have jobs for the years to come which is what I would like to say. So, but here is the part that really struck me the most is that out of the universities or the institutions that we thought would have won, we are not the ones that won at all, in fact, there were groups that we never heard of within my own program.

(Refer Slide Time: 18:33)



So, two of these challenges was won by colleagues at the University of Michigan and another challenge actually got won by a Group at Korea University. Now, this was a biologically driven challenge. This actually caught the attention now the Food and Drug Administration back in the United States, and we decided to do a new challenge with

them. So, now, the FDA has launched a very first regulatory proteogenomic based challenge.

(Refer Slide Time: 18:54)

precisionFDA NCI-CPTAC Multi-Omics Mislabeling Challenge

Global Crowdsourcing

Nature Medicine (journal partner) - supports submission of overview paper and insights that emerge from it.

FDA nature medicine CPTAC CLINICAL PROTEOMIC

precisionFDA NCI-CPTAC Multi-Omics Mislabeling Challenge

Challenge launched September 24, 2018

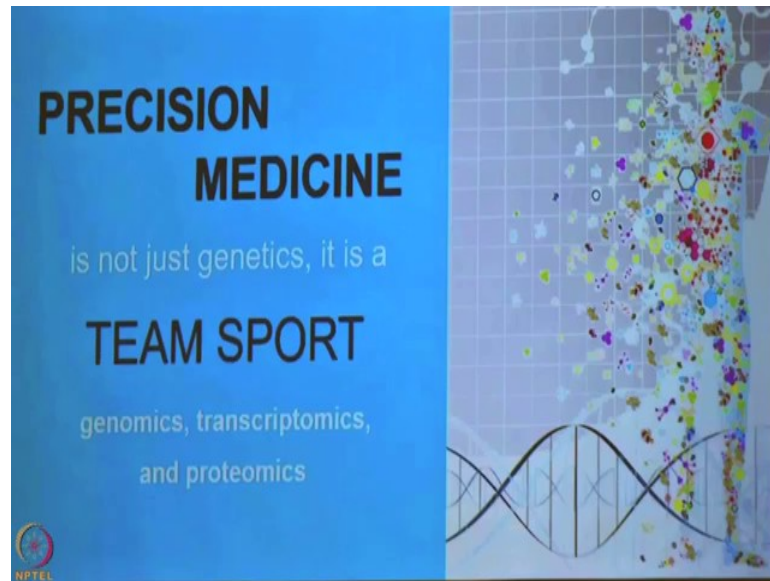
Challenge closes December 18, 2018

Join today: <https://precision.fda.gov/mislabeling>

And here it is, it is still technically ongoing. So, this one now basically again it looks like crowd sourcing, it no longer involves the dream challenge although it is a partnership with them still, but pretty much with the US FDA. And what we are doing here is something that technically can happen. If you take it in an individual sample and ultimately the sample goes to multiple laboratories one for genomics, one for may be metabolomics, or for one for proteomics, in our case genomics, transcriptomics and proteomics, what if one of these samples become misplaced or mislabeled and the information comes back.

So, we asked that question and using other genomics or then taking genomics and third proteomics how easily are you able to identify and most likely put the information back to identify where the mix up occurred. So, this is the very first one with the FDA and it is still open if you quickly want to join it, but hopefully the goal out of this is that it kind of shows is that when you put the information in the public domain, there is all the utilities behind it.

(Refer Slide Time: 19:55)



So, lastly what I want to say is I hope that I have been able to kind of demonstrate how what we have been doing at the NCI. And, now with these moonshot based efforts whether it be in the US or in an international level is that at least for me, I am one of these converts, to me I actually come from a genomics background if people will ever look at my old history. But I am actually convinced is that if people talk about precision medicine or precision oncology for me that is really what I like to define as a team sport, it is not genomics in isolation, it is not either metabolomic in isolation or it is not proteomics in isolation.

If these technologies are robust and mature and the quantitative, and there is the ability to combine them to me that is really what fulfills the underlying story of biology, and really could push precision oncology even further. So, with that I want to thank everyone for your attention and I will be more than the glad to address questions at this point. Thank you everyone.

Student: Sir though it might be more ambitious what I am asking for, why are we restricting NCI to only cancer research data commons rather than expanding to Diabetes and Aging.

Ok, so, if I understand it correctly the question is for the data commons of NCI can you include aging and diabetes. So, aging is part of it because that is part of the metadata that we want right. So, all the electronic health records comes along with it, but diabetes is a

very difficult one, because obviously. So, the way that I phrase it is if you look at NCI, C is cancer.

So, but you know it is interesting because one of the things that within our assay portal, 2 years ago people were asking me, we would like to deposit our assays into your portal. And actually they were coming from the diabetes institute. And I will be honest, my first reaction was no, because ours is oncology. But the more that I started to kind of think about it is really almost all biology plays a role in all these diseases. So, I had say as things evolved there is always these possibilities.

Student: And, what I would like to know is there any patient based information available which can be downloaded with the huge database you already managed to accumulate, do you have data regarding patient treatment of cancer, do you have data on colorectal cancer etc.

As ID identified yeah.

Student: Yeah.

Yeah.

Student: So, there is separate datasets.

There is a couple of in there yeah. So, within the US, there is a couple of sites that you are allowed to download certain types of information.

Student: Ok.

So, but when you apply to it, obviously, you have to identify what the information is going to be used for, because the higher level of criteria. So, typically treatment naive information that is more simplistic at your access to, but the more you start moving into that space. And if it is de-identified, and if it is made available, it is just a higher rigor to get access to it.

Student: NCI is now also investing in this alternate and medicines etcetera. So, is there any kind of you know future thinking of alternate medicine.

So, when you say alternative medicine, I am assuming you are talking more natural products.

Student: Yeah, yeah.

Yeah. So, now, one I actually do not know if that is it is in the population set, I do know NCI does have a big repository of natural products, and they are trying to identify more application that it could be applied to. Quite frankly, I do not see why not, if that would not be part of it, but at least I am not aware of any activity at the moment.

Student: One question.

Yeah.

Student: I am not from cancer background.

Alright.

Student: I am from a different background. But my curiosity to know about this ok, it is fine whatever you are doing to treat to get out of with this cancer but you have is there anything like that what kind of people would be most prone to cancer that kind of.

So, the question is and let me make sure that I understand it. So, yours is are you able to identify individuals that could be more susceptible to development of cancer.

Yeah.

So, those are epidemiology based studies right, where you are trying to understand the environment, the food, all those components I mean those are separate components within the NCI. Those things do exist as part of the organization that they go after, that is the number of question that I get.

So, how come you guys still do metabolites, oh my god you guys need to metabolites. So, I yeah, so, I personally like to simplify things and from what I have seen is that it is already so complex quite frankly just to mix proteomics with genomics and transcriptomics, I did not know another layer of complexity even makes it more complex.

So, what I tend to always want to know is a very simple question. You ask the people already that are understanding that disease and then you ask them do you already use metabolize first and foremost to screen something of that individual for that disease of interest. So, for example, GBM writes like brain cancer people develop panels now. So, we do have a GBM project metabolite is part of the formula, but at least what we try to avoid because, it just adds additional time and cost is we really want to make certain that you just do not do something for the sake of doing it, you have some logic on why you would like to do it.

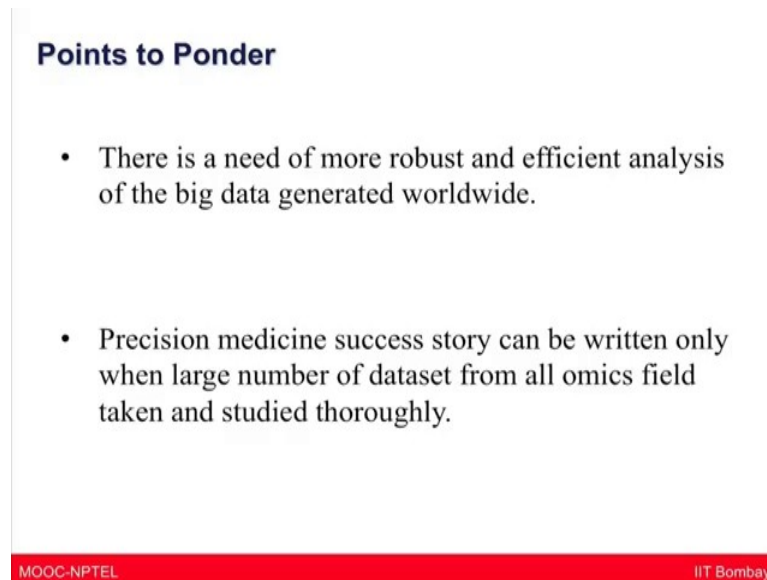
But the reality is we are playing with it, but not to the scale that we mix DNA and RNA and proteins, it is just a side project at the moment.

(Refer Slide Time: 25:54)

Points to Ponder

- ICPC, Apollo Network and Cancer Moonshot are some of the other organization and consortium that are managing Multimodal Data.
- The Genomics, Transcriptomics, Proteomics, Imaging data when come to under a same roof and analysed properly can give a huge number of unrevealed facts.

(Refer Slide Time: 26:05)



Points to Ponder

- There is a need of more robust and efficient analysis of the big data generated worldwide.
- Precision medicine success story can be written only when large number of dataset from all omics field taken and studied thoroughly.

MOOC-NPTTEL IIT Bombay

In conclusions, we understand that how much initiative has already been taken by the NIH to manage multimodal data in the form of repository and global databases. The genomics, transcriptomics, proteomics, imaging data, when kept under the same roof and analyzed properly could provide many new facets, many new unrevealed facts. Precision medicine success story could only be written when large number of data set from all omics field are together analyzed thoroughly, and then only meaningful conclusions can be drawn. Though we are generating large amount of data from NGS platforms and mass spectrometry technologies, but whether these big datasets are fully analyzed proper QC checks have been performed, we have to look into all of these very carefully.

So, let us thank Dr. Rodriguez for his wonderful lecture which has really setup a good stage for this course, why there is need to look at new approaches of proteogenomic analysis. Now, we will move onto the modules; the very first module on the genomic technologies. And, the first lecture of that will be given by Dr. Kelly Ruggles, next week talking about introduction to genomics.

Thank you.