

**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. Ratna Rajesh Thangudu**  
**Department of Bioscience and Bioengineering**  
**Indian Institute of Technology, Bombay**

**Lecture - 33**  
**Large-scale data Science-III**

Welcome to MOOC course on Introduction to Proteogenomics. We have here Dr. Ratna in last two lectures about large scale data sciences. Today, in the last lecture, he is going to continue sharing more information about Large scale data Sciences. Dr. Ratna we will talk about proteomic data commons – PDC and about its various dimensions such as ownership of the data, the quality management and life of the data.

He will also talk about FAIR principle for developing a proteomic data commons, fair stands for findable, accessible, interoperable and reusable - FAIR. It includes assigning data and patient, a UID unique identification which remains same across the world and hence reusable by the users. The importance soft management and up gradation of the repository with unique UID and versions with reason. So, to understand in depth, let us now welcome Dr. Ratna for his last lecture.

(Refer Slide Time: 01:34)



This is the minimally viable product that I talked about pdc Dot esacinc dot com. This is an alpha program. So, it is even before like I said. So, feel free to log in, I will show I

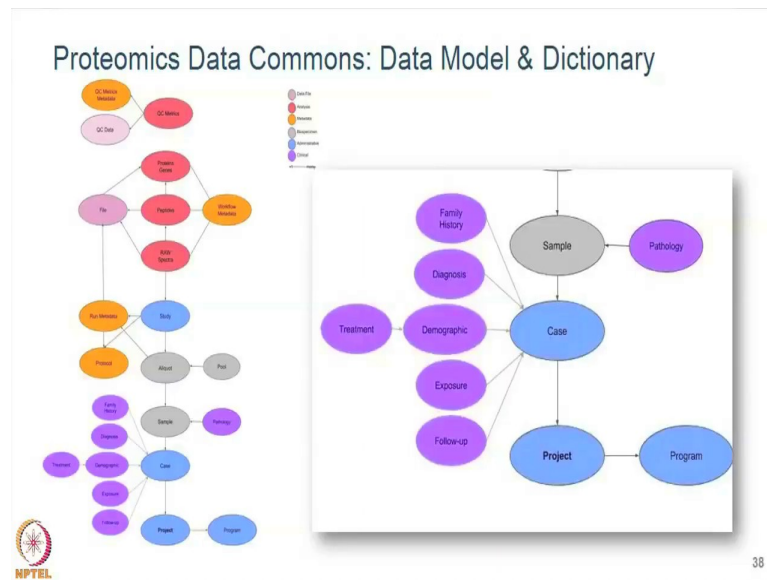
mean I will run you through that some screenshots of how the portal looks, but feel free to log in and see this is ok. Like I said this is the minimally viable product, all of the CPTAC data portal data sets that you see that colorectal cancer, ovarian and breast cancer data sets that are available where you just download the information. All that is actually harmonized to PDC, so you could actually explore all of those data sets here too much greater extent.

(Refer Slide Time: 02:22)



Just to give you an idea for what goes into the management, there is stewardship who owns the data. So, once you put it there it is done, am I the I mean is PDC the owner of the data or you still have it that is called stewardship. And then data governance what is the lifecycle of the data, who will what kind of policies that will guide the open access of the particular data. And then all other things about the standards and the processing and quality management, all those things are attached to the large scale programs all right.

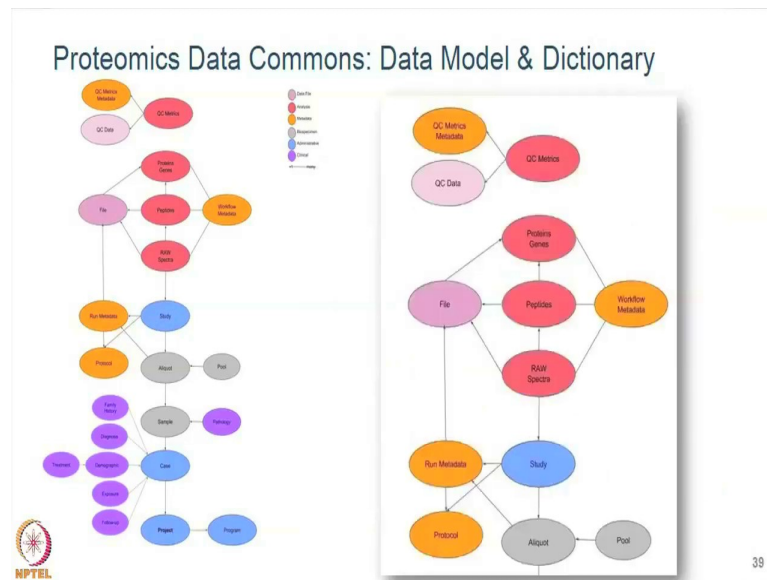
(Refer Slide Time: 02:55)



So, we have to represent the data in a certain way right, if I am getting data from so many different programs so many people are submitting how do I represent that in a common model. So, I need to have a conceptual model. So, this is our conceptual model. So, because all of this is cancer related data to begin with, so this is patient centric now talking about any model organisms and so on at this point. But even from model organisms the data can fit here it is pretty easy.

So, you have a program project a case is basically a patient or a donor who gives the tissue and then from that you get a sample. So, there is a lot of clinical data that is attached to a case all right. So, clinical information is attached to the patient.

(Refer Slide Time: 03:47)



And the next part I am showing this part right now. So, then you have an aliquot that is what actually goes into your mass spec all right. So, that from there you generate, so you group a bunch of samples run as run it is an experimental study. And then you have all of the run metadata that is nothing, but that you are experimental design, where we are capturing it and then you generate the raw files and then we run the workflow. So, then I run the workflow, it generates other informations.

So, every piece of information that the software generates we index it. So, it is captured in the model. So, when I ask a question about which proteins are expressed in this particular aliquot or a sample, I have an answer for it right that is because we have a model. So, we are trying we are working hard to actually define this. We try to fit all of the CPTAC data in this model and its working quite nice, but there are a lot of expectations happen right. So, if there is something that comes up that actually needs change in that are model that we have to do it, but for now it is working all right.

(Refer Slide Time: 04:58)

### Proteomics Data Commons: Data Model & Dictionary

**NATIONAL CANCER INSTITUTE**  
Proteomic Data Commons

Data Dictionary Viewer  
A small description about the dictionary can come here

Administrative	
date	The collection of all data related to a specific subject in the context of a specific project.
program	A broad framework of goals to be achieved. (NCI C32647)
project	Any specifically defined piece of work that is undertaken or attempted to meet a single requirement. (NCI C4782)
study	A detailed examination, analysis, or critical inspection of a subject designed to discover facts about it. (NCI C6353)

Biological	
gene	A functional unit of heredity which occupies a specific position on a particular chromosome and serves as the template for the synthesis of a specific protein.

Biopspecimen	
aliquot	Pertaining to a portion of the whole, any one of two or more samples of something of the same volume or weight.
pool	Any aliquot where multiple aliquots are combined to produce a reference. Sample pooling is commonly used for diagnostic testing.
sample	Any material taken from a biological entity for testing, diagnosis, propagation, treatment or research purposes including but not limited to cellular molecules, cells, tissues, organs, body fluids, embryos, and body excretory products.

40

So, then there is this data dictionary. So, every node in this graph is basically we have a description of it. So, you can go there and try to understand what it is. And each of these turns are actually we apply standards to that. So, for example, here diagnosis and demography. So, these are all clinical terms there. So, are we using certain standards for example, ICD codes to define the disease or terms, all right.

(Refer Slide Time: 05:28)

Findable Accessible Interoperable Reusable

NPTTEL

So, then I will briefly talk about the FAIR principles. So, when you are developing a data commons the guiding principles for such a kind of effort are called FAIR – Findable, Accessible, Interoperable and Reusable.

(Refer Slide Time: 05:48)

**F**indable

(meta)data are assigned a globally unique and persistent identifier

Data are described with rich metadata

Metadata clearly and explicitly include the identifier of the data it describes

(meta)data are registered or indexed in a searchable resource

```
{
  "id": "1194202f-0786-421e-a86d-72288e77c269",
  "type": "object",
  "size": 512,
  "sha1": "c3445a5d30f4667eb95b73ea9d7bad3dc4",
  "sha256": "20b599f90f5f98e9e120ba6d3665f753c662721f368"
}
```

So, what is findable? So, like I said every piece of data that comes in whether a patient or the file. So, we will assign a unique identify to that because we call it the UID a global id and then we attach lot of metadata to it. So, it is unique in the system, so that nothing will change ever. So, even if there is a change, we will version them that. So, a why it changed and how it changed, so that you can always refer back to that particular patient successfully.

And then there is at the file level, most of the times when you start downloading or uploading files, they get corrupted right. So, what you start uploading and what is received at the end, you do not know if they are the same or not. So, you think you uploaded some file and but it is not; it is not fully uploaded it get gets corrupted.

The same thing when you download and you do not realize that there is some error occur somewhere and you start processing those files and it will your software will throw error, but it is very difficult for you to understand why the software is throwing error right in the first place. So, that is why we record a lot of metadata that is called MD 5 or SHA is these are called check some values it says small datum that is attached to a digital object.

So, you generate that locally on your computer. You upload the file and the receiver you will give him this MD 5 value. So, on the receiver on his computer, he will generate the same value and compare this value the what you are giving and what he received is the same or not the basically it confirms that the data is end to end transfer completely. So, all that information will be captured as a metadata. So, basically you can find any file or any entity in this model by a particular unique id and with the attached metadata.

(Refer Slide Time: 07:34)

**A** Accessible

(meta)data are retrievable by their ID using standardized communication protocol

- The protocol is open, free, and universally implementable
- The protocol allows for an authentication and authorization procedure

metadata remain accessible even if data aren't

FAIR data-restricted access (D)

FAIR data-Open Access (E)

FAIR data-Open Access/Functionally Linked (F)

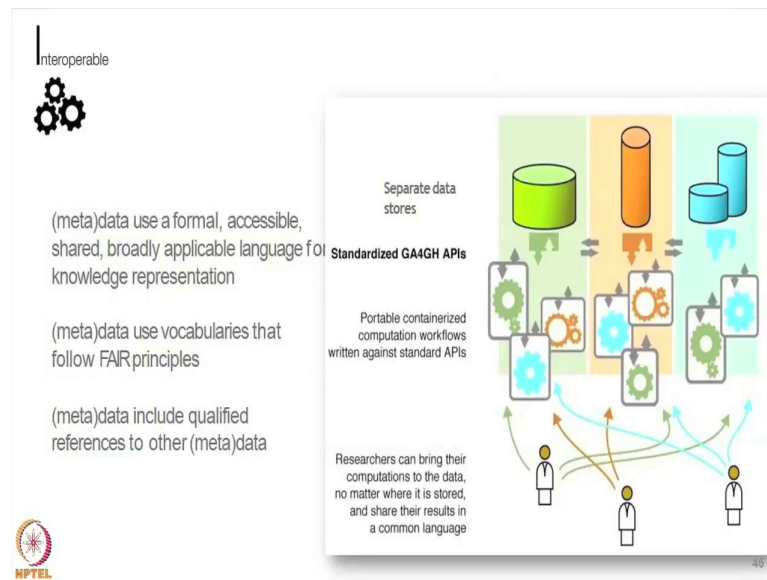
CIDR Authentication Authorization

NPTEL 45

And then accessibility, so who owns the data and what is the life cycle of the data. So, when you first generate the data and put it in a system like in a PDC, so you still own the data until you make it public right, so that is the user has the authority to tell ok, this, this is the time I want to make it public or my program requires me to publish in a particular journal, then I then I will have to make it public.

So, and some of it can be protected, especially in genomics we know the all the germline mutations are protected and you cannot just download. So, there is a data access committee and you have to go through that whole cycle to actually get the information.

(Refer Slide Time: 08:16)



And then interoperable. So, interoperability is basically about the controlled vocabularies, how we are defining the terms like if you are calling trypsin or you calling trypsin all the time the same way. So, it is confusing sometimes. So, what people do is the control vocabularies are standards, they will assign ids to those terms, so that you can just use the id instead of the word.

So, programmatically we will access that information when you use the id, we will get the information ok, this is called trypsin and then that is called metadata and. So, with that kind of information, these are for example, these are three different resources GDC, proteomic data commons and imaging data commons. If we are all using the same kind of terminology to describe the common terms across these resources, any tool I can use without much hassle.



(Refer Slide Time: 09:09)

Reusable

(meta)data are richly described with a plurality of accurate and relevant attributes

(meta)data are released with a clear and accessible data usage license

(meta)data are associated with detailed provenance

(meta)data meet domain-relevant community standards

HUPO  
Human Proteome Organization

NIH NATIONAL CANCER INSTITUTE  
Enterprise Vocabulary Services

Proteomics Standards Initiative

CDSR

Name	Controlled in	Description	Required	Schema
id	query	A unique identifier of the tool, exposed to this registry, for example, <code>123456</code> .	No	or string
registry	query	The image registry that contains the image.	No	or string
organization	query	The organization in the registry that published the image.	No	or string
name	query	The name of the image.	No	or string
toolname	query	The name of the tool.	No	or string
description	query	The description of the tool.	No	or string
author	query	The author of the tool (FOOD a thought cohort, are we assuming that the author of the CIVL, and the image are the same?)	No	or string
offset	query	Start index of paging. Pagination results can be based on numbers or other values chosen by the registry implementer for example, GSK values). If this exceeds the current result set return an empty set. If not specified in the request, this will start at the beginning of the results.	No	or string
limit	query	Amount of records to return in a given page.	No	or integer (int)

MPTEL

47

And finally, reusable. So, reusability comes with the standards like the I mention. So, for PDC we use a lot of standards from the proteomic standards initiative. So, PSI is a special body within the HUPO that formulates guidelines how to how do you represent the data and also both the formats of the files and also the controlled vocabularies that needs to be used.

It is not as mature as in genomics, but its slowly getting there, I think in the next several years. So, we will probably have more structure terms that people will start using. So, at this time, we are I mean the that itself actually has about 100 plus terms that you can use, but at this time we could only map about 10 of them to the metadata that we are collecting.

(Refer Slide Time: 10:00)

Why do we need standards?



- 01/04/2008
- 1/4/2008
- 1st April 2008
- (or January 4, 2008)
- 01042008
- 01.04.08

ISO 8601:2004  
Data elements and interchange formats – Information interchange – Representation of dates and times

2008-04-01.

MPTTEL 48

Just a general example of why we need standards, I just want to show you this. So, that plug points, yesterday came I came from US, so I do not have a converter, I do not need a converter I just need an adapter. So, I was looking everywhere I did not find one first ok. When I go back to hotel like I am just charging fully that my computer is now fully charged but anyways, so all these plugs they are actually based on some standards they are not random right.

Each country has different kind of standards and they are using. So, in such kind of situation, maybe you can have a converter or are you can have an adapter as a solution. So, it is possible because you know those people are using standards. And here on the right and we showing an example of how a date is being represented, you can write it in any way right. So, probably it will make sense for some of you, but when you give it to the computer or it will get confused.

So, then we come up with there you come up with the standards right. So, now, we have some converter plugs and then we have ISO code you have to represent the date in certain way. So, when you are submitting data to for example, PDC or GDC we will tell if you are putting a date it has to be like this right that is an example, simple example of standards all right.

(Refer Slide Time: 11:15)



So, why do we need standards, because bioinformaticians who are into bioinformatics they would actually understand this much easier. So, it is very before the same that is sitting in different, different forms it is a nightmare for that. So, we just have to I mean the idea is like you just you want to compare the protein quantitation data with the gene expression data that is you are receiving from GDC thing is like you have two files or why do not we just compare.

But if they are in different formats if you are calling the different gene names and you have to write parsers and programs and there is so much effort, so you waste all your time all right.

(Refer Slide Time: 11:49)

Clinical Data Standards out there..

Abbreviation	Full form
eMERGE	Electronic Medical Records and Genomics
EMR	Electronic Medical Record
DE	Data Element
UMLS	Unified Medical Language System
SNOMED CT	Systemized Nomenclature of Medicine-Clinical Terms
NCI	National Cancer Institute
caDSR	Cancer Data Standards Registry and Repository
SDTM	Study Data Tabulation Model
HL7	Health Level 7 ( <a href="http://www.hl7.org">http://www.hl7.org</a> )
CDISC	Clinical Data Interchange Standards Consortium ( <a href="http://cdisc.org">http://cdisc.org</a> )
IHE	Integrating the Healthcare Enterprise ( <a href="http://www.ihe.net">http://www.ihe.net</a> )

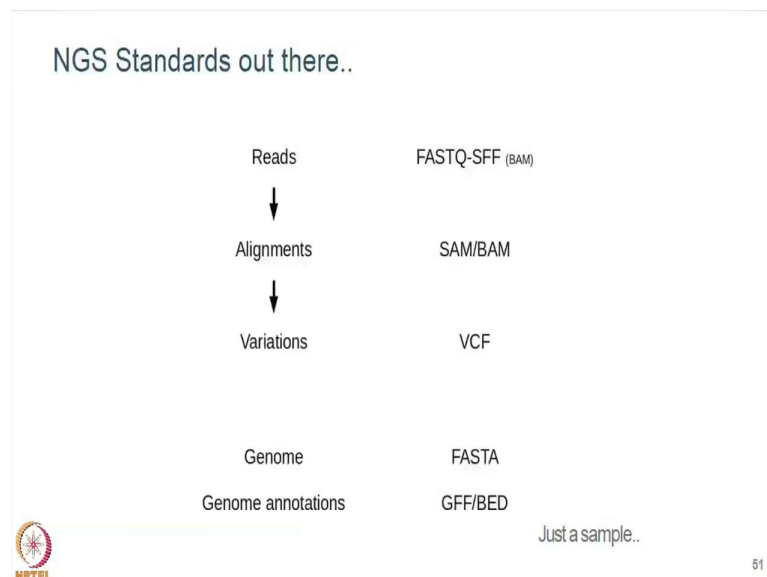
Just a sample..



50

So, just I am I have put something here to give you an idea of different resources these are this is not very extensive, it is just a sample. For example, we use caDSR the Cancer Data Standards Registry and Repository for representing all of the clinical data information in the PDC and GDC. So, every term has a I have mentioned in the caDSR.

(Refer Slide Time: 12:18)




And then we know for the next generation sequencing we all know there are some very well adapted formats FASTQ-BAMs and VCF and so on right.

(Refer Slide Time: 12:32)

Proteomic Standards out there..

- Minimum information (MIAPE) specifications: Format-independent specification of minimum information guidelines.
- Formats: Usually an XML schema (but also tab-delimited files) capable of representing the relevant Minimum Information, plus additional detailed data for the domain.
- Controlled vocabularies: Usually an OBO-style hierarchical controlled vocabulary precisely defining the metadata that are encoded in the formats.



52

So, proteomic standards, so I like a mention there are a few already there. So, there is something called probably some of you might have heard about MIAPE. So, minimum information about a microarray experiment that was came like several years ago that basically tells it is a recommendation it is not as it is not forced on you, it is just a recommendation when you are submitting some microarray data to some GEO dbGaP.



So, this is the minimum information you should provide. So, because it is just a recommendation or a guideline, if you go to geo database that is gene expression omnibus, it is a mess. There is so much data out there, but if you want to use it there are actually papers meta analysis people just did correcting the data in geo that itself is a scientific article right, so that is not worth it.

So, the same way I mean in proteomics also they started off with ok, we do not want to force anybody with standard set up to begin with what we will tell these are the minimal guidelines we have to fall, so that is MIAPE - minimum information about proteomic experiment right all right. So, then we already have a lot of formats, I will show you in the next screen and also we have some control vocabularies from the proteomics standards initiative from yes sir, all right.

(Refer Slide Time: 13:56)

Proteomic Standards out there..

MS data	• mzML
Identification	• mzIdentML
Quantitation	• mzQuantML
Final Results	• mzTab
SRM	• TraML

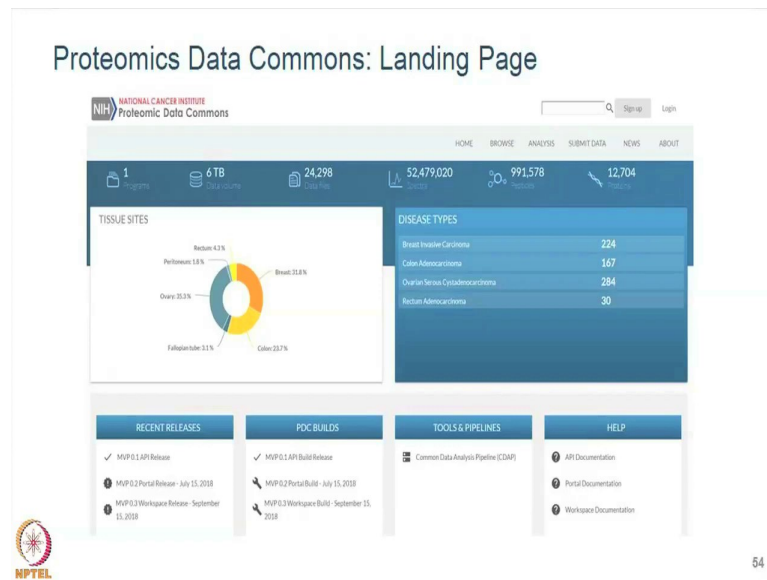
 Human Proteome Organization  
 Proteomics Standards Initiative

53

So, these are some standards like the deprecation mzML, mzIdentML. So, the lot of things came out, but the widely adapted one is the mzML so far because each piece of software that you are running each pipeline, it generates a different kind of output. So, it is thus the field is not as mature as in genomics yet, so but we are slowly going there. So, the expectation is that going forward maybe in the next several years, all these will be adopted.

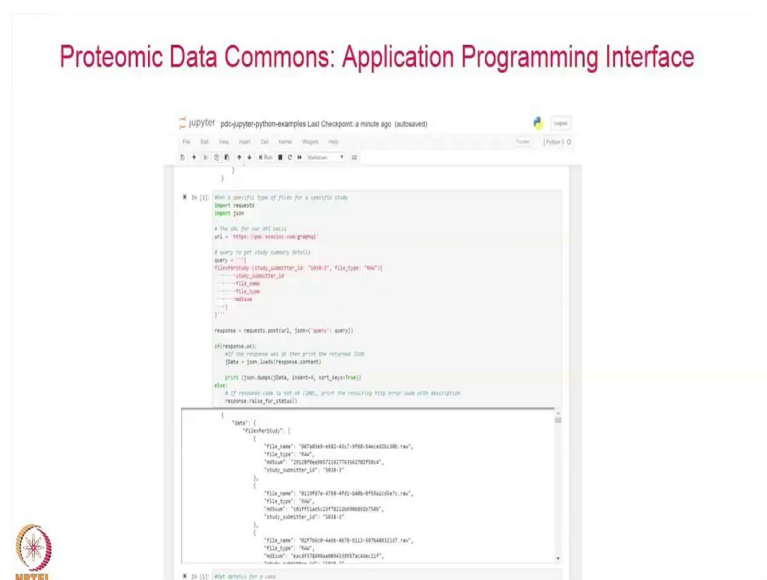
So, if they adopt like whether it is PRIDE or peptide atlas or proteomic data commons, if you have the same format of the file, you can easily compare two things is much easier ok.

(Refer Slide Time: 14:40)



So, with all that unconsideration we did build that minimally viable product that I mentioned earlier. So, this you can see there is only one program right now that is the CPTAC, we started off with that. We have about this much six terabytes of data and these many proteins and peptide identify, so lot of summary information on the home page.

(Refer Slide Time: 15:03)



We also have application programming interface basically whatever you do on the UI, you should be able to do programmatically also if you are efficient in the program.

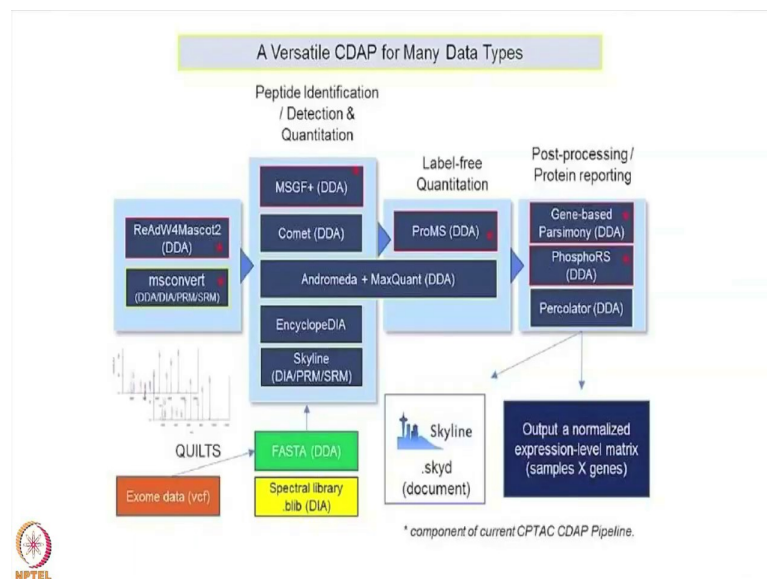
(Refer Slide Time: 15:16)

### Proteomics Data Commons: Workspace

ID	STUDY NAME	OWNER	PROGRAM	PROJECT	FILES	MODIFIED
121	S039 2 Prospective_Breast_RI_Phosphop...	Ratna Thangudu	Clinical Proteomic...	CPSC-Confirmatory	221	Aug 3, 2018
126	S039 1 Prospective_Breast_RI_Phospho...	Ratna Thangudu	Clinical Proteomic...	CPSC-Confirmatory	425	Aug 3, 2018
119	S038 3 Prospective_Ovarian_PANL_Phosp...	Ratna Thangudu	Clinical Proteomic...	CPSC-Confirmatory	144	Aug 3, 2018
118	S038 2 Prospective_Ovarian_PANL_Phosp...	Ratna Thangudu	Clinical Proteomic...	CPSC-Confirmatory	288	Aug 3, 2018
117	S037 3 Prospective_Color_PANL_Phosph...	Ratna Thangudu	Clinical Proteomic...	CPSC-Confirmatory	132	Aug 3, 2018
116	S037 2 Prospective_Color_PANL_Phosp...	Ratna Thangudu	Clinical Proteomic...	CPSC-Confirmatory	204	Aug 3, 2018
115	S028 4 TCGA_Ovarian_PANL_Phosphoproteo...					
114	S028 3 TCGA_Ovarian_PANL...					
113	S028 2 TCGA_Ovarian_PANL_P...					
112	S028 1 TCGA_Ovarian_PANL_G...					
111	S016 1 TCGA_Color_Cancer_P...					
109	S037 1 Prospective_Color_VG...					
110	S038 1 Prospective_Ovarian...					

And this is the workspace I was talking about I will show all this.

(Refer Slide Time: 15:18)



And this is the kind of common data analysis pipeline that we would implement within the system all right.



(Refer Slide Time: 15:32)

### Use Case : Proteogenomic Integration

- Identify variant protein sequences corresponding to somatic mutations and to evaluate the relationship between mutation frequency and variant protein expression.
- Determine how copy number variation translates into protein expression differences.
- Evaluate the impact of genomic features on the status of signaling networks through direct analysis of phosphoprotein intermediates
- Derive preliminary associations with clinical characteristics, such as platinum resistance in ovarian cancer.

60

So, before I am I just want to touch base on the proteomic integration because that is what it is CPTAC is about on the proteomic data commons is about.

(Refer Slide Time: 15:42)

### Use Case : Proteogenomic Integration

1) Sequence-centric Proteogenomics: describes aspects of sequence-centric proteogenomics and the combined use of genomic and proteomic data to augment gene or protein annotation

61

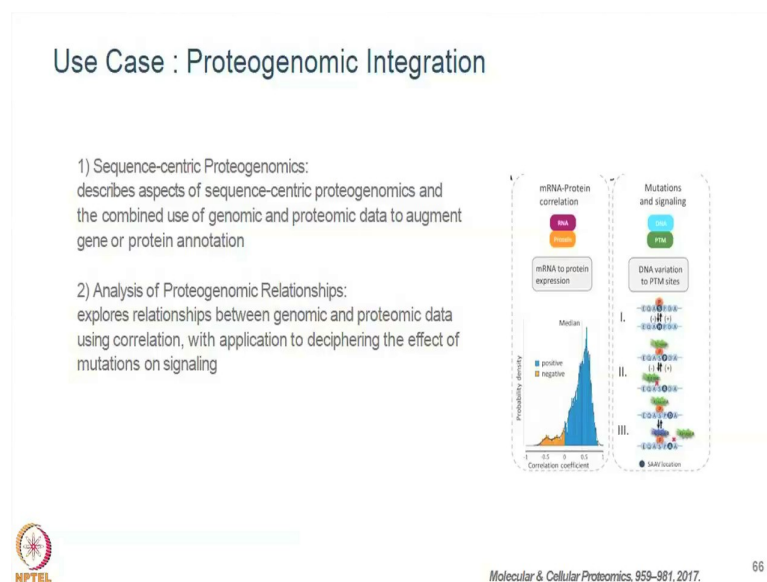
*Molecular & Cellular Proteomics, 959-981, 2017.*

So, at the system level when two systems are in interacting like a PDC and GDC, what do we need all right. So, there are some sequence centric Proteo-genomic approaches. So, that is where if because all of this information is patient derived patient centric. So, there is genomic information available for the patient on the GDC side right. So, you

want to use that information to run your such as a against right. So, how would you get that information?

So, in the model that we are developing or we are closely working with the a framework team from the NCI and also the genomic data commons and the cloud pilots, we are trying to figure out how to bring that information together, so that when I say a patient like when the PDC when I look for a particular patient, it will automatically tell, hey, there is some genomic data available for this patient at this resource you want it right. So, that way you can get that information and put as input for your proteogenomic experiment sorry data base search.

(Refer Slide Time: 16:51)



And then the other thing, so that is at the very high level even before you start your pipeline right. So, then if you actually generate the quantitation data already, so it is in the gene matrix. And then on the GDC side they already have the FPKM are became information gene expression with they have the read files they have the genomic variants.

So, you bring them and I start correlating them right, so that means, here you are actually comparing those results from two harmonization pipelines you are not comparing the raw data you are not doing anything there. You are using the information that came out of the pipelines and just comparing them. So, the R examples that you are trying their best trying to basically do that.

(Refer Slide Time: 17:34)

### Use Case : Proteogenomic Integration

- 1) Sequence-centric Proteogenomics:  
describes aspects of sequence-centric proteogenomics and the combined use of genomic and proteomic data to augment gene or protein annotation
- 2) Analysis of Proteogenomic Relationships:  
explores relationships between genomic and proteomic data using correlation, with application to deciphering the effect of mutations on signaling
- 3) Data Visualization:  
integrate mass spectrometry data with the genome.

Molecular & Cellular Proteomics, 959-981, 2017. 67

And finally, can we see all the peptides that we identified on the genome browser. So, you can download the information and you can generate the bed file and you can upload and do all those things. But we are trying to do that for you. For any given data set it will be automatically available, so these are the some visions I will show some data what is already there.

(Refer Slide Time: 17:58)

### Use Case : Proteogenomic Integration

Find all the **Program Project Study** in **PDC GDC IDC** that have **Proteomic Genomic Imaging** data

Get **RNA-Seq BAM file**, **Variant file**, **RNA-Seq Junction file** for a **Cases Sample's Aliquots**

Get **Copy Number Data**, **Expression Data** for a **Project Study**

*I know there is genomic information available for the PDC proteomic study I am interested in. How can I seamlessly integrate the somatic/germline variant, RNA-seq predicted junctions and fusion, etc to create a custom sample specific database to search against?*

MPTTEL

68

Student: Sir, here when you seen like we cannot upload our data in this repository am I right? So, which is whatever reason presented in CPTAC data whatever CPTAC data is

there, now we can see that data inside. And we if you want perform some kind of analysis all that is it right or not?

So, the idea is PDC we will do some basic analysis for you. So, when you go there, you already have all these reports generated. But if you want to change something, the personal workspace I talked about earlier, so you can do all those things in your own workspace without affecting the public site right. So, whatever we provide is actually for everyone.

Student: From that data only, not on the data which I am working on.

So, the high level goal is we have all this information on the PDC data portal which does not have a login you go and just go there and analyze, but you have some data that you want to actually see correlate with what is already there. So, you want analyze your data along side of the data its already there you should be able to do it in the work space. So, you load all your data provide all the metadata. So, that there will be tools available for you to do that kind of a process all right.

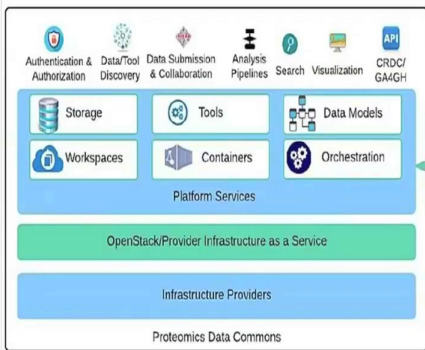
So, this is one use case we actually got somebody asked us I know there is the genomic information available for the PDC proteomics study and how can I actually seamlessly integrate all this information. So, use cases I was talking about. So, the way we implement at the systems level is like find all the projects in PDC that have genomic data that is a very easy way to ask, but we add this find all the programs that have in GDC that have proteomic data.

And you can ask that question whichever way you want right in different combinations, the system should be able to do that, so that is where we are trying to relate to. So, some more examples of the same thing all right.

(Refer Slide Time: 20:11)

## Summary

- Unsilo mass spectrometry data and bring compute and tools to data instead of the other way round
- A common vocabulary that paves the way for accurate and reliable communication among nodes - Semantic interoperability
- Improve reproducibility of proteomics data by making it possible to share tools and workflows with the release of the data.



The diagram illustrates the architecture of Proteomics Data Commons. It is structured into layers. At the top, there are several functional modules: Authentication & Authorization, Data Tool Discovery, Data Submission & Collaboration, Analysis Pipelines, Search, Visualization, and CRDC/GA4GH API. Below these are six core services: Storage, Workspaces, Tools, Containers, Data Models, and Orchestration. These services are grouped under 'Platform Services'. Below Platform Services is a layer for 'OpenStack/Provider Infrastructure as a Service', which sits on top of 'Infrastructure Providers'. The entire stack is supported by 'Proteomics Data Commons'.


69

So, I will just summarize here. So, we right now the proteomic data commons is in build phase. So, like I said the 6 months ago we build the MVP we started building the MVP and we release that on at HUPO meeting in October. So, we got some feedback, but in terms of system like I talked about all of these things, but basically will have storage, workspaces tools and containers was orchestration.

(Refer Slide Time: 20:47)

## What can you do?

- Planning early
- You have your data ... now what?
- You have no data ... now what?
- Where will the data live?
- Interested in finding openly shared data?

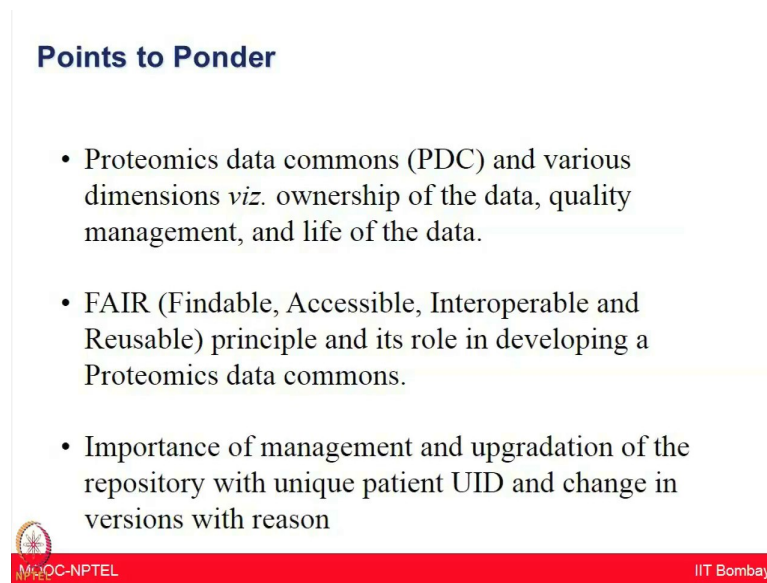


70

So, you plan early what can you do with this, this information right now. So, if you have data you have some ideas now, I think I convinced you to what to do. If you have no data

that is I guess some people came to me and said like we are interested, but I do not have any data, where do I start. So, there is much data out there you can start looking at that. And where the data will live? At least in the PDC side we say it is in the cloud. And if you do not make it public, you are the owner of the data.

(Refer Slide Time: 21:23)



**Points to Ponder**

- Proteomics data commons (PDC) and various dimensions *viz.* ownership of the data, quality management, and life of the data.
- FAIR (Findable, Accessible, Interoperable and Reusable) principle and its role in developing a Proteomics data commons.
- Importance of management and upgradation of the repository with unique patient UID and change in versions with reason

MOOC-NPTEL IIT Bombay

So, I hope today you have learned that the PDC data portal consists of all the omics data on a single platform with UID given to each data set or patient, which will remain same across the world. It enables users across the globe to access and reuse the data. You also learned about Proteomics Standard Initiatives or PSI and importance of such initiatives.

You also got glimpse that how difficult is the same data exists in different formats, hence a converter or a standard notion could play a major role in developing repositories which are accessible to all. We also learned about proteomic standards being used currently such as mzML, mzIdentML or mzQuantML and others. So, in the next lecture, we will shift topics we will talk about data independent acquisition and swath at least by another guest speaker.

Thank you.