

**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. Ratna Rajesh Thangudu**  
**Department of Biosciences and Bioengineering**  
**Indian Institute of Technology, Bombay**  
**Bioinformatics Lead, Enterprise Science & Computing, Rockville**

**Lecture - 31**  
**Large-scale data Science - I**

Welcome to MOOC course on Introduction to Proteogenomics. Today, we are going to have a new speaker Dr. Ratna Thangudu; he is a bioinformatics lead in Enterprise Science and Computing at Rockville, MD, USA. His company deals with large scale data management and provides bioinformatic solutions to various institutes and companies. Dr. Ratna going to talk to us about large scales data sciences. He will explain what exactly the term big data refers to and how it can be managed.

He will also talk about the major issue with the big data, and how one can overcome it by sharing the data from all the fields whether it is academia or industries. He will also talk about the importance of multi-omics data in understanding biology especially in context of precision medicine. So, let us welcome Dr. Ratna to talk to us about large scale data sciences.

My name is Rajesh Thangudu; I am the bioinformatics lead for the company called Enterprise Science and Computing. So, we are located in Rockville, Maryland, USA. So, we have kind of 10 miles away from the National Institutes of Health, so that is about 20 minutes drive and to put it in a context we are about 30 minutes from the White House that is where we are all right.

So, what are we doing actually as a company? So, we are into a large scale data management of and provide bioinformatics solutions for government clients, academics and also industry and we work pretty closely with the National Cancer institute. So, you heard a lot about the CPTAC program for the last few days. So, we actually built and managed all of the resources that they were using. So, I will talk a little bit about that. So, how many of you are actually aware of what is big data? I see a few hands just raising. So, I think I mean you are all part of big data, every day you are contributing to big data.

So, I will start from there and try to fit that into the perspective of big data in biology and what we are actually doing with the proteomics and how it how it is all going there all right.


(Refer Slide Time: 02:55)

**Big Data**

noun COMPUTING

extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.  
"much IT investment is going towards managing and maintaining big data"

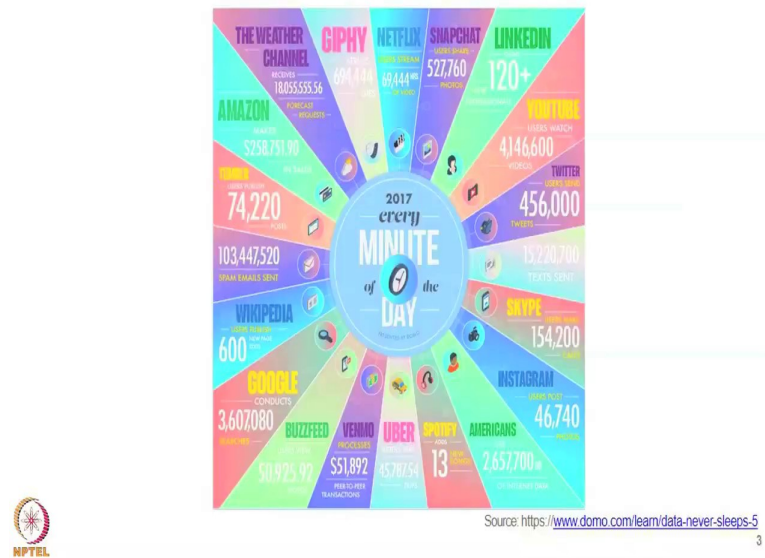
- Very large
- Dynamic
- Too complex for traditional data-processing application software
- Both structured and Unstructured
- Offers more statistical power
- Complexity may lead to higher false discovery rate

 2

So, say the word basically big data means it is an extent extremely large data set that may be analyzed computationally to reveal patterns, trends and also associations which are not easily seen by our regular kind of day-to-day analysis and as the datum or the name it actually implies, it is very large that is understandable that is big data and then it is very dynamic it keeps growing it keeps changing and it is very complex to actually use any of your traditional data processing techniques.

For example, take your excel file. So, excel can handle at the most 10 million rows that is it. So, if it crosses that, what would you do, so that is called big data, anything that is crosses the pad that you cannot actually analyze with the tools that you have on your desktop that is actually big data. So, with large amount of data it offers lot of statistical power and the complexity actually leads to some kind of false discovery rates, but that is ok, but we are getting lot of statistical power there.

(Refer Slide Time: 03:58)



So, to give an idea of what exactly is big data this is the social media big data that we all contribute to on a daily basis every minute. This is the number that is coming from 2017 for how many what kind of interactions that we do on a daily basis that contributed to the big data. So, I would not show you a lot of things, but for example, Google you see in every minute we almost do 3 million searches. So, every search that we do is a data point, it is not necessary data it is returning but the search itself is a data point. Similarly, you watch lot of YouTube videos.

So, the amount of time, you spend on YouTube that is a data point and the amount of video content that you upload, it is a data point. So, every interaction that we do on a day-to-day basis is a data point that actually contributes to the big data all right.

(Refer Slide Time: 04:51)



So, what is big data again, it certainly involves large quantities of data, but it has some characteristics. Everything if I give you some I say something like a big file, I say this is data that is not big data right just because the file is big it is not big data. So, it has some characteristics some features to it. So, this actually started off with describing volume that we just discussed, and then there is velocity, how fast you can access it, how fast the data can actually move from point A to point B. So, for example the video streaming, so Netflix, YouTube that you see.

And also the variety of the data, so is it all the same kind of data no there is a lot of structured data, there is lot of unstructured data. Structured data refers to basically anything that kind of interactions that you do the airline targeted system. So, the banking transactions all of the e-commerce things you do, they are all kind of structured, you know well which person is doing what at what time and what amount he is spending that is kind of structure, but what is unstructured.

So, all the emails that you send every day, it is unstructured. So, because it is all text. So, you cannot actually assign them the categorize them by words that is about just one example. But as people started looking into the big data, they these are I mean this is not enough, they came up with more descriptions ok, they added veracity. Is the data actually valued I mean does it make any sense at all? It should be some value data right.

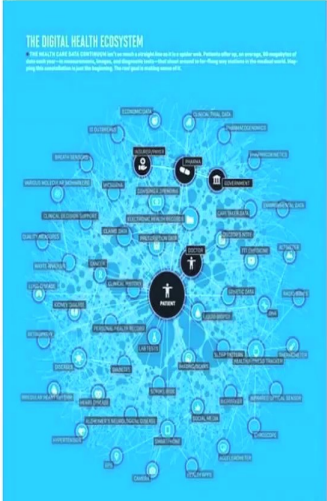
So, then there were some more people who came and they said let us add one more what is that it is called veracity, rigid veracity I think value. So, does it actually add any value right, at the end of the day you have so much data I am talking so much data like I showed so many points. If that is all not generating money for the companies Google or YouTube, it is of no value right. So, it has to have value also and then they added I mean people came up with more things just like they said after veracity they said visualization.

So, can we actually look at the data and say something about it, and then they have something called (which are missing) viscosity, does it stick, is it with you I mean does it make any sense at the end of the day? So, all this seems like what I am trying to say is basically big data is not just one big file that you have or a bunch of files that you have, it has to have some meaning to attach to it when you analyze all right. So, now, I jump to big data in biology. So, we all know I mean next generation sequencing, you are all aware of we have been discussing for the last several days maybe you are all doing your own research and in that area.


(Refer Slide Time: 07:30)

### Big Data in Biology

- Biologists are joining the big-data club.
- Many forms of life science data – genomics, proteomics, molecular pathways, healthcare, etc and from different populations of people.
- If we can handle the complexity of information, those data create a potential bonanza.
- Tools and techniques for analyzing big data promise to mold massive mounds of information into actionable intelligence.



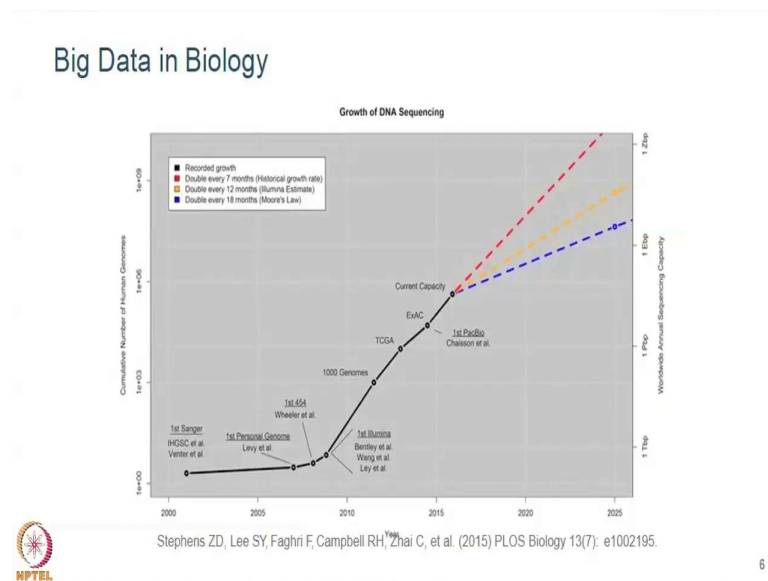
Source: <http://fortune.com/2018/03/19/big-data-digital-health-tech/>



So, biologists we joined a big data club long time ago, I think I would said the advent of the human genome about 10 years ago or 20 years ago and then so there are many forms of data right, even when within the biology there is genomics, proteomics, molecular pathways and there is one thing called healthcare. So, all the doctor visits that you do it is recorded in electronic health records. So, it is pretty popular back in the US and western

world, I am not very sure I mean probably the new corporate hospitals they are actually doing it already. So, there is a lot of EMR data and all the visits that you do, all the tracking the health care tracking devices that you use fitbit, it records lot of data. See the all if you bring all of them together, it tells something about you, you are predisposing to or you a particular disease right. So, it allows us to develop new tools, new techniques, new ways of understanding the data all right.

(Refer Slide Time: 08:33)



So, just to give you an idea of where we are actually going with the data that we have all right. So, this is 2000 that is around with the first Sanger sequencing came. So, then the human genome project came, and then we have TCGA we heard about that, so that is about 10. We have first 1000 genome projects that is about 1000 genomes, and then we have TCGA that is about 11000 patients' data.

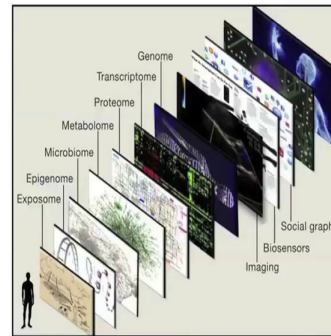
So, then the exome sequencing came that is about 68000. Now, there is an institute back in the US that does actually a 100,000 genomes and now people are talking about a million genome, and then there is a new initiative started back in US, it is called all of us. It is basically looking at almost a million and plus cases across the country.

So, basically when we are adding there a day by day, so as from more slide it has to increase like this, so every 2 years, the data will double but Illumina when they are down who are actually developing this NGS machines they basically said over no it will go like

this. But actually if you see it is actually going at that rate that is a projection, so the hexabytes of data. I think I convinced you about how much data is there.

(Refer Slide Time: 09:53)

- Multi-omics data integration is the key for precision medicine approaches.
- Considerable work has been done with the advent of high-throughput studies, which have enabled the data access for downstream analyses.
- A gamut of algorithms and tools developed to improve the clinical outcome prediction.

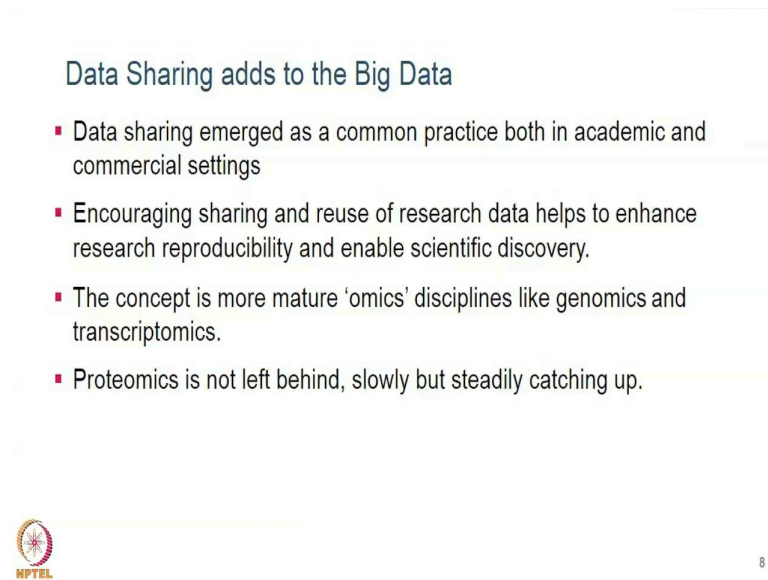


Eric Topol, Cell VOLUME 157, ISSUE 1, P241-253, MARCH 27, 2014



And coming back to the actual, actual multi-omics data, so it is not right at the example I showed before in the earlier slide is just the genomic data I am talking about. But there is so many facets to the multi-omics data you have transcriptome, proteome, metabolome and exposome and epigenome and also the social graph with the demography of the patient, are the people it is not just a patient. So, then there is imaging data there is biosensors. So, you have to bring all of that together to actually make sense to what if you want to achieve the goal of the precision medicine, so that is the personalized medicine all right.

(Refer Slide Time: 10:32)



**Data Sharing adds to the Big Data**

- Data sharing emerged as a common practice both in academic and commercial settings
- Encouraging sharing and reuse of research data helps to enhance research reproducibility and enable scientific discovery.
- The concept is more mature 'omics' disciplines like genomics and transcriptomics.
- Proteomics is not left behind, slowly but steadily catching up.

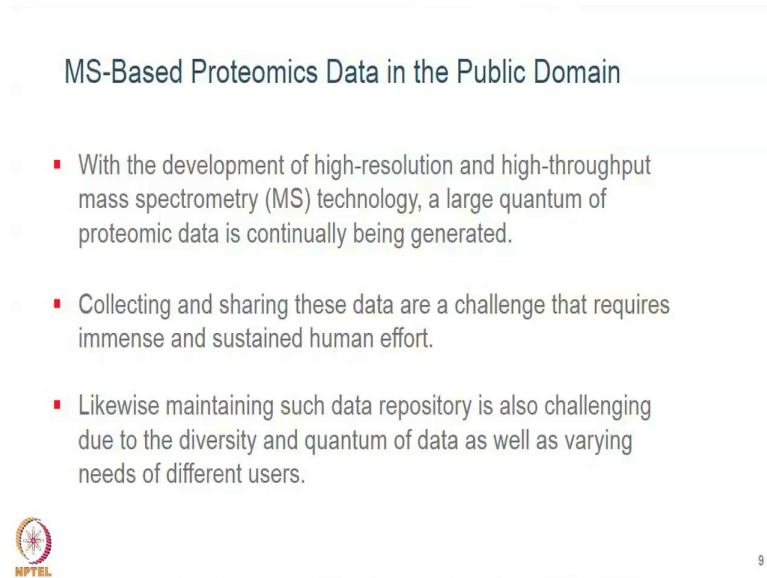
 8

So, where is this all data coming from? So, people are actually generating lot of data. If people are not sharing it is not big data anymore right. So, you generate for 10,000 patients, your group generates the data and you keep it with you, you do not share with anyone else. So, then it is not a 1000 genome that is your genome you are elapsing, but it is actually rapidly change, it actually changed I mean all of the databases that we have now just because of public sharing of the data.

So, if you run a blast and NCBI, you are running again say a reference genome where is it coming from because people shared the data publicly because it is all funded by the governments. So, whichever I mean all of the European and American country US subcontinent basically they whatever the data that is funded by the government that has to be in the public domain that is a requirement all right.




(Refer Slide Time: 11:32)



MS-Based Proteomics Data in the Public Domain

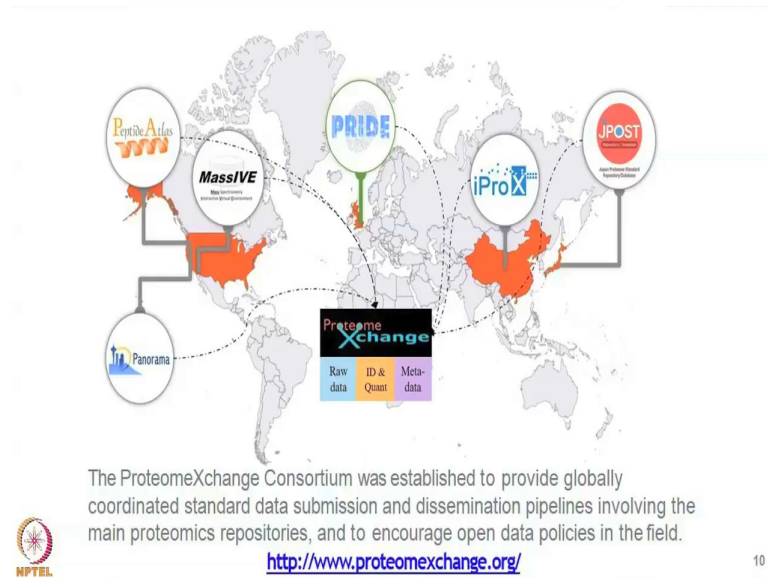
- With the development of high-resolution and high-throughput mass spectrometry (MS) technology, a large quantum of proteomic data is continually being generated.
- Collecting and sharing these data are a challenge that requires immense and sustained human effort.
- Likewise maintaining such data repository is also challenging due to the diversity and quantum of data as well as varying needs of different users.

 9

So, so but we are all into proteomics right. So, where does proteomics stand here? So, there is a lot of proteomic data out there in the public domain. I am not sure how many of you actually aware of that. I will show some slides there probably, you are all aware you might have seen it or some of you actually used it. But I mean with the advent of the high resolution mass spectrometry and collecting and sharing is a actually a big challenge right. You run the instruments and you saw how much time it took to understand the data. So, after that what you do? So, you run your experiment, you analyze it, probably you will have a publication and then what you will do?

So, the pub the publisher basically now wants you to actually put that somewhere that is accessible by both the publisher and also from all the people right. So, that is all actually contributing to the public data and likewise actually even though the sub publisher says you submits here, actually for the people who are managing the data that is the repositories it is a herculean task, it is extremely difficult to manage it that is coming from so many different places all right.

(Refer Slide Time: 12:50)



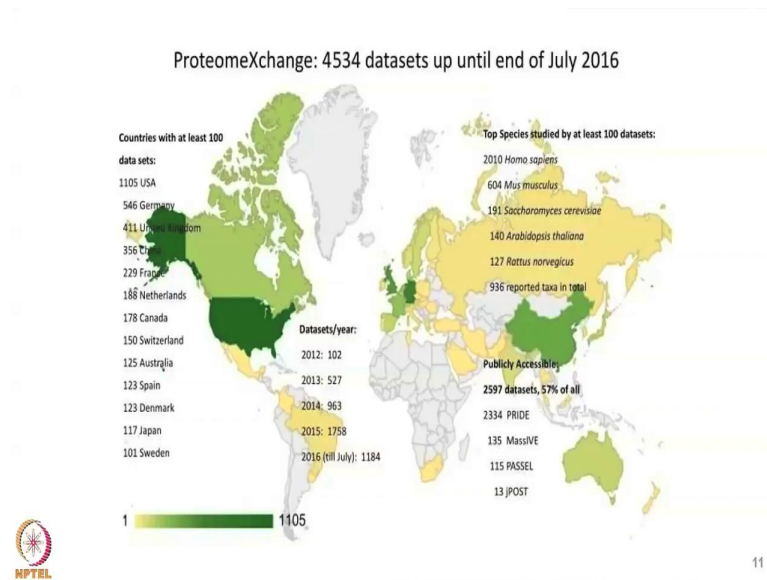
So, we did a very good progress in terms of the where we stand in the proteomics community. So, there is this consortium ProteomeXchange probably lost lot of you are already heard. So, this is a consortium of about 6 groups there are the 6 repositories. So, the pride is the central one, it is the largest public data resource, so that is sitting in Europe and then we have massive and peptide atlas I think from ISB David Campbell is here, it is part of that. And then you have Panorama most of you are actually accessed it and then we have from China and also from Japan. So, there are different resources out there. So, they have all of the data publicly available.

Student: Sir, I have a question when we submit the data to proteomeXchange, proteome Xchange, are they very validating our data or just they are sending the that is you know this is the ID what I mean. Because I am wondering like we are supporting after 3 days we are sending your data is successful.

Student: So, are they checking each and every thing like we have given the file or that search engine file or raw file accordingly or not.

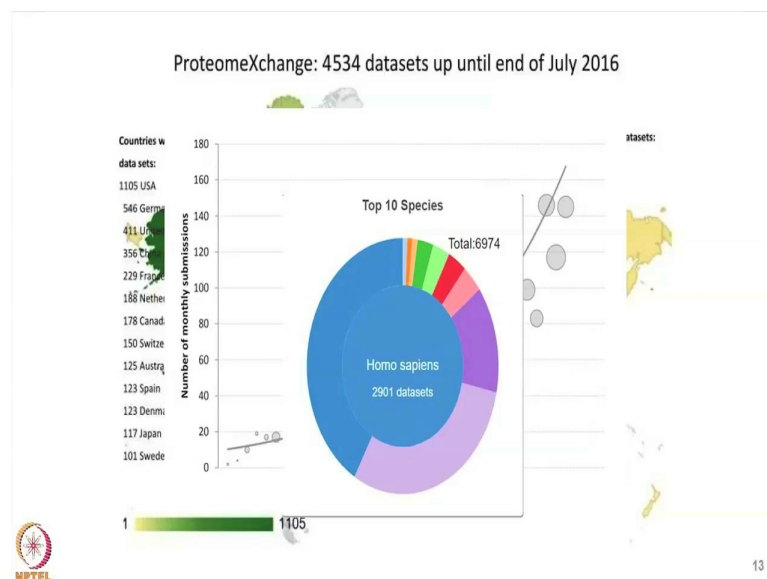
That is a very good question. So, I have a slide next slide, I will talk about that. Earlier like when these resources started, they actually started taking your raw files and your results right. So, the people and their so we do not know the format of it, so we do not know the validity of it.

(Refer Slide Time: 14:24)



But now let me go to the next slide all right. So, many data of are concerned, so many different places right. In 2016, there are about 4000 data sets are available. There by issue of the growth of the data.

(Refer Slide Time: 14:40)



So, they started with just take view whatever you have and also the results also. So, just see here it started in 2012, where we are and just see in 2015 this is a dated slide. This is in 2015, it went to that level and as of few days back I checked just a couple of days back there are 10 species, and they are about almost 7000 data sets public available through

that ProteomeXchange. It is not just from one resource, but all the six or seven resources I showed you.

(Refer Slide Time: 15:11)


### MS-Based Proteomics Data in the Public Domain

**Table 1** List of major MS-based proteomics resources

Category	Name	Link	Main features	Rating	Refs.
General	PRIDE	<a href="http://www.ebi.ac.uk/pride/archive">http://www.ebi.ac.uk/pride/archive</a>	Supports raw data storage and data submission	★★★★★	[3]
	PeptideAtlas	<a href="http://www.peptideatlas.org">http://www.peptideatlas.org</a>	Supports raw data storage, data submission, and data analysis	★★★★★	[1]
	Human Proteinpedia	<a href="http://www.humanproteinpedia.org">http://www.humanproteinpedia.org</a>	Supports raw data storage and data submission	★★★★★	[4]
	iProX	<a href="http://iprox.hupo.org.cn">http://iprox.hupo.org.cn</a>	Supports raw data storage, data submission, and data analysis	★★★★★	
	Tranche	<a href="https://proteomecommons.org/tranche">https://proteomecommons.org/tranche</a>	Supports raw data storage and data submission	★★★★☆	[5]
	GPMDDB	<a href="http://www.thegpm.org">http://www.thegpm.org</a>	Supports data analysis	★★★★☆	[6]
	MOPEd	<a href="http://moped.proteinspire.org">http://moped.proteinspire.org</a>	Stores protein expression information from MS-based proteomics experiments	★★★★☆	[7]
	YPED	<a href="http://yped.med.yale.edu">http://yped.med.yale.edu</a>	An integrated bioinformatics suite and database for proteomics research	★★★★☆	[8,9]
	Quantitative PTMs-focused	PaxDb	<a href="http://pax-db.org">http://pax-db.org</a>	Supports quantitative proteomics data storage	★★★★☆
Phospho ELM		<a href="http://phospho.elm.eu.org">http://phospho.elm.eu.org</a>	Supports phosphoproteomic MS data storage	★★★★☆	[11]
PhosphoSitePlus		<a href="http://www.phosphosite.org">http://www.phosphosite.org</a>	Stores raw data and MS-reported PTM sites	★★★★☆	[12]
dbPTM		<a href="http://dbptm.mbc.nctu.edu.tw">http://dbptm.mbc.nctu.edu.tw</a>	Stores raw data and MS/MS peptides associated with PTMs	★★★★☆	[13,14]
PHOSIDA		<a href="http://www.phosida.com">http://www.phosida.com</a>	Supports raw data storage and phosphoproteomic MS data storage	★★★★☆	[15,16]

Note: These web resources are rated based on their score against parameters such as storage of raw data, data submission support, and provision of data analysis pipelines MS, mass spectrometry; PTM, post-translational modification.

Chen et al Genomics Proteomics Bioinformatics 13 (2015) 36–39

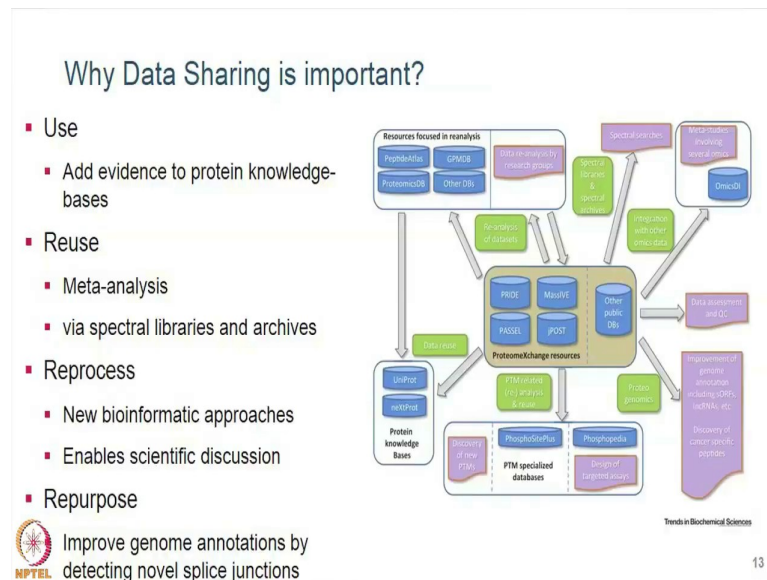


14

And I will be answering your question in a bit. So, here is another list of public databases that are there. So, some already I already listed there, and then some these are actually derived databases, these are the post translational modifications how do you capture. So, there are separate specialized databases over there and like I said it is not so easy to manage all though data resources, it needs a lot of money, a lot of manpower and a lot of expertise.

For example, here the tranche or tranche is one data resource, it does not exist anymore, it is not there anymore because it is lack a funding, so that is one place actually we took some of the data for the CPTAC, CPTAC used to deposit data in tranche before. So, as the part of the CPTAC data console data co ordinary center, I will discuss later but we got all of the data and made it available again ok.

(Refer Slide Time: 16:07)



So, this is I think I will try to answer your question here. So, what are the uses of this data out there, one thing you get a publication. So, publisher requires you to submit the data somewhere, so that he can validate them. So, the primary use is the why you generated the data. So, you have the publication that is the primary use of the data. So, and then it adds evidence. For example, in the uniprot you have all these manually created validated protein sequences how is it coming. So, it takes evidence from all the publications, people are actually publishing, and so that adds value that is the primary use.

And then there is reuse meta analysis, so you can take data from the 10 different data sets which are similar to your work, but you never knew that they exist, just because you went to one of the resources you could find you just search for example, colorectal cancer, maybe you will find something there right. So, you go to a publication, I mean you go to a PubMed you search for something. So, you get literature, you do not get data. So, there is a difference.

So, things are slowly changing, but as time goes by the idea is when you search in PubMed, it is not just the publication that comes out, it also tells where your data is sitting, what pipeline has been used, and how the pipeline has been run, and can you actually reproduce the results yourself with a click of a button. So, that is where people are trying to head. So, we long way to go there, but the vision is there.

So, to answer your question basically some of the resources they actually reanalyze all of your data. So, massive for example, they will reanalyze all of your data through their own pipeline and if they find something interesting that did not come through your results, they will they are nice to send you an email saying that we did find it maybe it will be interesting to you that is from the UCSD. That will be in the partial submission or complete submission. There are two options in the PRIDE. I would imagine like through the process of submission.

Student: Yeah

So probably I think it is once it is complete submission because it is very difficult to go back and reanalyze all of the data right. So, sometimes they try to analyze all of the data together, so that needs lot of computing power all right.

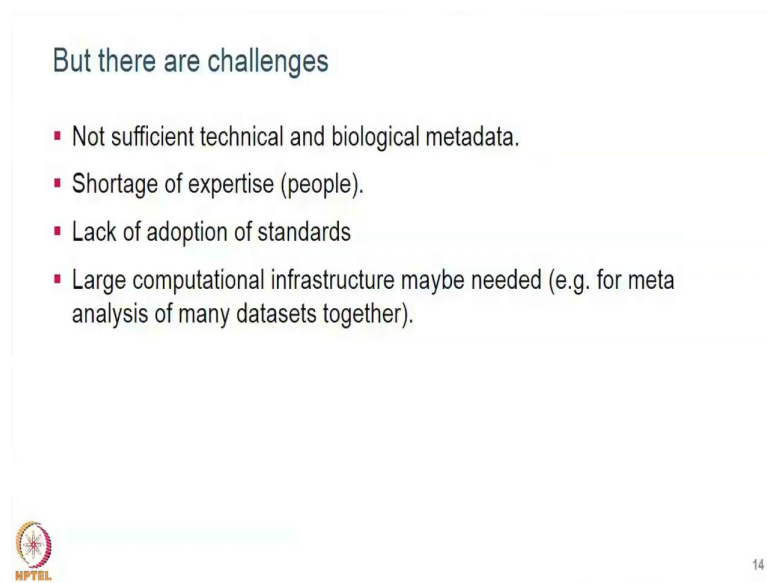
So, I talked about the meta-analysis. The reanalysis for example, here reprocessing. So, all these resources I mean these are the center you see all the primary consortium members ProteomeXchange depositories, once you deposit there are other resources like peptide Atlas, and all the other GPMDB and data actually take the data and reprocess them. So, that gives you insights. And it allows for example, this I already discussed and then it provides value to the adds value to the phosphoproteome post translation modification data sets, data repositories, and then finally, there is repurposing.

So, you start with the purposes of your experimenting something, but you use that to add value in a different way. So, one example is basically the proteogenomics approach right. So, you have all this data you take it and you deposit the data, but as a user I see this is very interesting, I will try to find if there are any novel splice junctions I can find. So, I will find so then I will add that value back to the either NCBI or somewhere I deposit them.

So, basically I am trying to say that data sharing actually helps in use of the data, and reuse, and reprocessing and also repurposing, I did not talk about reprocess ok. Here it helps when you always use some pipeline. So, the pipelines continuously and constantly evolving new algorithms are coming. So, in the first attempt, you might miss it, but the new algorithms actually might find a new information from the same dataset, so that is one thing the massive does all right.

So, we talked about I mean everybody is generating data depositing there I showed all most 6000 datasets there but if you actually take a look back at any of these resources they collect very minimal metadata because it is very onerous. Metadata is data above data you describe your data, what patient is coming from, what samples what protocols you used, how did you do that, what is the experimental design. So, if you do not provide all the context about your datasets it is all useless right. So, it just sits in there and I go there and I try to get all the data and I cannot do anything. I do not even know which patient it is coming from.

(Refer Slide Time: 21:00)



But there are challenges

- Not sufficient technical and biological metadata.
- Shortage of expertise (people).
- Lack of adoption of standards
- Large computational infrastructure maybe needed (e.g. for meta analysis of many datasets together).

HPTeL 14

So, there is not sufficient. The problem here is always the data submission making available, making them available in the public domain it starts after you and your research goal, right. So, you finish your research objective you achieve that and then you go there, because somebody else tells you the journal says you have to deposit somewhere, so that is like a burden for a lot of people right. So, I am done with this, now I had to do all this.

Like if you actually go and submit in any of these resources, it is not so easy; you have to collect the data metadata in a certain way, and reformat it and submit to them. So, then they validate it if there are something missing, they will come back to you and ask this is missing, we cannot actually support unless you provide this. So, the amount of metadata

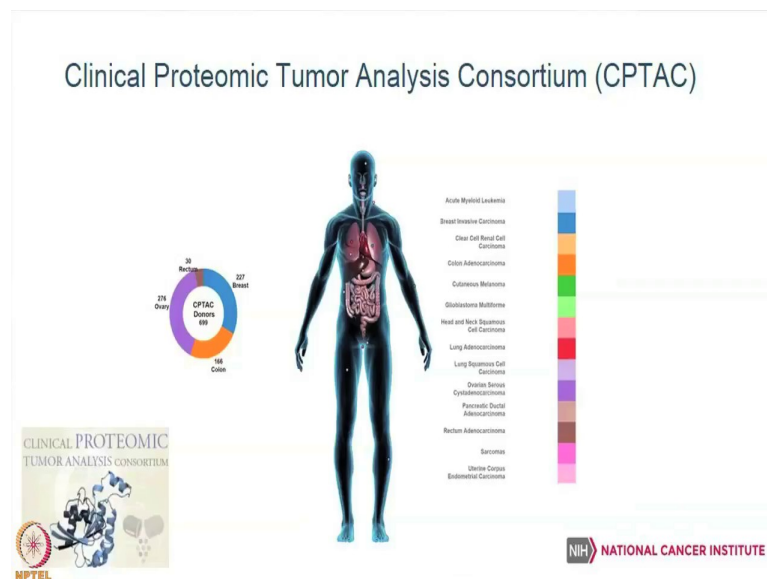
that they require actually they try to shrink it, because it is becoming burden and people will stop submitting it.

So, we want the data to come, so let us I mean you said give us a minimal metadata and we will just keep it there and see how we can process it, so that is not really helping and then there is short shortage of expertise, so as the data sets is growing, so many datasets, so much volume; so we need expert people to handle that. So, now the expertise is limited to the resources like pride and peptide atlas and all these.

And then like adaption of standards, I will talk a lot more about the standards in a little later, but standards is a how do you represent your data. So, are you using any of the existing controlled vocabularies to define what it is. So, you are saying TMT, and somebody says capital letters or small letters that is a very basic example I am giving. But if you actually go into the clinical data side, it is the same disease breast cancer there are so many subtypes. If you just say breast cancer, it does not help; you have to tell exactly what it is.

So, there are standards to help to do that but when you submit you have to do that actually, all right. So, I talked a lot about what is already there, I did not talk anything about the CPTAC, all right.

(Refer Slide Time: 23:11)





The CPTAC produces a lot of data, so that is the conservation we are part of it, we develop the we manage the data coordinating center and we also distribute all of the data. So, we started pretty low, so the CPTAC two program, three cancer types; breast, colon and ovarian about 600 patients. So, they started off with the some TCGA samples which are already there ok, there are genomics data is already existing. So, why do not we take them some 100 samples from each of these patients and reanalyze with proteomics and they try to combine. So, it is a phase 1 after that they realize ok, these samples are not optimized for proteomics. So, we need to collect more. So, then they collected 300 more for each of those cancer types.

So, there is lot of success and now the current running program is CPTAC 3. So, there are more at least 10 more cancer types they added. So, it is a very ambitious and large program, it is not in terms of the volume of the data that they are generating, because proteomic that is much smaller compared to the genomic data. But this the breadth of the coverage of the cancer types and what they are trying to do in terms of the proteogenomics it is pretty big, all right.

(Refer Slide Time: 24:26)

**CPTAC Data Coordinating Center**

CPTAC Private Data Portal (<https://cptacdcc.georgetown.edu/cptac/>)

- Coordinate proteomic data produced by the CPTAC data producers.
- Coordinate genomic, biospecimen, and clinical data on corresponding samples with the BCR, CDR and GDC.
- Establish standard operating procedures (SOP) for data transmission from CPTAC data producers to the DCC.
- Design and implement data analysis procedures that perform quality checks on incoming data

CPTAC Public Portal (<https://proteomics.cancer.gov/data-portal>)

- A centralized repository for the public dissemination of CPTAC proteomic datasets
- Analyze all of the CPTAC data through a Common Data Analysis Pipeline (CDAP) for public release
- Enable high speed transfer through UDP technology (Aspera)
- Provide support to the user community

CPTAC Assay Portal (<https://proteomics.cancer.gov/assay-portal>)

- The CPTAC Assay Portal serves as a centralized public repository of "fit-for-purpose," multiplexed quantitative mass spectrometry-based proteomic targeted assays.

ESAC  
Enterprise Science And Computing

HPTeL

So, we manage the CPTAC data coordinating center, this is basically the consortium has about 15 to 20 different groups or institutions. For example, for the last 3 days you saw at least 3 groups here representing; so one is broad institute, you have NYU – New York University and Bing Zhang is from Vanderbilt University. So, there are three different

groups they are working, so they just represented three groups, but then we have another 15 groups sitting there. So, all is all these people are actually generating data, so we have to actually coordinate that; so the data actually refers to clinical data, of the bio-specimen data, genomic data, proteomic data, imaging and so many aspects to it.

So, the private portal is specifically for the consortium members. So, it is a controlled access they can only login and they exchange the data and then we have a public portal, can you how many of you actually went to use this resource. Can you raise your hands?

Alright, so that's three. That's not good. So, the idea is I really encourage you to go to that resource. So, I mean the purpose of this stock is basically to introduce you to all these resources, there is so much data that is already there; start taking a look and you do not have to generate anything, it is already some so much is sitting there. Just to try, you do not even have to have a new discovery, just to try and then we also have a assay portal.

(Refer Slide Time: 25:52)



So, we did this in collaboration with George town university, that's in Washington DC.

Student: the Assay portal it shows you can see the method, but later on it shows no method is available here.

Is that true? Can you send an email to our help help email if you see on assay portal? We will definitely address

Student: It might be that they don't want to show the protocols that they have used in it.  
Ok so usually I mean it is very extensive

Student: In assay one other one is ok

Ok that's interesting to know because as far as I know we will put a lot of effort on accept what to make it, because for each essay there is stamp from NCI. So, you might have seen there is a check mark there, saying that it is kind of branded. So, if there is some information missing we will definitely take a look all right. So, this is the portals landing page, you can go to the proteomics dot cancer dot gov, and it will tell some information there and you can click a link.

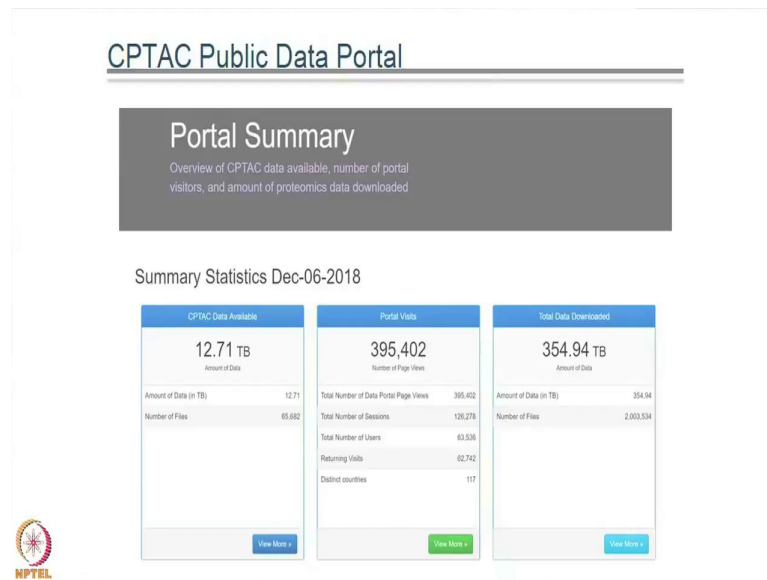
Student: Excuse me, Many a times we face issues with downloading few files from the site. Might be because of the issues within the servers.

So, it is we use a ASPRO technology to transfer, ASPRO its supports very high volume data transfer at very high speeds. So, but it uses a UDP technology. So, most of the universities academic institutions, they block it.

Student: limited access port

Right. So, it is not too hard like if you reach your IT department. For IIT and ask them to open a particular port, so we can provide sufficient information if it is a problem like, you cannot just go and ask them to open a port as a security whole. But if you want some information from outside, we can actually write to them saying that because all of the US we use that resource ok, all right.

(Refer Slide Time: 26:59)



So, we have about 13 terabytes of data right now, so that is about 43 studies. So, I talked about only about 3 cancer types, but there are 43 the data is organized into 43 different studies and so far from the last 6 years we have so many visits, like people coming and clicking and browsing our resource. And we have only 13 terabytes data, but it is been download if the actual download amount is about 400 terabytes close to 400 terabytes that is that is not a lot when you compare to compare with the genomics.

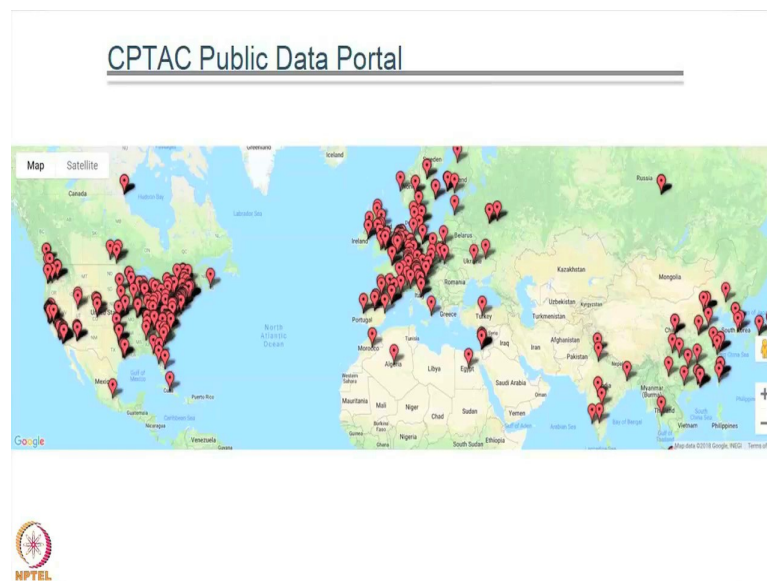
But what is interesting here is this number, the number of files that are downloaded is a almost 2 million or close to 3 million. What it means is along with the raw data we actually provide the results from the common data analysis pipeline that we run on the portal. So, these files are just the gene sample matrices that you were using for all the Morpheus hands on a session, so all right. So, what.

Student: Sorry, I did not get the last point. You said that people are not only downloading the raw file as well as what they are downloading?

I will just come there, basically for each study we run a common data analysis pipeline, so that generates the result files these are the summary reports there, routine parsimony results the identified peptides, identified proteins at a certain threshold. So, most people probably want those files, they do not need raw files. So, if you download raw files, it is not helping you if unless you have a established pipeline and resources to actually analyze reanalyze the data.

So, what I am trying to say here is basically just the volume of the downloads you see, it is basically a lot of people are actually interested in the results files that is what you want, you do not just go there and see there is a lot of data I have to download. It is always there, you do not have to worry about it; it is not going anywhere and you do not have to actually download if you do not want to, just go through the results and use data information, all right.

(Refer Slide Time: 29:14)




So, this is just trying to show how many people actually access the data from throughout the world and I see very few dots from here. But I think that will increase when I go back home, I think I will see a lot of dots.

(Refer Slide Time: 29:29)

**Points to Ponder**

- Sharing correct and curated data among industry and academics may facilitate solving big data and reliability issues.
- Important role of multi-omics data in precision medicine
- Data repository, ProteomeXchange for accessing publically available data for correlation studies and for reliable candidates.



NPTEL IIT Bombay

I hope you have learned that why sharing correct data is important, and how it can help people across the world to find solutions to the problems where individuals failed to solve. You got a glimpse of how contributing to big data, also helps in obtaining the reproducible datasets which could help in finding the most reliable candidates or even potential biomarkers for very datasets from different studies. I hope you have also learned about proteomicsXchange and its evolution, in terms of data with time. Hence one should emphasize, further on sharing the correct data with society.

In the next lecture we will continue Dr. Ratna's lecture, we will talk to you about larger scale data sciences and give you few examples.

Thank you.