Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Ratna Rajesh Thangadu
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 32
Large-scale data Science-II

Welcome to MOOC course on Introduction to Proteogenomics. In last lecture a guest scientist Dr. Ratna Thangadu, started providing you an overview for a large scale data sciences. Today he is going to continue his talk and mainly focus on the clinical proteomic tumor analysis consortium or CPTAC. I think it is really important to know that there are some resources publicly available which are sharing large amount of data set.

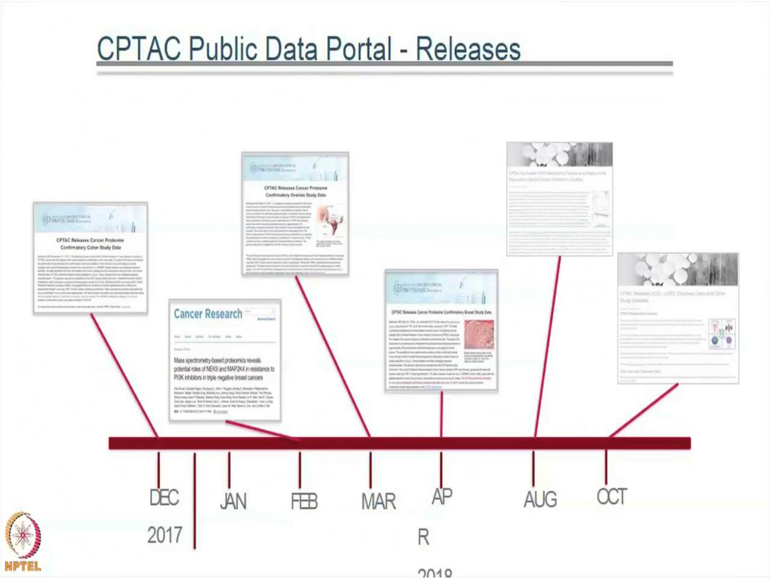
I think TCGA, the cancer genome atlas from National Cancer Institute was one such an initiative which is really provided large data set to the broad community. Thousands of patient's entire genome for various type of tumors were sequenced and then that data made publicly available. It was so interesting to note that many of the data set while the original data was published already in nature, but from the same data set many people started probing a specific question that what could be the effect of let us say survival for based on different genes and then you know they did metadata analysis publish papers based on that.

So, not only the original papers which generated data those are published, but also many associated papers came out solely based on the data available in public repository. So, the cancer genome atlas really made a huge impact to the broad community for sharing the data and making the publicly available. So, in this light the CPTAC provides a very good resource for the community to look at the proteomic data of different tumor types. So, Dr. Ratna will talk about the common data analysis pipeline of CPTAC and cancer imaging archive.

Various omics data repository such as even imaging data for various cancers, genomic data from the NCI-GDC portal as well as the DBGAP. Some of these are good resource repository is available for obtaining large data set which are already available from the you know very nicely done experiments using next generation sequencing as well as mass spectrometry. The huge amount of resource have been already put in to obtain these data and now the data is may available to public for doing further analysis.

So, Dr. Ratna will actually illuminate you, provide you more information about how these repository resources could be accessed and what kind of features are there and how you can make use of them for your own research, in which way you can analyze the data or use the data for your various comparisons. So, let us welcome Dr. Ratna for his second lecture on the big data sciences.

(Refer Slide Time: 03:25)



CPTAC Public Data Portal - Releases

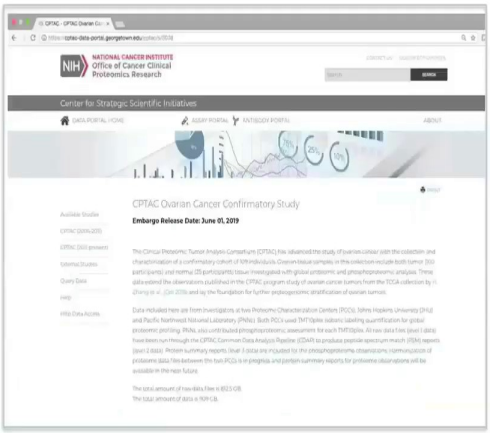
CPTAC Releases Cancer Proteome Confirmatory Colon Study Data
Cancer Research
Mass spectrometry-based proteomics reveals potential roles of NEK9 and MAP2K4 in resistance to PI3K inhibitors in triple negative breast cancers
CPTAC Releases Cancer Proteome Confirmatory Ovarian Study Data
CPTAC Releases Cancer Proteome Confirmatory Breast Study Data
DEC
2017
JAN
FEB
MAR
AP
R
2018
AUG
OCT
NPTEL

Alright, so, the interesting thing with CPTAC program is we do release all of the data even before the publications come out. That is a very interesting way to look at the data alright. So, if you are generating the data you are most of the times you are already worried if I put my data out there somebody else will publish, but that is actually hindering the progress of the research in terms of how the government says.

So, there is some of guidelines that come input fourth saying that we will give you the data, but there is an embargo date. So I mean I am just showing like we are pretty active, like every other month we will release some one study we harmonize all the study and at least on the top these 1, 2, 3, 4 those studies they do not have any publications yet. The groups are working. Dr. Mani is working on one publication, Dr. David Fenyo is working on another publication right now and Bing Zhang is working on another publication alright here.

(Refer Slide Time: 04:29)

CPTAC Public Data Portal - Releases



CPTAC - CPTAC Ovarian Ca...

https://cptac-data-portal.georgetown.edu/cptac/s/S038

NATIONAL CANCER INSTITUTE
Office of Cancer Clinical
Proteomics Research

CONTACT US | CLICK FOR DEFINITIONS

Search | SEARCH

Center for Strategic Scientific Initiatives

DATA PORTAL HOME | ASSAY PORTAL | ANTIBODY PORTAL | ABOUT

Available Studies
CPTAC (2006-2011)
CPTAC (2011-present)
External Studies
Query Data
Help
Http Data Access

CPTAC Ovarian Cancer Confirmatory Study

Embargo Release Date: June 01, 2019

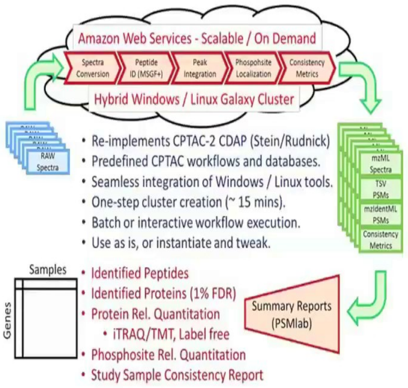
The Clinical Proteomic Tumor Analysis Consortium (CPTAC) has advanced the study of ovarian cancer with the collection and characterization of a confirmatory cohort of 109 individuals. Ovarian tissue samples in this collection include both tumor (100 participants) and normal (25 participants) tissue investigated with global proteomic and phosphoproteomic analyses. These data extend the observations published in the CPTAC program study of ovarian cancer tumors from the TCGA collection by H. Zhang et al. (Cell 2016) and lay the foundation for further proteogenomic stratification of ovarian tumors.

Data included here are from investigators at two Proteome Characterization Centers (PCCs), Johns Hopkins University (JHU) and Pacific Northwest National Laboratory (PNNL). Both PCCs used TMT10plex isobaric labeling quantification for global proteomic profiling. PNNL also contributed phosphoproteomic assessment for each TMT10plex. All raw data files (level 1 data) have been run through the CPTAC Common Data Analysis Pipeline (CDAP) to produce peptide spectrum match (PSM) reports (level 2 data). Protein summary reports (level 3 data) are included for the phosphoproteome observations. Harmonization of proteome data files between the two PCCs is in progress and protein summary reports for proteome observations will be available in the near future.

The total amount of raw data files is 812.5 GB.
The total amount of data is 909 GB.

(Refer Slide Time: 04:53)

CPTAC Common Data Analysis Pipeline



- Re-implements CPTAC-2 CDAP (Stein/Rudnick)
- Predefined CPTAC workflows and databases.
- Seamless integration of Windows / Linux tools.
- One-step cluster creation (~ 15 mins).
- Batch or interactive workflow execution.
- Use as is, or instantiate and tweak.

- Identified Peptides
- Identified Proteins (1% FDR)
- Protein Rel. Quantitation
  - iTRAQ/TMT, Label free
- Phosphosite Rel. Quantitation
- Study Sample Consistency Report

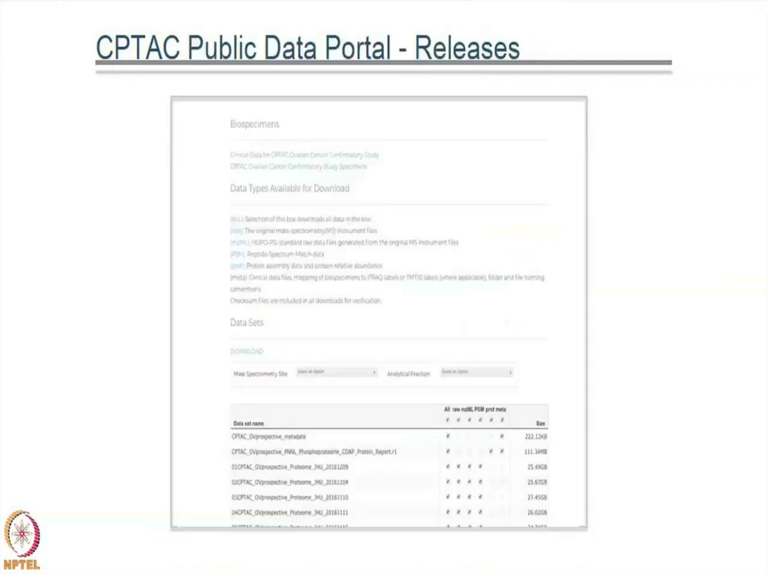
J Proteome Res. 2016 Mar 4;15(3):1023-32. doi: 10.1021/acs.jproteome.5b01091. Epub 2016 Feb 25.

So, there is something called a Embargo date. So, the expectation is that all the data is free, but we except use it, but you wait until that date before you publish because; that means, basically we are giving credit to the data generators from the conservation to publish first alright. So, this is the common data analyses pipeline that we have done. Every piece of data that comes into CPTAC portal, we have developed a Amazon based, galaxy based infrastructure to run a pipeline on that. This is basically as you can see here it is energy of search, then we have a custom built protein report summary generation and at the end the files that I am pointing out earlier is basically identified peptides, identified proteins and the quantitation data and also the study sample consistency report which is basically QC matrix. So, this is lot of information out there.

So, we thought downloading any of these raw files you can just use this information, but if you are not satisfied with the results always get those raw files and generate. And, another thing you would appreciate if you notice is basically all of the mass spectrometry files they are proprietary, that that they come from the instrument alright. So, you need to have a windows machine that is lot of complications how you convert them into open source; so, most of the pipelines run very smoothly and seamlessly and in Linux environment.

So, there is a bottom like that we have to do the first steps in the in the windows environment. So, what we do we follow open standards and we make all of the we convert all of the raw data into mzml. Mzml is the re representation of the proprietary formats in the open data formats. So, all of that information is available as mzml also and all the other information that comes as peptide spectral match and other things we convert them into the open source formats mzml mzidentml and so on wherever it is possible alright.

(Refer Slide Time: 06:41)



CPTAC Public Data Portal - Releases

Biospecimens

Clinical Data for CPTAC Ovarian Cancer Confirmatory Study
CPTAC Ovarian Cancer Confirmatory Study Specimens

Data Types Available for Download

(ALL): Selection of this box downloads all data in the row
(raw): The original mass spectrometry(MS) instrument files
(mzML): HUPO-PSI standard raw data files generated from the original MS instrument files
(PSM): Peptide-Spectrum Match data
(prot): Protein assembly data and protein relative abundance
(meta): Clinical data files, mapping of biospecimens to iTRAQ labels or TMT10 labels (where applicable), folder and file naming conventions
Checksum files are included in all downloads for verification.

Data Sets

DOWNLOAD

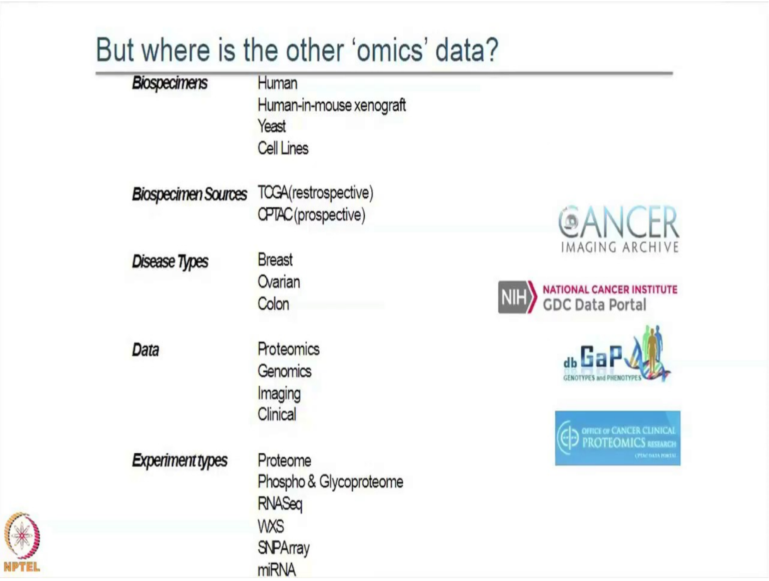
Mass Spectrometry Site: Select an Option   Analytical Fraction: Select an Option

Data set name	All ✓	raw ✓	mzML ✓	PSM ✓	prot ✓	meta ✓	Size
CPTAC_OVprospective_metadata	✓					✓	222.12KB
CPTAC_OVprospective_PNNL_Phosphoproteome_CDAP_Protein_Report.r1	✓				✓	✓	111.34MB
01CPTAC_OVprospective_Proteome_JHU_20161209	✓	✓	✓	✓			25.49GB
02CPTAC_OVprospective_Proteome_JHU_20161104	✓	✓	✓	✓			25.67GB
03CPTAC_OVprospective_Proteome_JHU_20161110	✓	✓	✓	✓			27.45GB
04CPTAC_OVprospective_Proteome_JHU_20161111	✓	✓	✓	✓			26.02GB

So, this is once you are on a particular study page, you see lot of 43 studies listed there. You can click on a study once you go there, so the first thing is because it is all cancer related patient derived so, we take utmost care to actually curate all of the data, the clinical information and also the experimental design. So, we will put all that information in these files. They are just simple excel files you can just download, but it have lot of information that connects you to the files in here and then we have also metadata packages. So, that has a all the protocol information, mass protocol information and also, we have the protein reports that I have mentioned earlier.

So, they are all packaged into one single packet that you can just download those two for example, you do not need anything else with the bottom ones are raw files and as you would might have noticed here it is like dot or 1. So, what that means, is we have we version all of the protein reserved report results. What happens sometimes we find something interesting that, we need to update the pipeline itself. So, then we will read on the pipeline and we report the results as the new version. So, always you can trace back what that means, alright.

(Refer Slide Time: 08:03)



But where is the other 'omics' data?

Biospecimens	Human Human-in-mouse xenograft Yeast Cell Lines

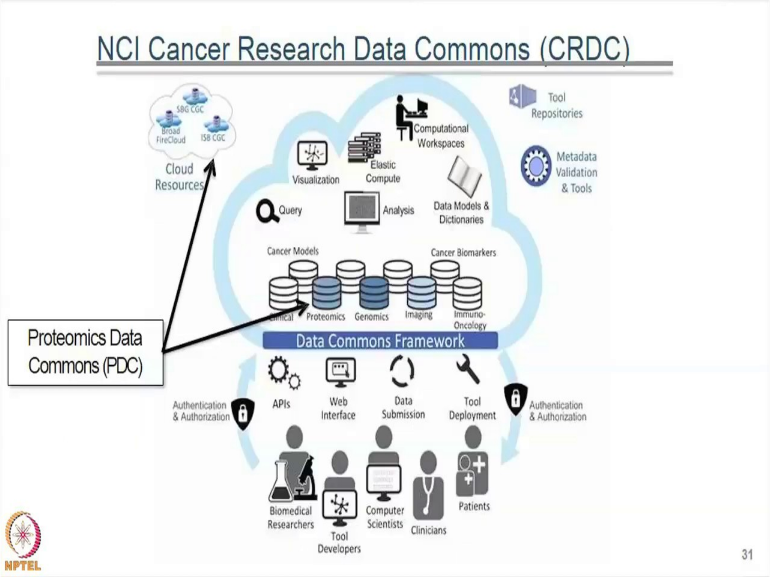
Biospecimen Sources	TCGA(restrospective)
	CPTAC(prospective)

Disease Types	Breast Ovarian Colon
Data	Proteomics Genomics Imaging Clinical
Experiment types	Proteome Phospho & Glycoproteome RNASeq WXS SNPArray miRNA

CANCER IMAGING ARCHIVE
NIH NATIONAL CANCER INSTITUTE GDC Data Portal
dbGaP GENOTYPES and PHENOTYPES
OFFICE OF CANCER CLINICAL PROTEOMICS RESEARCH
NPTEL

So, I talked a lot about all this mass spectrometry data I showed but where is all this other omics data, where is a genomics data, where is the clinical data, where is the imaging data? So, that is all CPTAC is producing, where is it sitting let us see ok. The imaging data sits in cancer imaging archive, the genomic data is sitting in GDC data portal, the other part of the genomic data that is a SNP array data it is sitting in DBGAP, that is the database of genotypes and phenotypes and the proteomic data is sitting here excellent.

(Refer Slide Time: 08:57)



NCI Cancer Research Data Commons (CRDC)

So, we have data from one single patient setting in four different areas. So, how do you connect all of them? So, we are generating data it is extremely difficult even for me to connect all the dots and it is not helpful and it is lot of information being lost. So, what to do? So, Dr. Henry actually mentioned about a precision medicine initiative that came to light in the last 4 years, say Joe Biden's or cancer moonshot initiative.

So, as a part of that the national cancer institute it is developing a cancer research data commons. It is a big ecosystem in theoretical ecosystem where all the repositories here these stacks are basically kind of repositories. So, genomics, proteomics, imaging so, on they all coexist together.

So, physically they are at different locations on different servers, but in the ecosystem so, they are together and then we provide the ecosystem provides analysis of all of the information. So, now, I talked only about earlier the common data analysis pipeline for proteomic data, but we will have tools the expectation that will have tools to analyze the proteomic component and then we will have data models and dictionaries to represent the data.

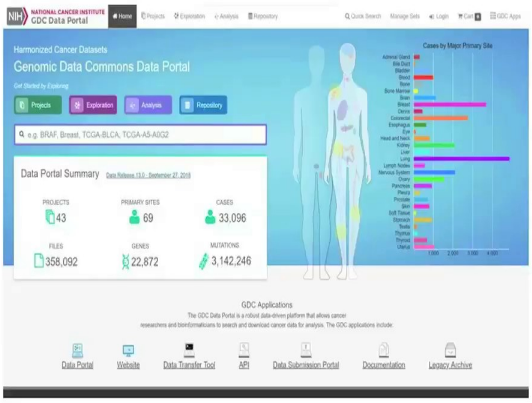
So, that I when I call a patient, I am calling that patient from all different resources at one time alright. So, then we have visualization, we have portal basically if you go to any portal you have all these features and the kind of users that we are expecting is you see there wide variety of users we are looking at. The patients, clinicians, computer scientists and tool developer and biomedical researchers so, everyone's export is different everyone's expectations are different.

So, such an ecosystem trying to support so, many different kinds of users, it is magnanimous support is needed alright. So, and then there are cloud resources. So, I have all this data sitting there, but I want to analyze the data myself. So, there are 100 datasets sitting I want I will pick and choose 3 or 4 different data sets and I want to analyze by myself. How will I do it? So, that is where we have cloud resources. So, Dr. Mani has talked about the fire cloud the other day.

So, fire cloud is one such resource. So, you just need to have a login, you do not have to take anything other than your credit card right. So, you go there, and you log in and you have the tools existing there, this is pipeline and you have data sitting there, everything is there. You pick a pipeline; we attach your data and you are on the pipeline that is it. You click a button only thing is I mean you have to understand what you are doing, but all the tools are available. That is that is the vision NCI has in terms of this research data commons.

(Refer Slide Time: 11:37)

NCI Genomic Data Commons (GDC)


Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 13.0 - September 27, 2018

PROJECTS	PRIMARY SITES	CASES
43	69	33,096
FILES	GENES	MUTATIONS
358,092	22,872	3,142,246

GDC Applications

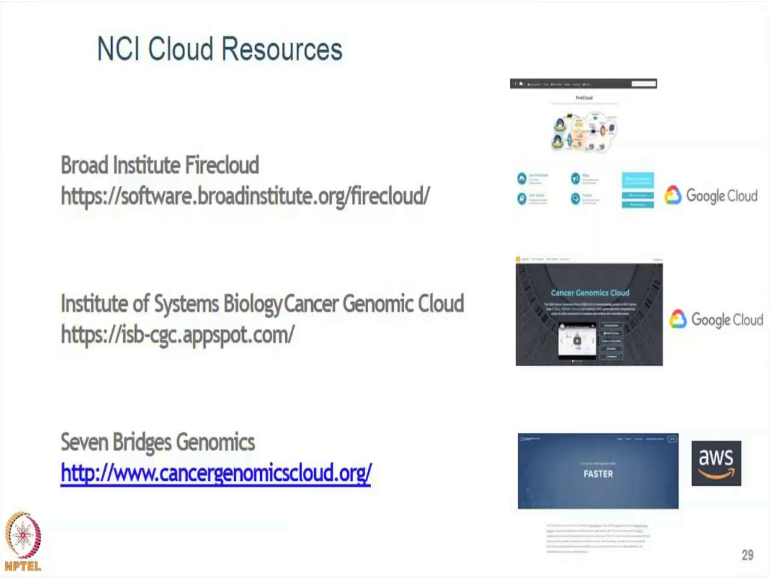
The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

https://portal.gdc.cancer.gov/

So, already there is one NCI genomic data commons which came to right in 3 years ago. So, most of you might have heard about the TCGA the cancer genome atlas and all of the data used to be available in a TCGA data portal. That is very specific to TCGA so, but now there are so many programs coming up, NCI thorn we have to bring all of this genomic data together at one resource. It is not just one program specific resource, but rather a common resource where all the genomic data will be there.

So, that is what they did and right now there are about 40 plus programs. There were the humongous program that takes So much data there are out there and this is free resource, there is no login required and you can see there are so many different cancer types available already. So, this is all about the genomic data.

(Refer Slide Time: 12:35)



NCI Cloud Resources

Broad Institute Firecloud
https://software.broadinstitute.org/firecloud/

Institute of Systems BiologyCancer Genomic Cloud
https://isb-cgc.appspot.com/

Seven Bridges Genomics
http://www.cancergenomicscloud.org/

29

And then I talked about the NCI cloud resources so, Dr. Mani already talked about the broad’s fire cloud. So, cloud is it is a public cloud. This is somebody else is actually offering you services so, you do not have to do anything, you do not need to have a cluster right. So, cluster yesterday when you are asking me I have this much data so, I do not have any disk space what do I do.

So, I buy a new disk I will attach it, reconfigure it and then 10 days later more clients come, and you generate more data. So, what do you do and there is no solution? So, people are slowly moving towards public clouds, these are Amazon web services and Google cloud. I can I mean anyone can actually have a free account, you can login try to export. The good thing with that is you do not need to have a lot of informatics expertise to begin with. Do not be scared just go out there, see what is there a lot of them are made as services.

So, I want to do this so, you click a button and you pay a price for that is minimum. So, so, Broad institute of system biology they have cancer genomic cloud and then seven bridges. Seven Bridges is a private company, it is a commercial partner with NCI in developing what they determine AWX. The Fire Cloud and the Seven Bridges I would recommend you to take a look at those two resources.

Institute of system biology is more of a program centric approach they took. So, they made all the data available in a certain way and you can access through API's programmatically. So, you need to have a little expertise understanding that, but they do have a some kind of a UI where you can actually look at the results. So, now, the genomic data commons and the cloud resources are the only things from the ecosystem I just showed they are existing as of today.

(Refer Slide Time: 14:45)

Proteomic Data Commons

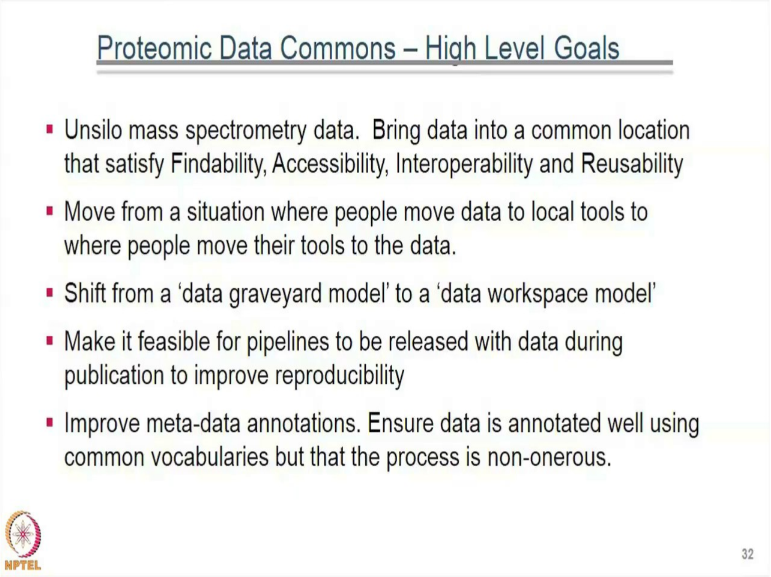
- The overarching goal of the NCI Proteomic Data Commons (PDC) pilot project is to ***democratize access*** to cancer-related proteomic datasets as well as to ***provide sustainable computational support*** to the cancer research community.
- NCI aims to enhance data sharing within the ***cancer proteomics community*** and ***beyond***.
- Also the PDC will co-localize proteomics data and analysis tools within ***a larger data ecosystem*** – CRDC, that supports multiple research and clinical data types.

30

So, last year this proteomic so, NCI thought although CPTAC that is sitting just like the ways TCGA used to sit. It is in its own specific resource. So, what do we do about that? That is why they come up with why do not we have a proteomic data commons because there are more programs that are coming out it is not just a CPTAC, but there are Apollo and ICPC that I will talk a little bit later and doctor Henry talked about that.

So, we need to have a similar resource, but we are asked to actually have both the GDC and the cloud resources combined. So, we have the data and also analysis tools at one place for the proteomic data. That is pretty ambitious, yeah that is what I mean. So, now, that is the proteomic norm, but it needs to do both the things that the GDC and the cloud resources doing together alright.

(Refer Slide Time: 15:43)



Proteomic Data Commons – High Level Goals

- Unsilo mass spectrometry data. Bring data into a common location that satisfy Findability, Accessibility, Interoperability and Reusability
- Move from a situation where people move data to local tools to where people move their tools to the data.
- Shift from a 'data graveyard model' to a 'data workspace model'
- Make it feasible for pipelines to be released with data during publication to improve reproducibility
- Improve meta-data annotations. Ensure data is annotated well using common vocabularies but that the process is non-onerous.

32

So, the very high level goes off the PDC at the proteomic data commons, Unsilo the mass spectrometry data, everybody is storing on their own local spaces. Do not do that, share it publically and the move from a situation where people move the data to the local tools. So, same thing I am telling do not download the data you bring your tool to the data because data is so humongous. It is not just your local storage, it is about how you transferred the data.

Does IITs network allow you to transfer so, much of data right? I do not think So, it is not worth it. So, instead you bring your tool go there to the cloud and analyze and then this is an interesting thing. I always I am shift from a data graveyard model to a data workspace model what this means is graveyard the one I talked about earlier you go you deposit your data after the life cycle of your research. That is like you are dumping it somewhere I am done with this I do not want this anymore alright, that is the data graveyard.

So, nobody looks at that kind of data, but the workspace model that we are proposing with the proteomic data commons is, you connect your instrument directly to the work space, on the PDC. So, the data directly moves from the instrument directly to the workspace. So, there you attach all your metadata. When you start when you are ready to analyze the data at that point you actually attach all the metadata, all the samples and the study design and which tools you want to use and click a button and run the pipeline.

Student: right now atleast for proteomics we do not have such tools where we are allowed There may some open resource for when it comes to proprietary software which has been used. There is no such I mean there is no way that we could analyze the files one by one cloud.

Yeah.

Student: So, in which case it becomes very difficult I mean so you first store the data on the cloud then you need to download it, analyze and then again store it back. So, it takes us back to the same problem. So, there is there any initiative that has been taken to being more and more company based softwares like to develop any such cloud system wherein the analyses could also be done and then there is no need for a physical drive that a data analyses can go.

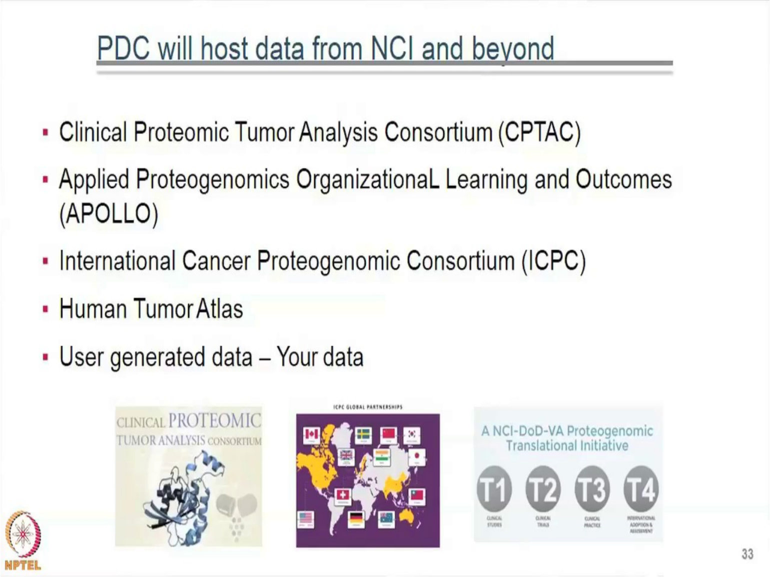
Yeah, so, that is the cloud thing actually doing. So, you can whatever the tool that you develop, you can dockerize that. So, dockerization may be you have to learn a little bit about that, but basically it is not just a tool, but the environment the computer system where it runs, you package the whole thing and put it on somewhere it is like a another database it is called dock store. So, once you have the tool you can take the that particular tool to the cloud and run it and like here I said these are the high level goals their goals.

So, we are not there yet there are lot of there we have to cross lot of hurdles to reach to those goals and I mean PDC we are starting small. So, we will make a couple of pipelines available initially and based on the user’s interest we will add more. So, that way we know what is interested in the community, you just I mean I cannot we cannot have a resource that tries to do so, much and then give it to you and you see are the users come and see like 90 percent of the stuff we are not interested in this.

So, that is not the good use of money or time or therefore, right. Did I answer your question? And then we have to improve the metadata annotations and ensure the data is annotated well like the standards I was talking about.

Alright, just to give you give you an idea of what kind of data you will see in future. So, CPTAC data that's I already talked about then we have Apollo and the international cancer photogenic consortium that is Dr. Srivastava is part of and then human tumor atlas and lastly I said user generated data, that is your data fine. So, whenever you start generating the data you can upload alright.

(Refer Slide Time: 19:45)

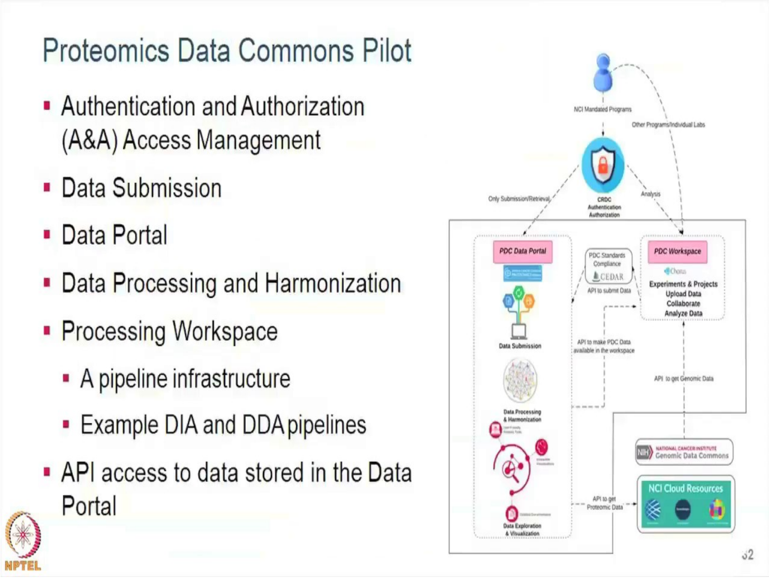


PDC will host data from NCI and beyond

- Clinical Proteomic Tumor Analysis Consortium (CPTAC)
- Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO)
- International Cancer Proteogenomic Consortium (ICPC)
- Human Tumor Atlas
- User generated data – Your data

CLINICAL PROTEOMIC TUMOR ANALYSIS CONSORTIUM
ICPC GLOBAL PARTNERSHIPS
A NCI-DoD-VA Proteogenomic Translational Initiative
T1
T2
T3
T4
NPTEL
33

(Refer Slide Time: 20:15)



Proteomics Data Commons Pilot

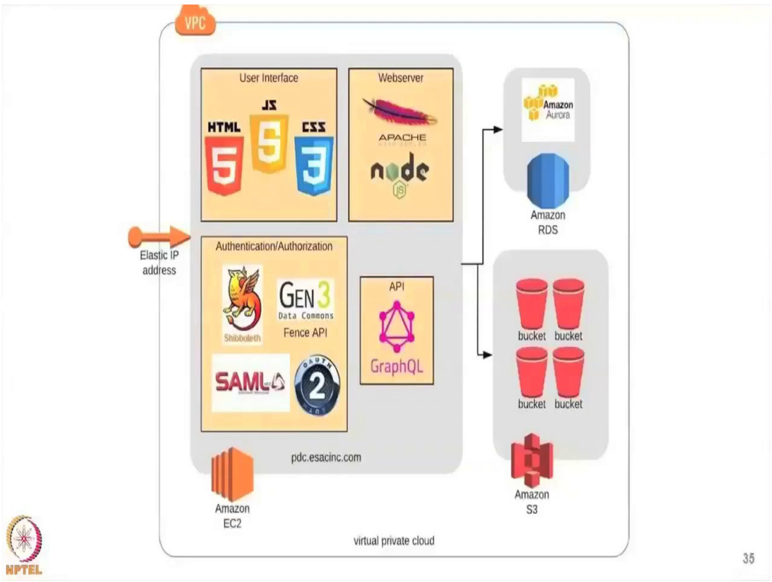
- Authentication and Authorization (A&A) Access Management
- Data Submission
- Data Portal
- Data Processing and Harmonization
- Processing Workspace
  - A pipeline infrastructure
  - Example DIA and DDA pipelines
- API access to data stored in the Data Portal

NCI Mandated Programs
Other Programs/Individual Labs
Only Submission/Retrieval
CRDC Authentication Authorization
Analysis
PDC Data Portal
Data Submission
Data Processing & Harmonization
Data Exploration & Visualization
PDC Standards Compliance
CEDAR
API to submit Data
API to make PDC Data available in the workspace
PDC Workspace
Chorus
Experiments & Projects
Upload Data
Collaborate
Analyze Data
API to get Genomic Data
NIH NATIONAL CANCER INSTITUTE Genomic Data Commons
API to get Proteomic Data
NCI Cloud Resources
NPTEL
32

So, what we did about 6 months ago, we started building a prototype. That is what we call proteomic data commons pilot or in other words we call MVP or a Minimally Viable Product. So, what would it is like I talked about a lot of things right now what we did is like build a minimal product and put it out there, it has some basic features that I am talking about does it will not do everything, but bits and pieces of everything I talked about and then see its take the feedback from the community and develop on those alright.

So, rather than I develop the whole product and I put it there like I said, and nobody is interested in that its waste of our effort. So, we will the portal like the commons will have a data submission system, a portal where you will see all the information and a workspace that I talked about where you can actually upload your data and run pipelines and then programmatically or the computationally access that information through the programs alright.

(Refer Slide Time: 21:21)



Student: What do you mean by air data processing? Like we are uploading our data and then you take the data processing in harmonization.

Yeah, data processing and harmonization is the common data analysis pipeline that we run all of the data. So, what happens like within CPTAC so, each of the cancer types is being the data is being generated by different groups and it is analyzed by different softwares, they are and different protocols. So, when we start putting out the data and there is the portal. So, it is we have one single consortium, but we are actually providing results from so, many different kinds of pilots.

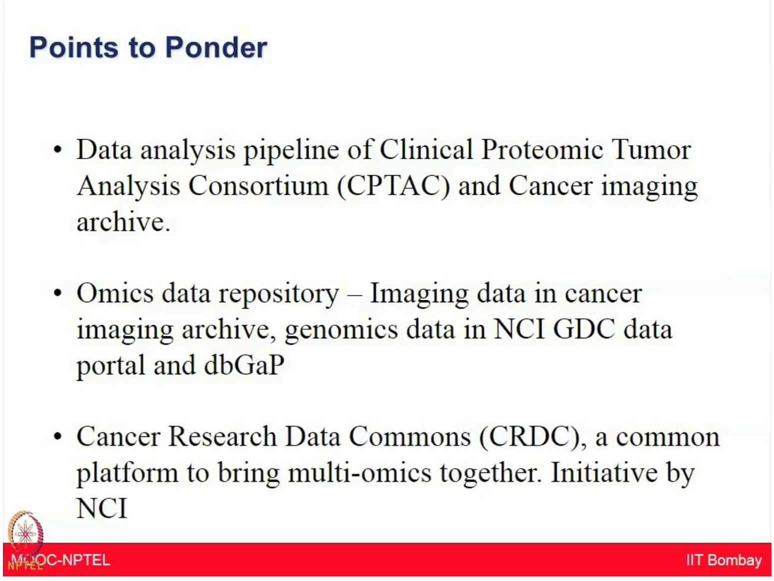
So, what the even when you go the genomic data commons so, we have both these portals what we try to do is this called data harmonization process. Any data that comes to these resources, we will use it once in common pipeline and all of them. So, it is not an ideal way, but it is one way to look at all the information on once. So, that is something what the PRIDE is doing what Peptide Atlas are doing right.

So, you are depositing all the data and one resource of your publication is done and they are running through the one common pipe line because when you have so much data if is not harmonized as for one single pipeline it is very difficult to make sense out of it. And, then that is that is what I call processing and harmonization and the processing workspace is basically you would be able to run the pipeline by yourself right. We run the pipeline as a starting step, but you want to change something some parameters ok, I am not very happy with this you want to change some parameters; you should be able to do that.

So, as an example we will have encyclopedia as a one DIA pipeline and also the pipe DDA pipeline that I described earlier. It is just showing the software architecture of our system and you do not have to actually spend a lot of time here, but just to give you an idea of this is called S3 this is a hard drive you can say the cloud hard drive where you put all your data. So, it is scalable. So, as your data is increasing, they will just make available how much ever you want. So, there is no limitation here. The same thing like you want more compute power with a click of a button you can add as many processors as you want and this thing it is called authorization.

The idea of this is basically because there are so, many portal I talked about genomic data commons, proteomic data commons, imaging and cloud pilots there are so many things are there, but the user has to remember all his user ID's and passwords and how do these commons actually talk to each other if it is not a single sign off. So, the purpose of this box is basically telling that we will have a single sign on. Once you sign into any of this resource so, you will be able to access the data from the other resources, seamlessly.

(Refer Slide Time: 24:33)



Points to Ponder

- Data analysis pipeline of Clinical Proteomic Tumor Analysis Consortium (CPTAC) and Cancer imaging archive.
- Omics data repository – Imaging data in cancer imaging archive, genomics data in NCI GDC data portal and dbGaP
- Cancer Research Data Commons (CRDC), a common platform to bring multi-omics together. Initiative by NCI

MOOC-NPTEL IIT Bombay

So, in conclusion today you have learnt about one of the large initiative from NCI about the CPTAC which has contributed the scientific knowledge for the cancer research immensely. You also learnt that to obtain omic data for a single patient, you need to search four different repositories. Hence, the NCI has taken initiative to make a common data portal so, that you do not have to look for variety of different places to search for data and all data could be commonly accessed from the Cancer Research Data Commons or CRDC.

You also got a glimpse of how CRDC has different features such as visualization, analysis, query and many more features. We also learnt that how cloud platform which is being freely available to everyone can be very useful in handling and analyzing big data or even metadata analysis. In the next lecture Dr. Ratna will continue his talk about large scale data sciences and inform us more about different publicly available portals.

Thank you.