

Interactomics: Basics and Applications
Prof. Sanjeeva Srivastava
Prof. Deeptarup Biswas
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 58
Data Repositories and Databases

Hello students. As we are you know reaching close to finish this course, you have already got a glimpse of variety of high throughput technology is available. For example, we talked about microarrays, we talked about label free biosensors, some exposure of next generation sequencing and mass spectrometer. All of this are contributing towards big data generation. And many times you might be thinking that until unless you have access to a lab, which has all these kind of you know high end gadgets, you cannot do your project, you cannot do your research in this area without availability of these equipments.

Partially it is true, if you want to generate some new data, on a kind of sample which nobody has tried yet then you have to run your sample and analyse data. But what is very important here in this case, nowadays there are many public repositories are available and those public repositories are able to provide all the raw data file, to the entire scientific community.

And that is one of the big you know shift and throughout in the world, that all the generals are now making it mandatory to upload the raw data files. And eventually after publication the data becomes in public repositories. So, you can access lot of NGAS data, you can access lot of mass spec data, micro data all of this in the raw format.

So, you always need not to have perform your own experiment and generate data, with your own equipments. You know you can always start with the public repositories, try to obtain the data file process them in very uniform manner; look at what is the hypothesis which you want to test out and try to build that particular hypothesis.

And see whether these data set are able to answer that question or not. So, this kind of you know, I must say that availability of various public repositories. And the omics data set

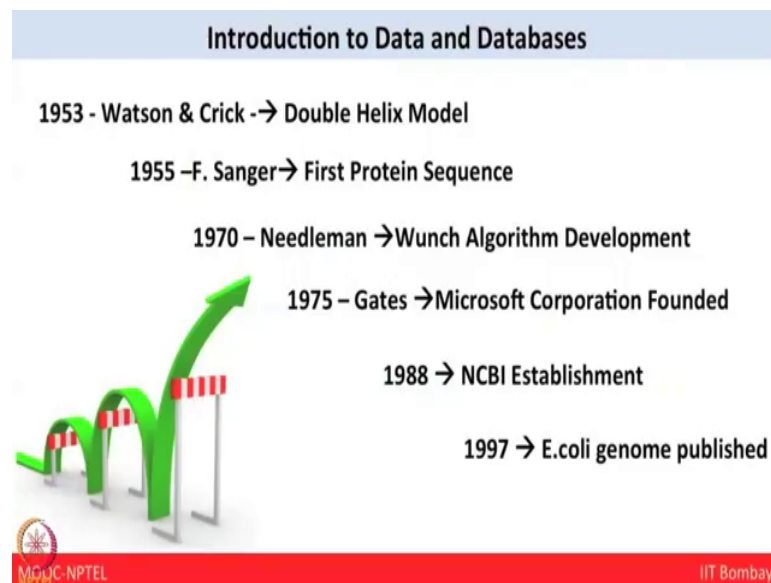
availability has really shifted the gear of the entire omics community. Where now the set of people who are still doing wet lab experiments and generating data, but there are set of scientists who are starting to do a lot of core bioinformatics and proteogenomic analysis to start putting the things together.

And it start utilizing the datasets looking at in a very very uniform manner. So, the question arises what are these resources from where you can obtain these dataset right? And information for them is very very important; because once you know that you know from which places you can go and you can obtain the data set then you can start doing many of these interesting analysis and processing data and visualization yourself.

So, in today's lecture in hands on session Deeptarup Biswas, the research scholar in my proteomics lab at IIT Bombay; will take you through different portals from where these data can be extracted for further analysis. So, let us have today's lecture and demo session.

So, first we all know that even before 10 years even before 20 years, the amount of data generated was not that much huge that is getting generated today. So, in 1953 Watson and Crick was the first one, who proposed the double helix model for DNA based on X ray.

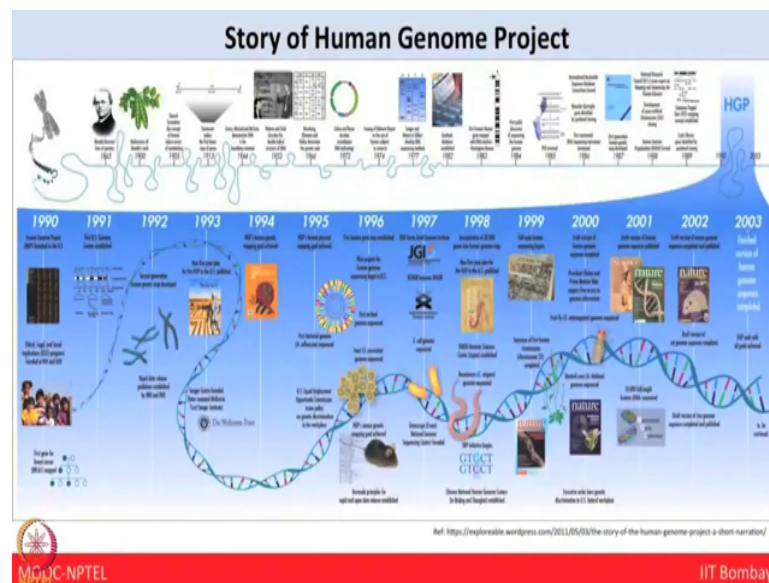
(Refer Slide Time: 03:40)



And the data and the generation of the data on the basis of that, can be taken as a first milestone of data generation. After that in 1955 the sequence of the first protein was to be analysed, which was bovine insulin. Followed by 1970, the first algorithm that is Needleman Wunch algorithm was came into play; in 1975 a major breakthrough happen when Microsoft corporation was founded by Bill Gates and Allen.

Further in 1988 the National Center of Biotechnology Information that is the NCBI which we all are familiar with was established.

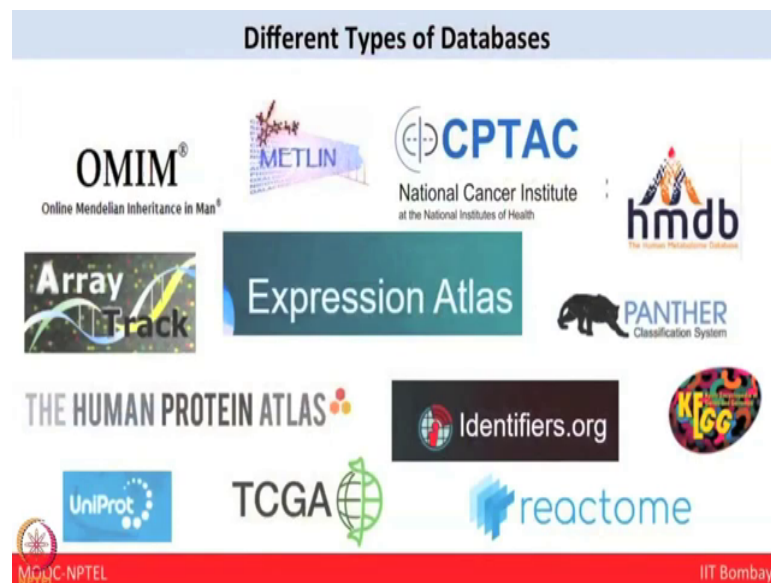
(Refer Slide Time: 04:31)



In 1970 in 1977 the at the Genome of E coli was published; after that we all know the biggest breakthrough that happened with the publication of Human Genome Project around 2004. After the publication of the Human Genome Project, we never moved back and the amount of data generated was huge and still now, we are generating huge amount of data.

So, now, let us talk about different databases that are available. So, if I give a very simple example that, database is a collection of data; any data that can be protein proteomics data, genomics data, that can be metabolomics data.

(Refer Slide Time: 05:08)




And if we are talking about parts, if we are talking about other backgrounds then there are many databases that are available. Like in terms of astronomy, in terms of ecology, in terms of cosmic sciences the databases are available. So, what is the main role of these databases?

(Refer Slide Time: 05:35)

Role of Databases

- Data Availability
- Data Systemization
- Data Analysis



MOOC NPTEL IIT Bombay

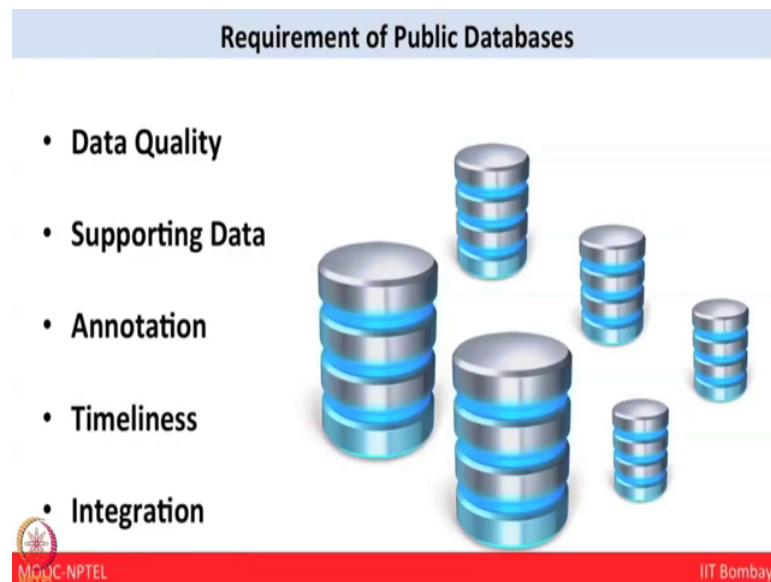
So, the main important role of the databases are availability of biological data systemization of the data and analysis of the computer biological data. So, as we know with the advancement of the technology and experimental strategy omics experiments and omics platform has really developed, in such a way that the amount of data we are generating is huge.

On the basis of that, from a long time database has been segregated into different forms and different levels on the basis of the data types, data sources, different data design and data bases. On the basis of data type, there are a number of databases that are present; like genomic databases, microarray databases.

Where you will get the microarray processed and pre processed files pathway databases; like example KEGG reactome disease databases maybe OMIM and this kind of databases, which

are already available on the basis of data type. So, now, what are the principal requirement of a public databases?

(Refer Slide Time: 06:38)

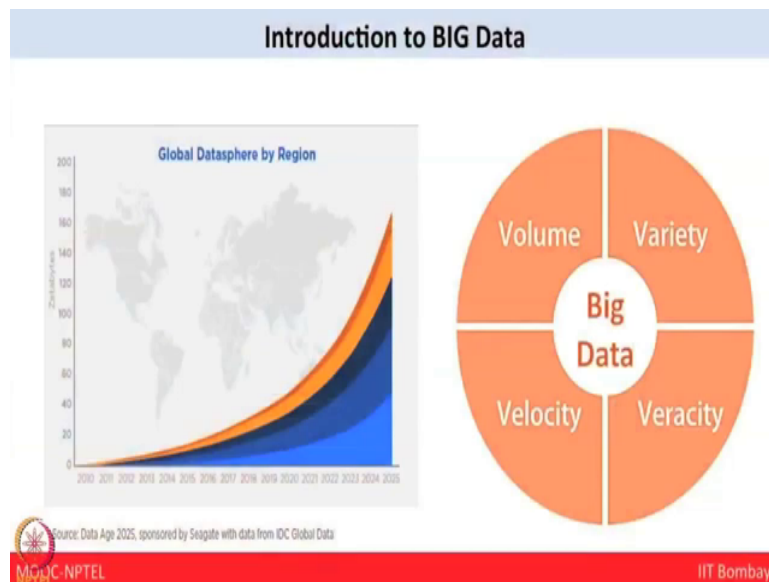


So, first data quality; the data should be curated should be of high quality data. Next is supporting data, the database users will need to examine the primary experimental data, either in the database itself or by following cross reference back to the network accessible laboratory database. Third is the deep annotation supporting and ancillary information should be attached to each basic data object in the database.

So, next is the time timeliness; that means, the data which you are putting into a database, should be available on an internet accessible server within or after few days of publication or submission. And finally, integration; each data object in the database should be cross referred to the representation of the same or related biological entities in other databases. The amount

of data we are generating now is huge and from where we are using a very frequent term about the data is like big data.

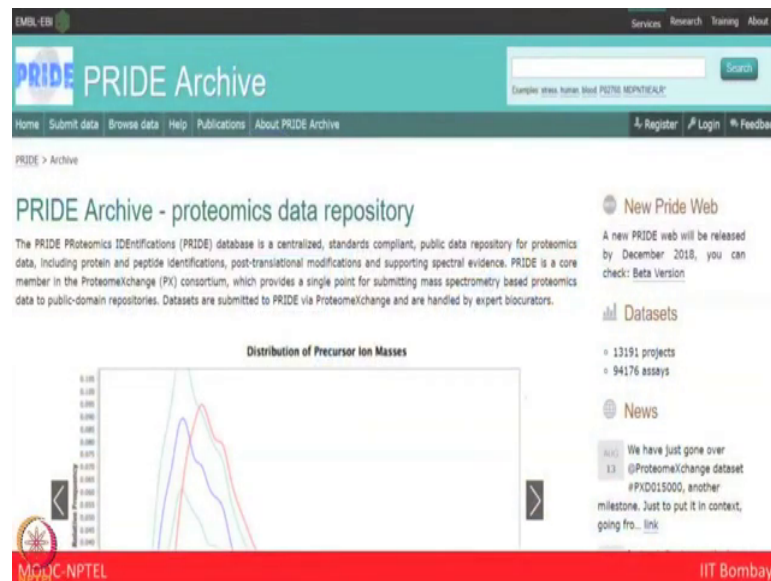
(Refer Slide Time: 07:45)



So, from where the word big data comes from. So, term has been used since 1990's which was given by John Massey. So, big data is usually includes data set, with size beyond the ability of a commonly used software tools to capture curate manage and process the data which are tolerable elapsed type.

So, the big data comes with four v; that means, volume, variety, velocity and veracity. So, now, let us come back to the databases that are available for proteomics and genomics. So, let us take an example a very popular database that is available is PRIDE that is proteomics identification database.

(Refer Slide Time: 08:27)



Which is a public user populated which a public database and user populated proteomic data repository. The repository contains the data generated by mass spectrometry proteomics experiment, which includes rows spectral data, peptides, protein identification and associated statistics; even different parameter files that are used for generating the data or processing the data.

PRIDE suppose the submission of data generated from many platforms in specific data format known as PRIDE xml file formats.

(Refer Slide Time: 08:57)

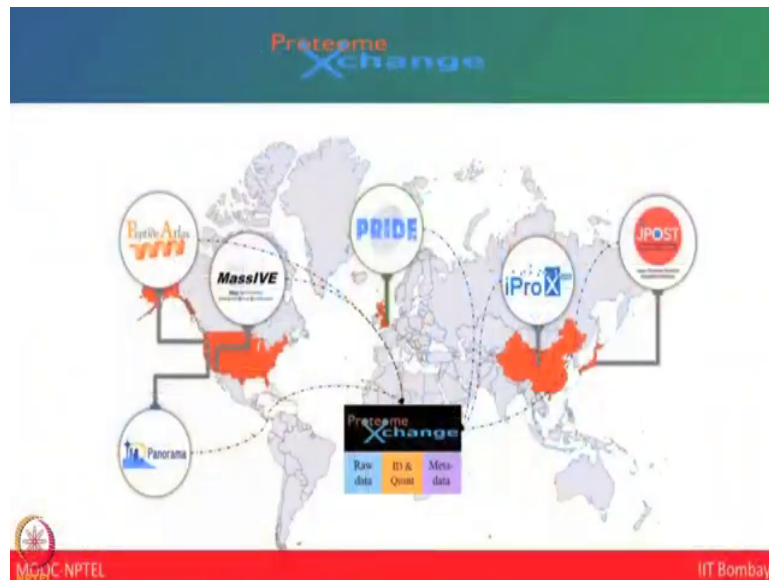


So, let us come to another database that is Peptide Atlas. The long term goal of the peptide atlas project is a full annotation of eukaryotic genome through a thorough validation of expressed protein. Different related databases that are available in the website like SRM atlas, which contains; all the data base all the data sets of different targeted approaches of mass spectrometry faecal. It is also same which contain different data sets of targeted proteomics. Phospho pep the name itself suggests that this part of the database contain the phosphoproteomics data, uni pep and ms spec line.

So, this databases contain different levels of proteomic data information, which can be accessed and downloaded. Let me show you a glimpse about proteome exchange which is a repository that is a collection of different databases. Like peptide atlas, massIVE PRIDE,

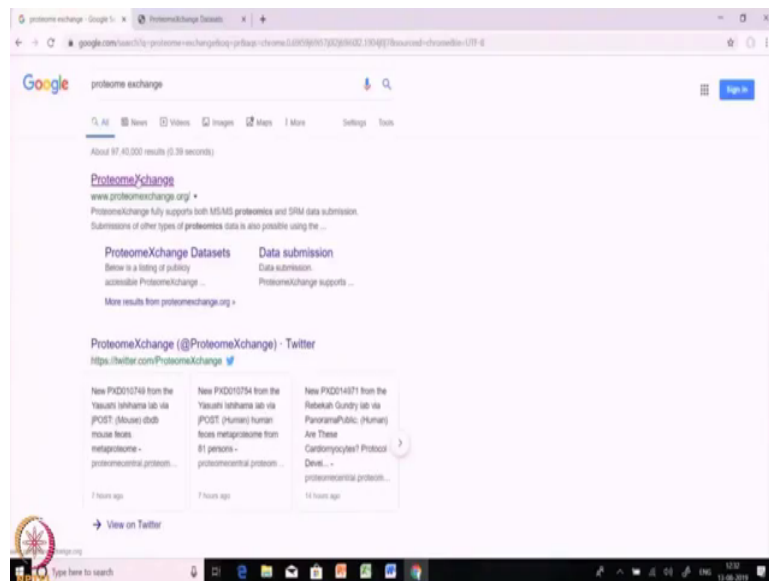
JPOST and Panorama; this proteome exchange will help you to download process and pre processed data.

(Refer Slide Time: 09:57)



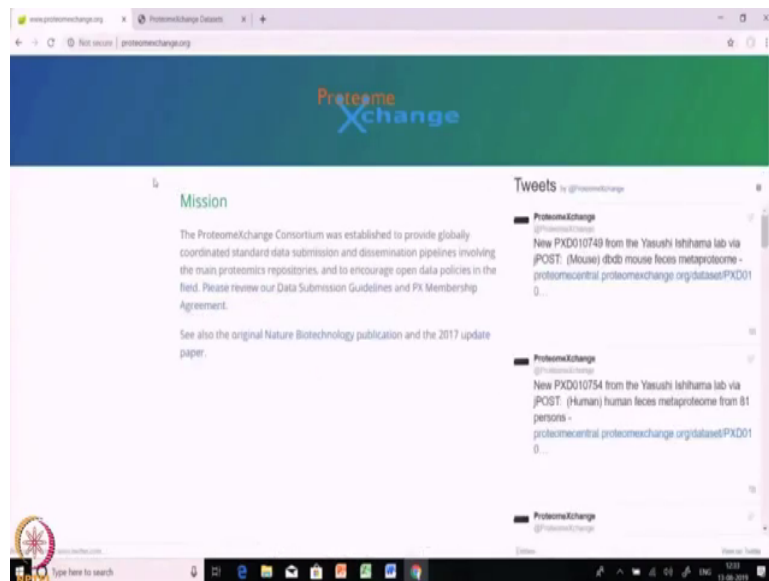
And let me give you a small hands on proteome exchange.

(Refer Slide Time: 10:11)



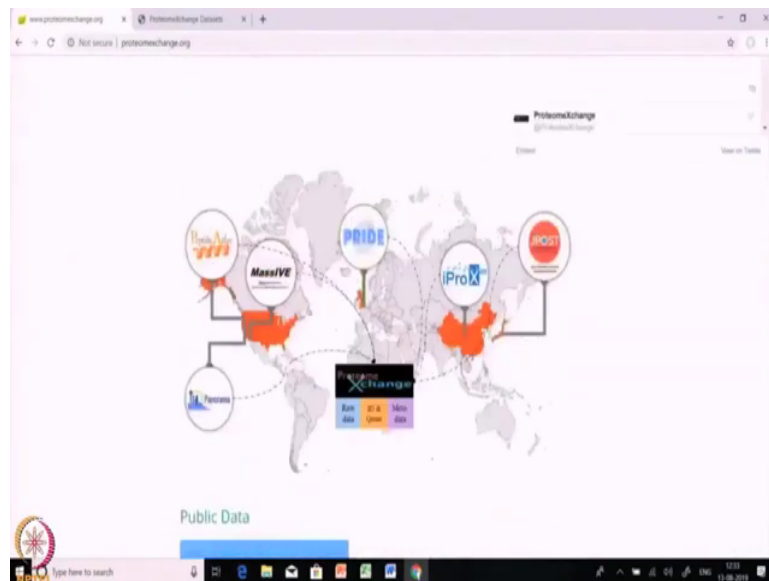
So, let us try to explore proteome exchange and I will show you how you can use proteome exchange for downloading different proteomics experiment data set. And how you can use those proteomics experiment data set in your further experiments. So, first let us first search proteome exchange in Google.

(Refer Slide Time: 10:31)



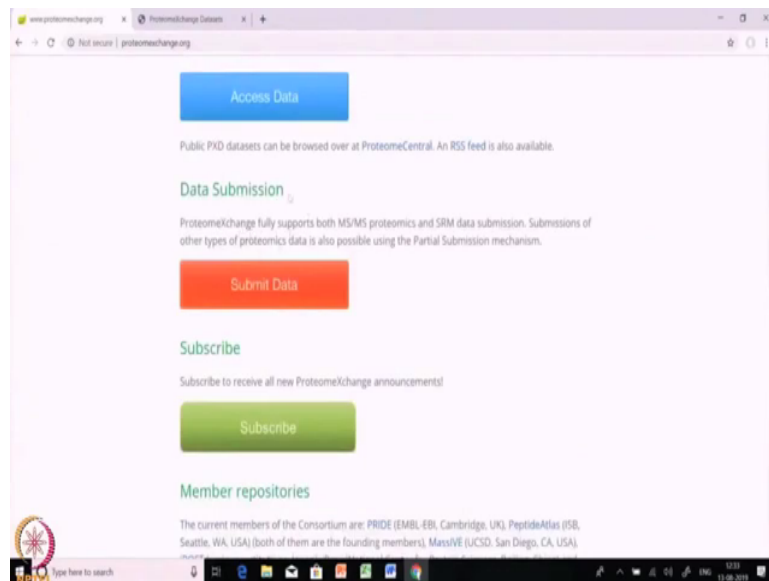
And click the first click the first one and as you can; as you can see that the proteome exchange consortium webpage is available. And you can read about the proteome exchange consortium over here.

(Refer Slide Time: 10:44)



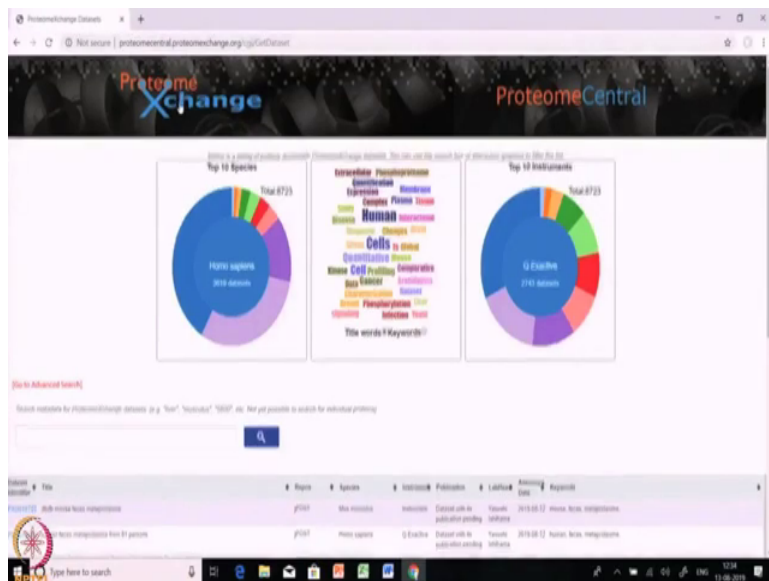
And you can understand how proteome exchange is a hub where all the different data sets. And repositories which are available worldwide has been interconnected to give you a better access for downloading data.

(Refer Slide Time: 10:58)

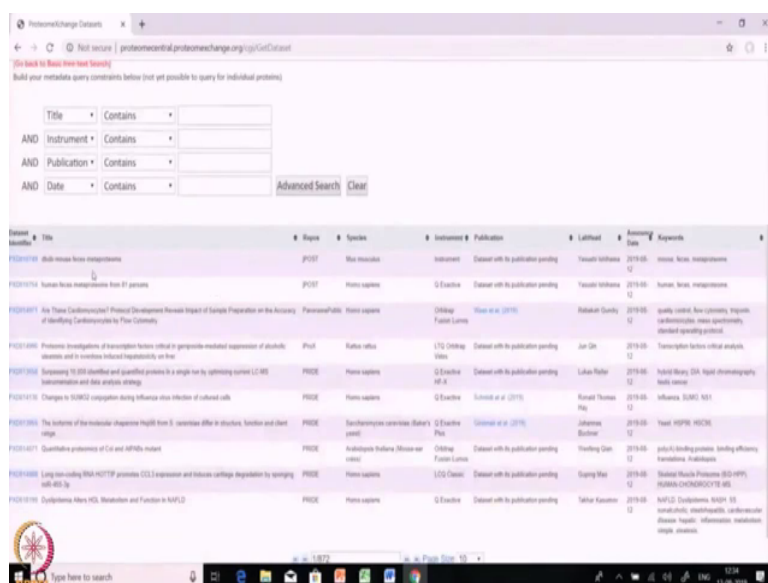


So, in proteome exchange there are three tabs are there; first one is for the public data, second one is for the data submission and third is the subscribe.

(Refer Slide Time: 11:23)



(Refer Slide Time: 11:32)



So, data submission is a data submission is important. And you will require this tab when you want to submit your own data of your own experimental a proteomic experimental generated data.

But for now, we will choose the public data for the for accessing the data. So, we will click the access data tab and we will found that a search a query search page will open, where you need to search your keywords for searching datasets. So, here we can use the advanced option and we can search on the basis of title data set even on the basis of instrument.

If you want to download the data, only a few the experimental data sets for a thermo fusion or you want data from (Refer Time: 11:57). So, you can mention the respecitic term respective term over here and you can do a advancers.

So, let me show you taking just a just an example of how to download a datasets. So, if I am if you come into the dataset result data set you will find that the there is a data set identifier. So, the data set identifier contains an unique PRIDE id.

So, this unique this PRIDE id is unique for each experiment and if you click this one it will redirect you to the page for download your data. Apart from this and the title option you will find a small title regarding the data set the repository name whether its a PRIDE repository j post repository or a massive repository.

So, the next is the species. So, even you can use species filtering or. So, if you want your data only from homo sapiens; then the instrument publication and whether the data is from which lab. So, the lab head name will be here let us choose the third one and I will and let us try how to download this datasets.

(Refer Slide Time: 13:15)

The screenshot displays the ProteomeXchange DataCite website interface. The browser address bar shows the URL: proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD014971. The main content area is titled "Full experiment listing" and "PXD014971". Below this, a "Dataset Summary" section provides key information:

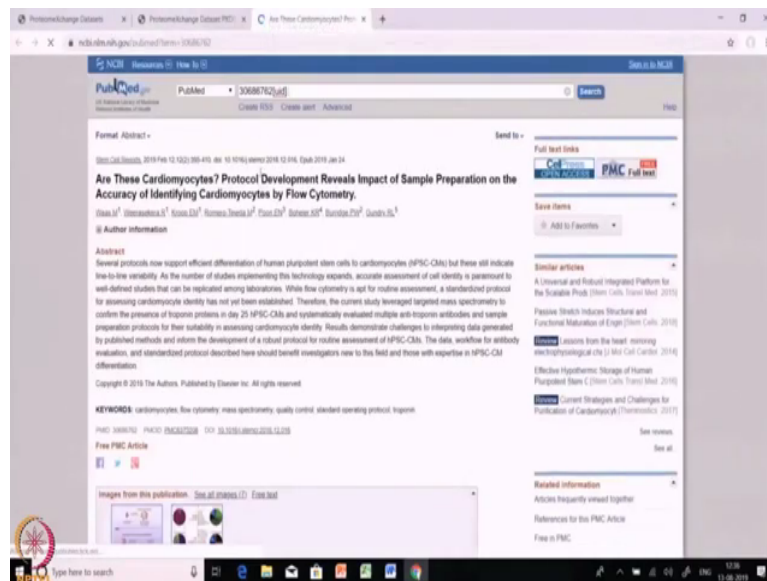
- HostingRepository:** ProteomeXchange
- AnnouncementDate:** 2019-08-12
- AnnouncementURL:** [Submission_2019-08-12_20:44:18.xml](#)
- DatasetIdentifier:** [Link]
- ReviewLevel:** Peer-reviewed dataset
- DatasetOrigin:** Original data
- RepositorySupport:** Supported dataset by repository
- PrimarySubmitter:** Matthew Weiss
- Title:** Are These Cardiac-specific? Protein Development Reveals Impact of Sample Preparation on the Accuracy of Identifying Cardiac-specific by Flow Cytometry
- Description:** Several protocols now support efficient differentiation of human pluripotent stem cells to cardiomyocytes (hPSC-CMs) but these still indicate low to low variability. As the number of studies implementing this technology expands, accurate assessment of cell identity is paramount to well-defined studies that can be replicated among laboratories. While flow cytometry is apt for routine assessment, a standardized protocol for assessing cardiomyocyte identity has not yet been established. Therefore, the current study leveraged targeted mass spectrometry to confirm the presence of known proteins in day 25 hPSC-CMs and systematically evaluated multiple anti-isoform antibodies and sample preparation protocols for their suitability in assessing cardiomyocyte identity. Results demonstrate challenges to integrating data generated by published methods and inform the development of a robust protocol for routine assessment of hPSC-CMs. The data, workflow for antibody evaluation, and standardized protocol described here should benefit investigators new to the field and those with expertise in hPSC-CM differentiation.
- SpeciesList:** scientific name: Homo sapiens, NCBI TaxID: 9606
- ModificationList:** Label TSC2(P102), Label TSC2(S104K)
- Instrument:** Orbitrap Fusion Lumos

Below the summary, a "Dataset History" table is shown:

| Version | Submission | Status | ChangeLog Entry |
|---------|---------------------|-----------|-----------------|
| 0 | 2019-08-09 15:55:20 | Requested | |
| 1 | 2019-08-12 08:44:18 | Announced | |

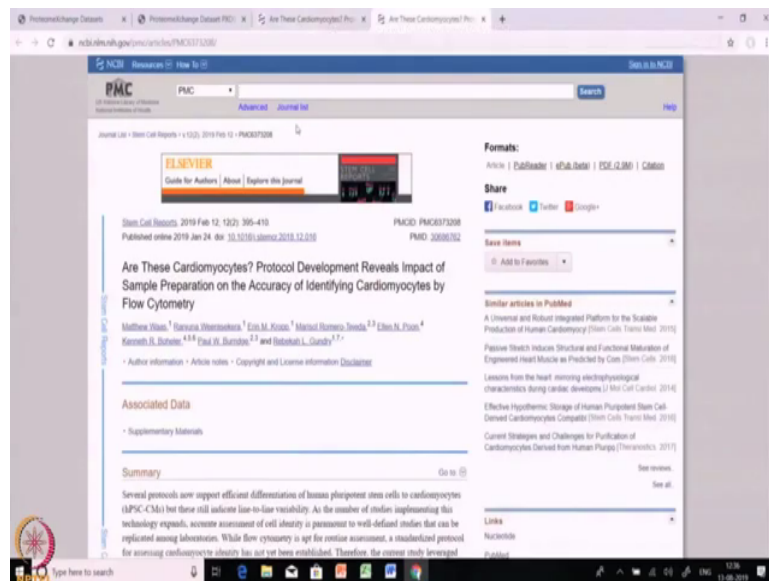
So, after clicking that you will be redirect to a interface of the PRIDE and you will find that the PXD014971 is the PRIDE unique PRIDE id for these datasets. There are a couple of informations are available about the data set like announcement date. Where, what is the description of the experiment spectral list modification list like this.

(Refer Slide Time: 13:44)



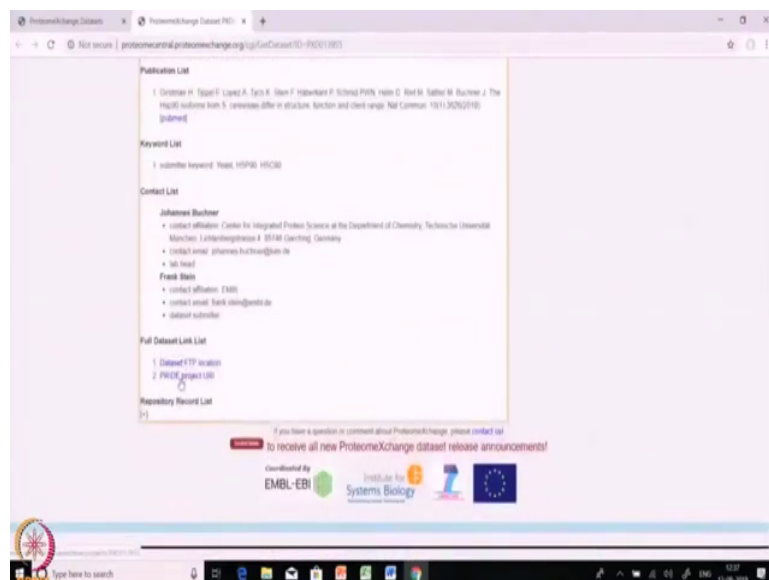
But if you come to a publication list and if you click here the page will redirect you to the public verification page of the NCBI. And you can check directly which paper it is and what is the material methods.

(Refer Slide Time: 13:52)



Or the results or the experimental evidences of this paper and this you can relate with the data set available over here.

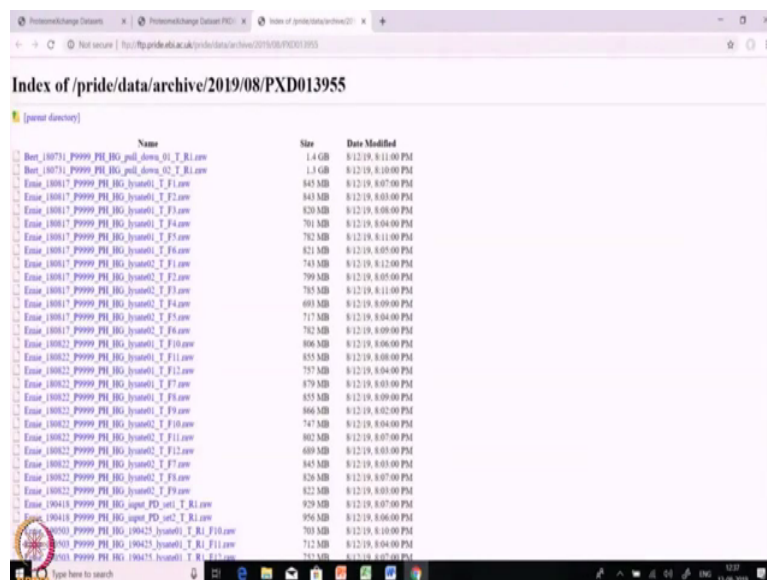
(Refer Slide Time: 14:02)



And it will help you to download the datasets. So, as we can see that this is the panorama public datasets; that means, this data say its available these datasets is available in the panorama. So, let us try to get some data set for from pride. So, we can use this one.

So, this is the another dataset whose unique PRIDE PRIDE id is PXD013955. So, here also you can see the all the details are available and the publication list is available. And there are two links are available one is dataset FTP location and the PRIDE project URL. So, you can use any of this link and that and it will redirect you to the directory from where you can download the data set.

(Refer Slide Time: 14:51)



Index of /pride/data/archive/2019/08/PXD013955

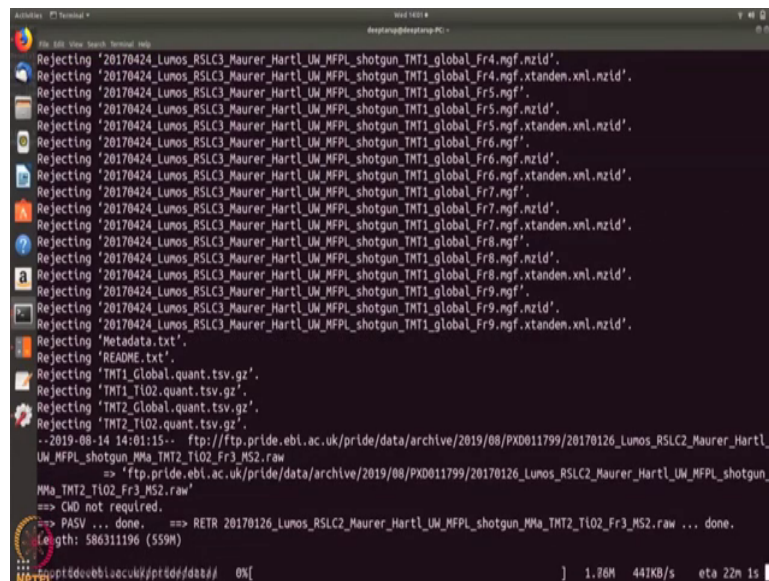
[parent directory]

| Name | Size | Date Modified |
|--|--------|--------------------|
| Best_180717_P9999_P01_H02_pull_down_01_T_R1.raw | 1.4 GB | 8/12/19 8:11:00 PM |
| Best_180717_P9999_P01_H02_pull_down_02_T_R1.raw | 1.3 GB | 8/12/19 8:10:00 PM |
| Enase_180817_P9999_P01_H02_hysat01_T_F1.raw | 843 MB | 8/12/19 8:07:00 PM |
| Enase_180817_P9999_P01_H02_hysat01_T_F2.raw | 843 MB | 8/12/19 8:03:00 PM |
| Enase_180817_P9999_P01_H02_hysat01_T_F3.raw | 830 MB | 8/12/19 8:08:00 PM |
| Enase_180817_P9999_P01_H02_hysat01_T_F4.raw | 701 MB | 8/12/19 8:04:00 PM |
| Enase_180817_P9999_P01_H02_hysat01_T_F5.raw | 782 MB | 8/12/19 8:11:00 PM |
| Enase_180817_P9999_P01_H02_hysat01_T_F6.raw | 821 MB | 8/12/19 8:05:00 PM |
| Enase_180817_P9999_P01_H02_hysat02_T_F1.raw | 743 MB | 8/12/19 8:12:00 PM |
| Enase_180817_P9999_P01_H02_hysat02_T_F2.raw | 799 MB | 8/12/19 8:05:00 PM |
| Enase_180817_P9999_P01_H02_hysat02_T_F3.raw | 785 MB | 8/12/19 8:11:00 PM |
| Enase_180817_P9999_P01_H02_hysat02_T_F4.raw | 680 MB | 8/12/19 8:09:00 PM |
| Enase_180817_P9999_P01_H02_hysat02_T_F5.raw | 717 MB | 8/12/19 8:04:00 PM |
| Enase_180817_P9999_P01_H02_hysat02_T_F6.raw | 782 MB | 8/12/19 8:09:00 PM |
| Enase_180822_P9999_P01_H02_hysat01_T_F10.raw | 806 MB | 8/12/19 8:06:00 PM |
| Enase_180822_P9999_P01_H02_hysat01_T_F11.raw | 835 MB | 8/12/19 8:08:00 PM |
| Enase_180822_P9999_P01_H02_hysat01_T_F12.raw | 757 MB | 8/12/19 8:04:00 PM |
| Enase_180822_P9999_P01_H02_hysat01_T_F7.raw | 879 MB | 8/12/19 8:03:00 PM |
| Enase_180822_P9999_P01_H02_hysat01_T_F8.raw | 855 MB | 8/12/19 8:09:00 PM |
| Enase_180822_P9999_P01_H02_hysat01_T_F9.raw | 866 MB | 8/12/19 8:02:00 PM |
| Enase_180822_P9999_P01_H02_hysat02_T_F10.raw | 747 MB | 8/12/19 8:04:00 PM |
| Enase_180822_P9999_P01_H02_hysat02_T_F11.raw | 802 MB | 8/12/19 8:07:00 PM |
| Enase_180822_P9999_P01_H02_hysat02_T_F12.raw | 680 MB | 8/12/19 8:03:00 PM |
| Enase_180822_P9999_P01_H02_hysat02_T_F7.raw | 845 MB | 8/12/19 8:03:00 PM |
| Enase_180822_P9999_P01_H02_hysat02_T_F8.raw | 826 MB | 8/12/19 8:07:00 PM |
| Enase_180822_P9999_P01_H02_hysat02_T_F9.raw | 822 MB | 8/12/19 8:03:00 PM |
| Enase_190418_P9999_P01_H02_super_P02_s01_T_R1.raw | 929 MB | 8/12/19 8:07:00 PM |
| Enase_190418_P9999_P01_H02_super_P02_s02_T_R1.raw | 956 MB | 8/12/19 8:06:00 PM |
| Enase_190401_P9999_P01_H02_190425_hysat01_T_R1_F10.raw | 700 MB | 8/12/19 8:10:00 PM |
| Enase_190401_P9999_P01_H02_190425_hysat01_T_R1_F11.raw | 712 MB | 8/12/19 8:04:00 PM |
| Enase_190401_P9999_P01_H02_190425_hysat01_T_R1_F12.raw | 744 MB | 8/12/19 8:05:00 PM |

So, for downloading these data set, first you need to know that whether what is the terminology or what are the files that is available; most of the time you will get this information from the publication supplementary files. So, you need to download the supplementary files.

And check the sample table or the sample information present over there with these data sets and on the basis of that you need to take the call which data sets you need to download. So, for downloading a data set if you just click any of the if you just click any of the files it will start getting download.

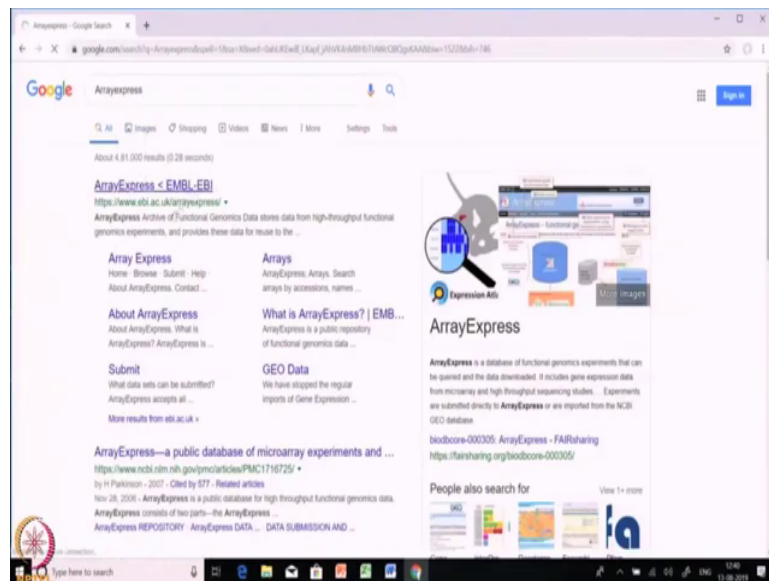
(Refer Slide Time: 15:28)



```
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr4.ngf.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr4.ngf.xtanden.xml.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr5.ngf'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr5.ngf.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr5.ngf.xtanden.xml.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr6.ngf'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr6.ngf.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr6.ngf.xtanden.xml.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr7.ngf'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr7.ngf.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr7.ngf.xtanden.xml.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr8.ngf'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr8.ngf.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr8.ngf.xtanden.xml.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr9.ngf'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr9.ngf.nzid'.
Rejection: '20170424_Lumos_RSLC3_Maurer_Hartl_UW_MFPL_shotgun_TMT1_global_Fr9.ngf.xtanden.xml.nzid'.
Rejection: 'Metadata.txt'.
Rejection: 'README.txt'.
Rejection: 'TMT1_Global.quant.tsv.gz'.
Rejection: 'TMT1_Tl02.quant.tsv.gz'.
Rejection: 'TMT2_Global.quant.tsv.gz'.
Rejection: 'TMT2_Tl02.quant.tsv.gz'.
..2019-08-14 14:01:15.. ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2019/08/PXD011799/20170126_Lumos_RSLC2_Maurer_Hartl_UW_MFPL_shotgun_MMA_TMT2_Tl02_Fr3_M52.raw
==> 'ftp.pride.ebi.ac.uk/pride/data/archive/2019/08/PXD011799/20170126_Lumos_RSLC2_Maurer_Hartl_UW_MFPL_shotgun_MMA_TMT2_Tl02_Fr3_M52.raw'
==> CWD not required.
==> PASV ... done. ==> RETR 20170126_Lumos_RSLC2_Maurer_Hartl_UW_MFPL_shotgun_MMA_TMT2_Tl02_Fr3_M52.raw ... done.
Length: 586311196 (559M)
[Progress bar] 1.76M 441KB/s eta 22m 1s
```

As simple as this, but when the file size are huge and you need to download the complete file size, clicking each files and downloading each file will be little tough. So, for that reason you can use Linux operating system. So, as you can see that downloading a complete experimental data set for an publication is. So, easy and even you can go for multiple data sets and you can download those you can integrate the data set and you can do further analysis.

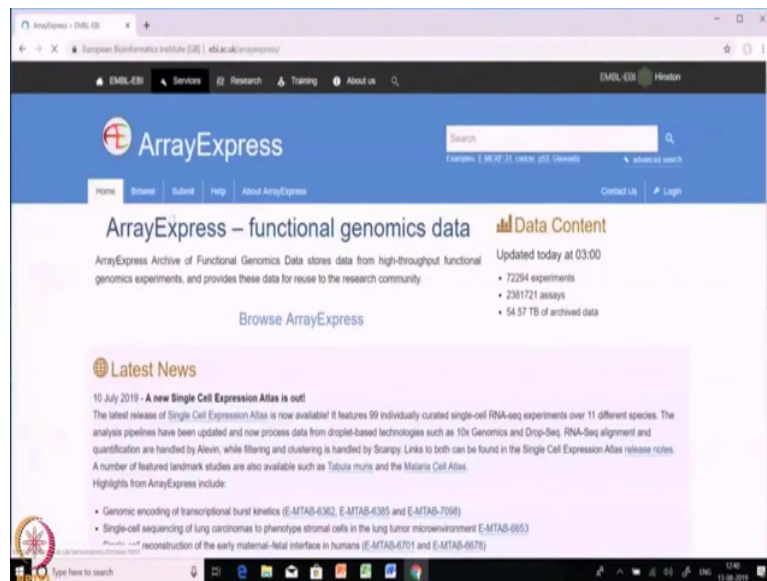
(Refer Slide Time: 16:03)



Let us go to another database that is array express. So, ArrayExpress is a kind of functional genomic data stores, where the high throughput functional genomics experiments provides this data and you can use this you can download this data.

And you can use it for your analysis. So, let us try to check how you can download your data from array express. First write array express in the Google and click the first search option of the ArrayExpress.

(Refer Slide Time: 16:35)



So, this is the web page of ArrayExpress where you can see that its a functional genomic data, but here you will find data from both proteomics and also genomics. Most of the data are available here from RNA sequence RNA seq microarray and even from mass spectrometry. So, let us try one example how to download a data from ArrayExpress. So, I am typing glioma and I am searching glioma in the array express.

(Refer Slide Time: 17:07)

The screenshot shows the ArrayExpress search results page for the query 'glioma'. The page header includes the ArrayExpress logo and navigation links. The search results are displayed as a table with columns: Accession, Title, Type, Organism, Assays, Released, Processed, Raw, Views, and Alias. The table shows 1168 experiments in total, with the first five results visible. The first result is E-MTAB-7804, titled 'Infant high grade gliomas comprise multiple subgroups characterised by repeat targetable gene fusions and better survival'. The second result is E-MTAB-7802, titled 'Infant high grade gliomas comprise multiple subgroups characterised by repeat targetable gene fusions and better survival'. The third result is E-MTAB-7724, titled 'Transcriptomic analysis by AmplicSeq ion Torrent to understand the role of LKB1 and PARC3 in Glioblastoma biology'. The fourth result is E-MTAB-7800, titled 'Profiling of BP188 ATTK isogenic glioma cells'. The fifth result is E-MTAB-6881, titled 'RNA-Seq of human intracranially implanted U87MG-GFP cells as a function of distance from'.

| Accession | Title | Type | Organism | Assays | Released | Processed | Raw | Views | Alias |
|-------------|---|--------------------------------|--------------|--------|------------|-----------|-----|-------|-------|
| E-MTAB-7804 | Infant high grade gliomas comprise multiple subgroups characterised by repeat targetable gene fusions and better survival | methylation profiling by array | Homo sapiens | 304 | Yesterday | - | - | - | - |
| E-MTAB-7802 | Infant high grade gliomas comprise multiple subgroups characterised by repeat targetable gene fusions and better survival | methylation profiling by array | Homo sapiens | 158 | 09/08/2019 | - | - | - | - |
| E-MTAB-7724 | Transcriptomic analysis by AmplicSeq ion Torrent to understand the role of LKB1 and PARC3 in Glioblastoma biology | RNA-seq of coding RNA | Homo sapiens | 18 | 31/07/2019 | - | - | - | - |
| E-MTAB-7800 | Profiling of BP188 ATTK isogenic glioma cells | methylation profiling by array | Homo sapiens | 4 | 01/06/2019 | - | - | - | - |
| E-MTAB-6881 | RNA-Seq of human intracranially implanted U87MG-GFP cells as a function of distance from | RNA-seq of coding RNA | Homo sapiens | 4 | 30/04/2019 | - | - | - | - |

So, after the search gets over, we can see there are multiple data sets that are available and total 1168 experiments are total available with the search in the search result of glioma. And this 1168 is a huge number and we cannot download all the files or; that means, we need to put some filter.

(Refer Slide Time: 17:32)

The screenshot shows the ArrayExpress website interface. The search bar at the top contains the term 'glioma'. On the left, a 'Filter search results' sidebar is visible with the following settings: 'By organism' set to 'Human', 'By experiment type' set to 'Protein assay', and 'By array' set to 'Mass spectrometry assay'. The main content area displays 'Search results for glioma' with a table of results. The table has columns for Type, Organism, Assays, Released, Processed, Raw, Views, and Atlas. The results are filtered to show 1-20 of 1169 experiments. The first few rows of the table are as follows:

| Type | Organism | Assays | Released | Processed | Raw | Views | Atlas |
|---------------------|-------------|--------------|----------|------------|-----|-------|-------|
| proteomics multiple | methylation | Homo sapiens | 304 | Yesterday | - | - | - |
| proteomics multiple | methylation | Homo sapiens | 158 | 09/08/2019 | - | - | - |
| proteomics multiple | methylation | Homo sapiens | 18 | 31/07/2019 | - | - | - |
| proteomics multiple | methylation | Homo sapiens | 4 | 01/06/2019 | - | - | - |
| proteomics multiple | methylation | Homo sapiens | 4 | 30/04/2019 | - | - | - |

So, for this we need to select the filter search results over here. And as you can see there are multiple filters that we can put. Let us put homo sapiens let us put human then for the experimental tribe.

Let us choose protein assays; further let us choose micro mass spectrometry assay. And filter the data and after filtering the complete data with the following given filters.

(Refer Slide Time: 18:02)

The screenshot shows the EMBL-EBI search results page for the query 'glioma'. The page is filtered by organism 'Human', experiment type 'protein assay', and experiment type 'mass spectrometry assay'. A table of results is displayed, with the first entry being 'E-PR07-2' titled 'Proteomic profiling of NCI60 cell lines from Cancer Cell Line Encyclopedia'. The table has columns for Accession, Title, Type, Organism, Assays, Released, Processed, Raw, Views, and Atlas. Below the table, there are links to export data in various formats and a note about the ELIXIR infrastructure. The footer contains navigation links for Services, Research, Training, Industry, and About EMBL-EBI, along with a cookie consent banner.

| Accession | Title | Type | Organism | Assays | Released | Processed | Raw | Views | Atlas |
|-----------|--|--|---------------|--------|------------|-----------|-----|-------|-------|
| E-PR07-2 | Proteomic profiling of NCI60 cell lines from Cancer Cell Line Encyclopedia | proteomic profiling by mass spectrometry | Human sapiens | 60 | 27/01/2017 | 1 | - | 863 | - |

We can see that there is only one file that is present only ones data set that is present which is having the proteomics data as this database as this repository is completely based on is mainly based on genomics data.

(Refer Slide Time: 18:18)

The screenshot shows the EMBL-EBI search results page for the query 'glioma'. The page features a sidebar with filters on the left, a main search results area in the center, and a footer with navigation links. The filters include 'By organism' (Human), 'By experiment type' (RNA assay), 'All technologies', and 'By array' (All arrays). The main search results area displays a table with columns: Type, Organism, Assays, Released, Processed, Raw, Views, and Atlas. The table shows one result for 'glioma' with 60 assays, released on 27/01/2017, and 663 views. The result is labeled 'glioma' and 'glioma'.

Search results for glioma

experiment type protein assay experiment type mass spectrometry assay

| Type | Organism | Assays | Released | Processed | Raw | Views | Atlas |
|--------|----------|--------|------------|-----------|-----|-------|-------|
| glioma | glioma | 60 | 27/01/2017 | | | 663 | |

EMBL-EBI

Services

Research

Training

Industry

About EMBL-EBI

So, let us try something with genomics. So, maybe we can use this one as RNA assay and we can choose this as all technology and let us filter the data.

(Refer Slide Time: 18:38)

The screenshot displays the ArrayExpress website interface. At the top, there is a search bar with the text 'glioma' entered. Below the search bar, the results are filtered by organism 'Human' and experiment type 'RNA assay'. The results are displayed in a table with columns: Accession, Title, Type, Organism, Assays, Released, Processed, Raw, Views, and Atlas. The table shows three results, each with a title describing a transcriptomic analysis or RNA-seq experiment related to glioma. The first result is E-MTAB-7724, the second is E-MTAB-6881, and the third is E-MTAB-6882. The table also indicates that there are 671 experiments in total, with the first 25 shown. The page size is set to 25. At the bottom of the page, there is a cookie consent banner.

| Accession | Title | Type | Organism | Assays | Released | Processed | Raw | Views | Atlas |
|-------------|---|-----------------------|--------------|--------|------------|-----------|-----|-------|-------|
| E-MTAB-7724 | Transcriptomic analysis by Amplicon Sequencing to understand the role of LKB1 and PARC3 in Glioblastoma biology | RNA-seq of coding RNA | Homo sapiens | 18 | 31/07/2019 | - | - | - | - |
| E-MTAB-6881 | RNA-Seq of human intracranially implanted U87MG-GFP cells as a function of distance from blood vessels | RNA-seq of coding RNA | Homo sapiens | 4 | 30/04/2019 | - | - | - | - |
| E-MTAB-6882 | RNA-Seq of human U87MG-GFP cells from sub-ocan xenografts as a function of distance from blood vessels | RNA-seq of coding RNA | Homo sapiens | 9 | 29/04/2019 | - | - | - | - |

And after filtering after the filtering gets over, you can see there are 671 experiments we found. And there are multiple data datasets and experiments are available. So, now, on the basis of if we click the assays to it will help in sorting out on the basis of the maximum number of assays; you can select which experiments.

(Refer Slide Time: 18:55)

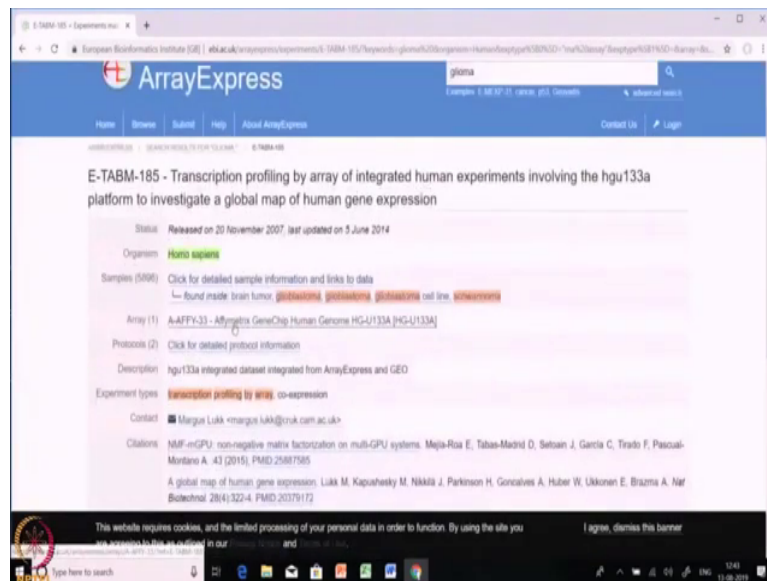
The screenshot displays the ArrayExpress website interface. At the top, there is a search bar with the text 'glioma' entered. Below the search bar, the results are filtered by 'organism: Human' and 'experiment type: microarray'. The results are displayed in a table with columns: Accession, Title, Type, Organism, Assays, Released, Processed, Raw, Views, and Atlas. The table lists four datasets:

| Accession | Title | Type | Organism | Assays | Released | Processed | Raw | Views | Atlas |
|--------------|--|----------------------------------|----------------------------|--------|------------|-----------|-----|-------|-------|
| E-MTAB-3732 | A comprehensive human expression map | transcription profiling by array | Homo sapiens | 27871 | 23/07/2015 | | | 21875 | - |
| E-TABM-186 | Transcription profiling by array of integrated human experiments involving the hgu133a platform to investigate a global map of human gene expression | transcription profiling by array | Homo sapiens | 5886 | 20/11/2007 | | | 12887 | - |
| E-MTAB-42 | Human gene expression atlas of 5372 samples representing 369 different cell and tissue types, disease states and cell lines | transcription profiling by array | Homo sapiens, Mus musculus | 5372 | 04/08/2010 | | | 48738 | - |
| E-GEOD-36139 | SNP and Expression data from the Cancer Cell Line Encyclopedia (CCLE) | transcription profiling by array | Homo sapiens | 1864 | 20/03/2012 | | | 1877 | - |

At the bottom of the page, there is a cookie consent banner that reads: 'This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our privacy policy and cookie policy.' There is a button to 'I agree, dismiss this banner'.

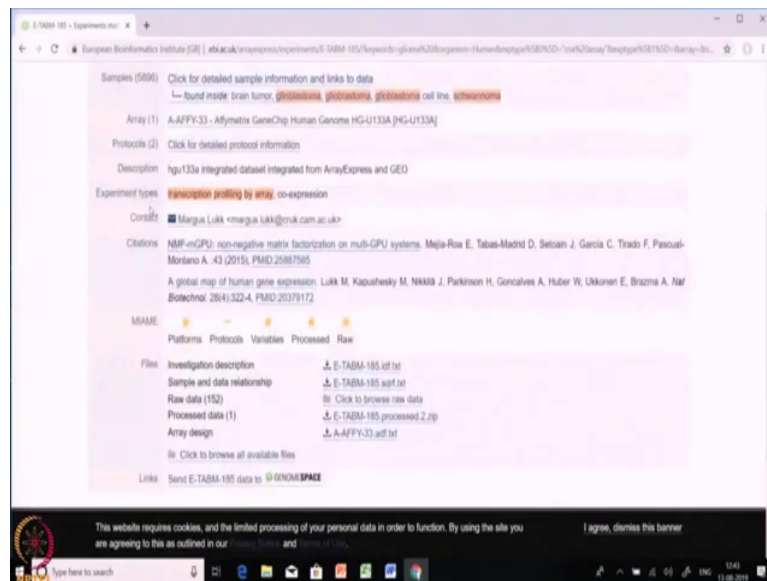
You want then which dataset you want and you can download though data set by just clicking the accession number.

(Refer Slide Time: 19:10)



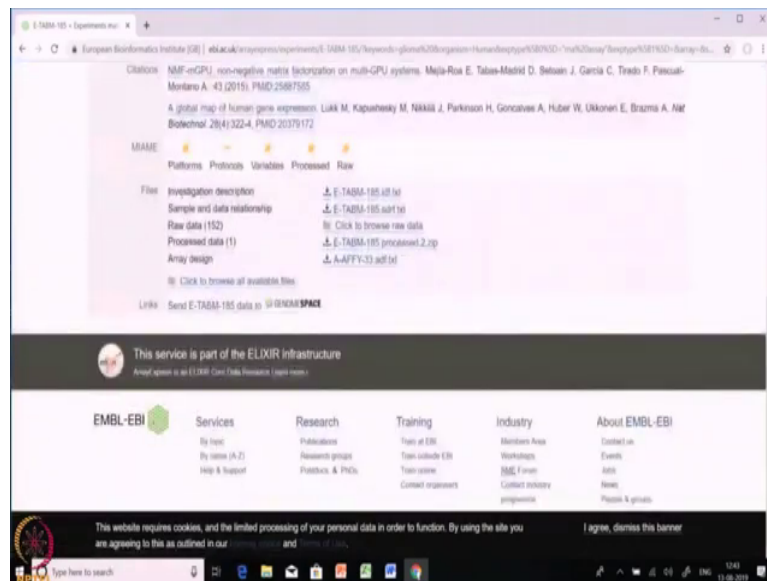
So, after clicking the accession the it will redirect you to the to another page where it will give you the details of the experiments.

(Refer Slide Time: 19:21)



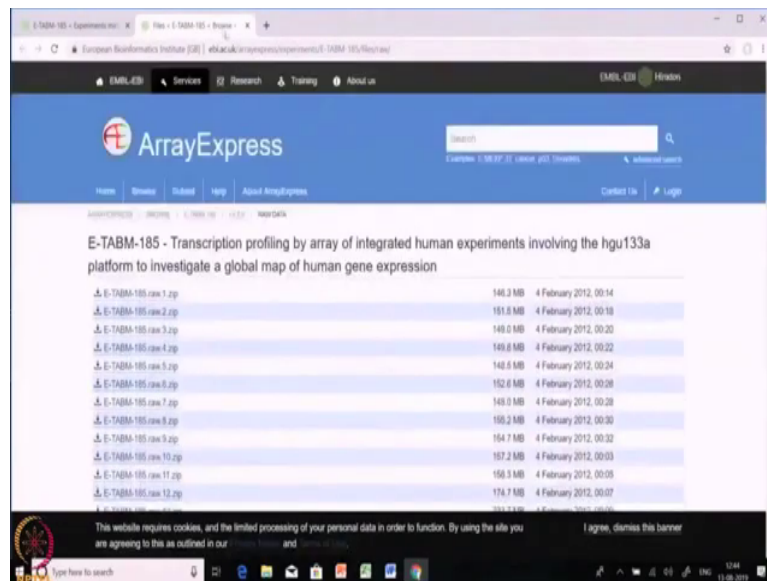
So, like the name of the title of the experiment followed by contacts description of the experiments and different samples and what are the files available.

(Refer Slide Time: 19:28)



So, the main important part is what are the files available over here. So, the first is the investigation description second is sample and data relationship this is the most important. And crucial file that we need to download from array express because it will give you its a kind of a metadata file, which will give you a complete in total information of sample and data relationship. Next is the raw data if you want to download the raw data. As you can see there is 152 you need to click the raw data.

(Refer Slide Time: 20:02)

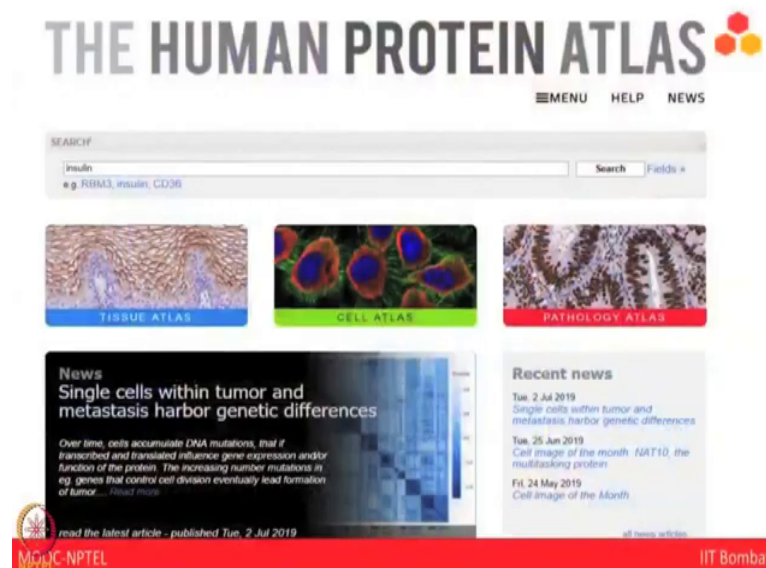


The screenshot shows the ArrayExpress website interface. The main heading is "E-TABM-185 - Transcription profiling by array of integrated human experiments involving the hgu133a platform to investigate a global map of human gene expression". Below this, there is a table listing 12 raw data files for download. Each row includes a download icon, the file name, the file size, and the upload date and time.

| File Name | File Size | Upload Date/Time |
|-----------------------|-----------|------------------------|
| E-TABM-185.raw.1.zip | 146.3 MB | 4 February 2012, 00:14 |
| E-TABM-185.raw.2.zip | 151.6 MB | 4 February 2012, 00:18 |
| E-TABM-185.raw.3.zip | 149.0 MB | 4 February 2012, 00:20 |
| E-TABM-185.raw.4.zip | 149.6 MB | 4 February 2012, 00:22 |
| E-TABM-185.raw.5.zip | 142.6 MB | 4 February 2012, 00:24 |
| E-TABM-185.raw.6.zip | 152.6 MB | 4 February 2012, 00:28 |
| E-TABM-185.raw.7.zip | 145.0 MB | 4 February 2012, 00:28 |
| E-TABM-185.raw.8.zip | 155.2 MB | 4 February 2012, 00:30 |
| E-TABM-185.raw.9.zip | 154.7 MB | 4 February 2012, 00:32 |
| E-TABM-185.raw.10.zip | 157.2 MB | 4 February 2012, 00:03 |
| E-TABM-185.raw.11.zip | 156.3 MB | 4 February 2012, 00:05 |
| E-TABM-185.raw.12.zip | 174.7 MB | 4 February 2012, 00:07 |

And it will take you to a page where you can download the complete raw data over here. So, you can see like how array express can also be used for downloading a data set. Let us move towards an database publicly available database which is very much informative and in proteomics.

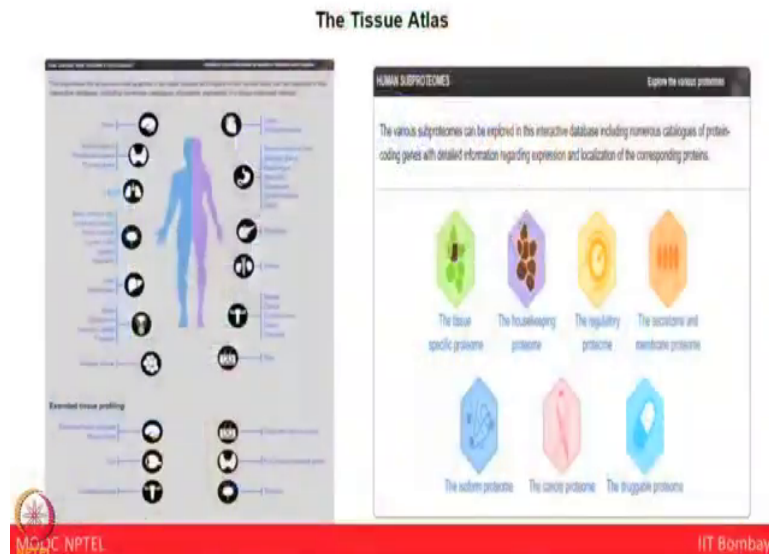
(Refer Slide Time: 20:25)



We should know about this database to a large extent that is Human Proteome Atlas HPA it is a Swedish based is a Swedish based program, which were started in 2003 with the aim to map all the human protein in cell tissue and organ; using integration of various omics technologies.

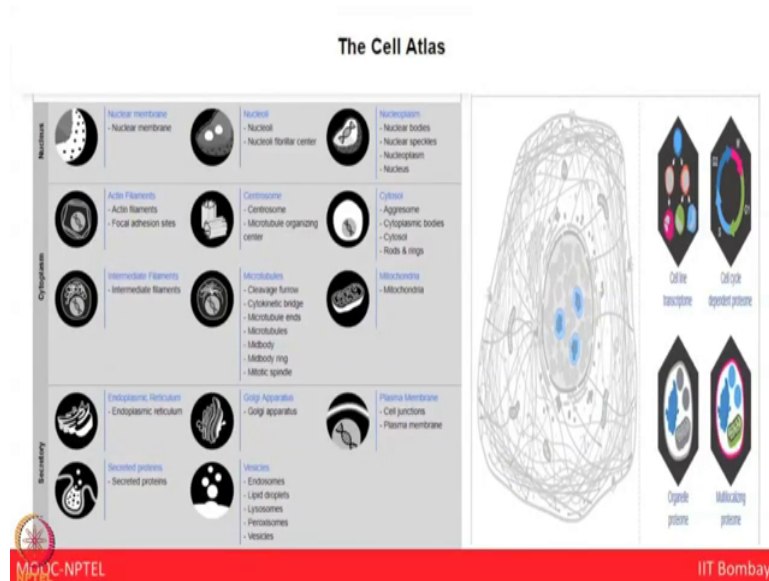
Including antibody based imaging mass spectrometry, based proteomics, transcriptomics and system biology. All the data in the knowledge resources is open access to all the scientists both in academia and industry to freely access the data for exploration of the human proteome. So, the human proteome database. So, the human proteome atlas has been broadly classified into 3 major atlases.

(Refer Slide Time: 21:15)



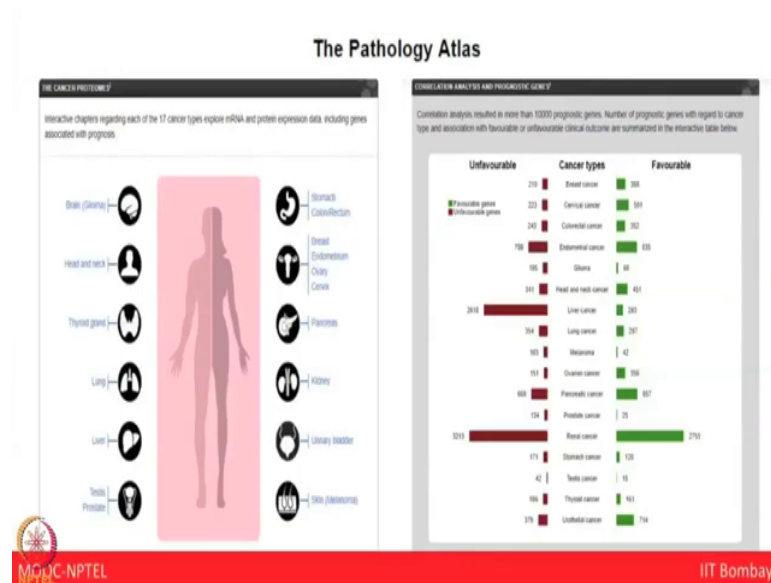
The first is a tissue atlas which contains the information regarding the expression profile of human genes both on mRNA and protein level. The protein expression data from 44 normal human tissues type is derived from antibody based protein profiling using immunohistochemistry.

(Refer Slide Time: 21:30)



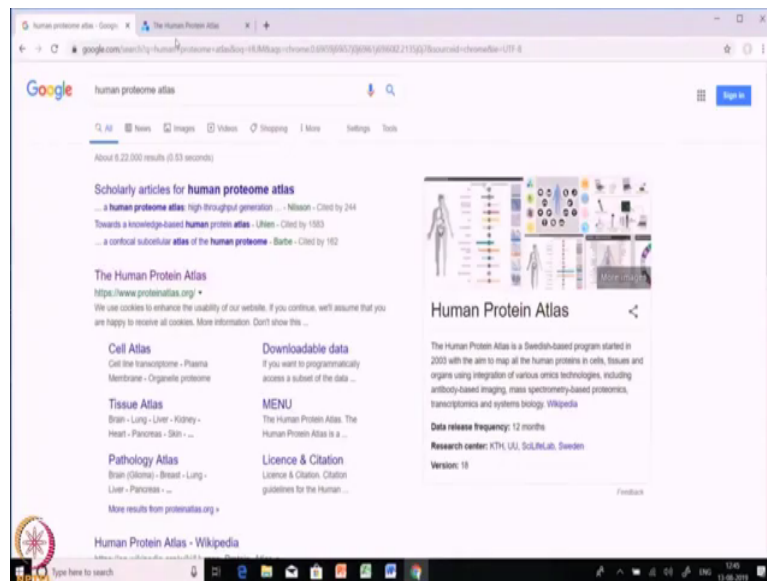
Next the cell atlas, the cell atlas provide high resolution insights into spatter temporal distribution of proteins within human cells. The protein localization data is derived from antibody based profiling.

(Refer Slide Time: 21:42)



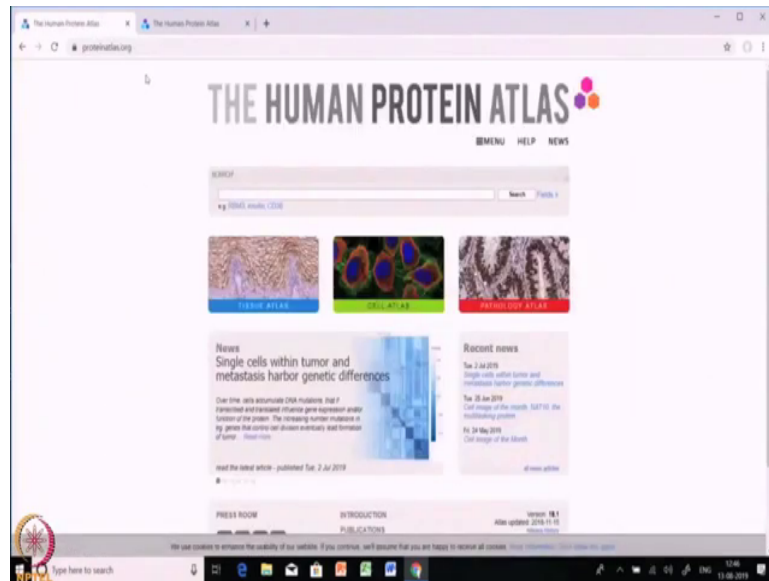
By immunofluorescence confocal microscopy using a panel of 64 cell lines to represent various cell population in different organs and tissue of the human body.

(Refer Slide Time: 21:53)

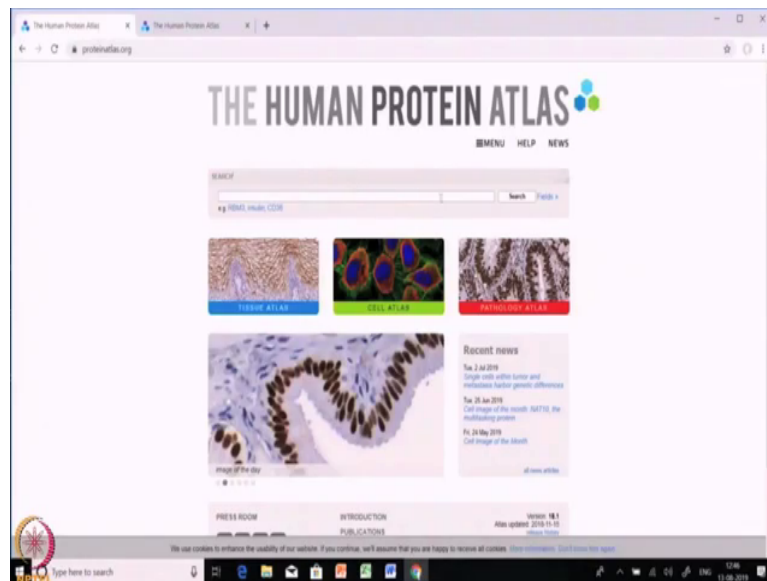


So, let us try to explore the human proteome atlas. So, as we have already understand human proteome atlas is having cell atlas, tissue atlas and pathology atlas. Now we let us try how one protein can be searched in all this part in all this atlas and we can get huge amount of information about a protein.

(Refer Slide Time: 22:16)

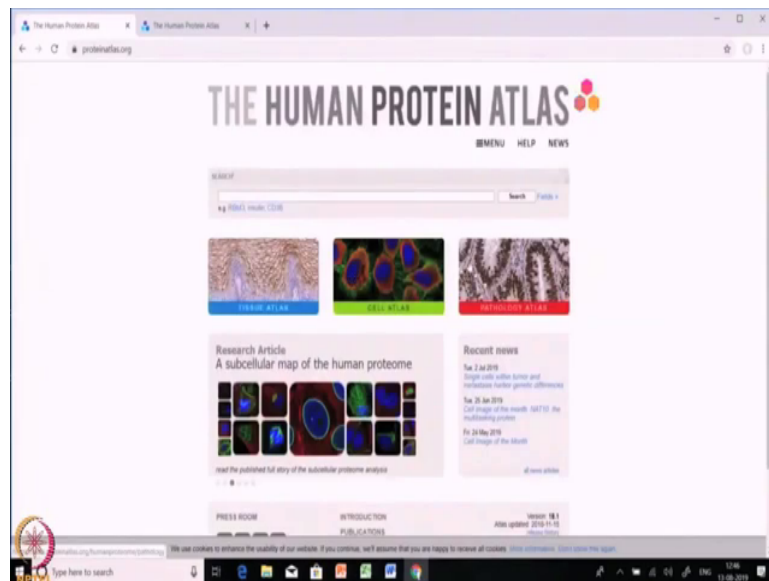


(Refer Slide Time: 22:22)



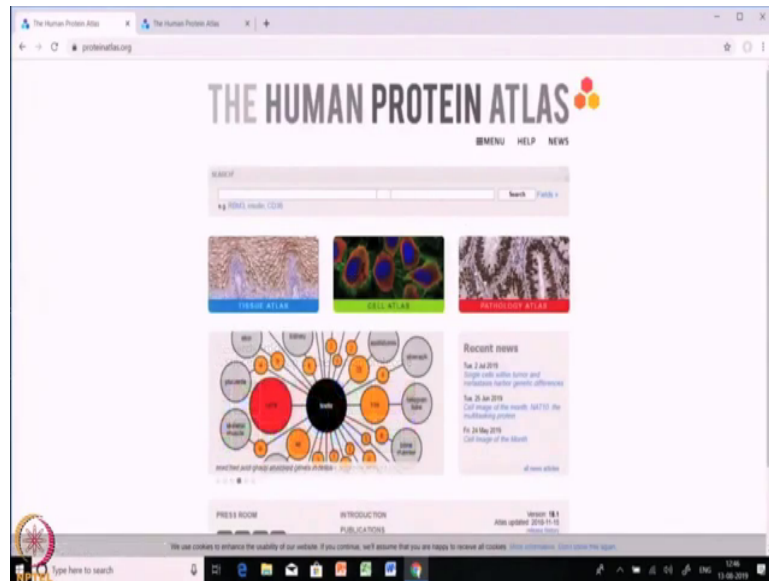
So, search human atlas human proteome atlas in Google and click the first one and you will find a search dialog box in the first page.

(Refer Slide Time: 22:28)

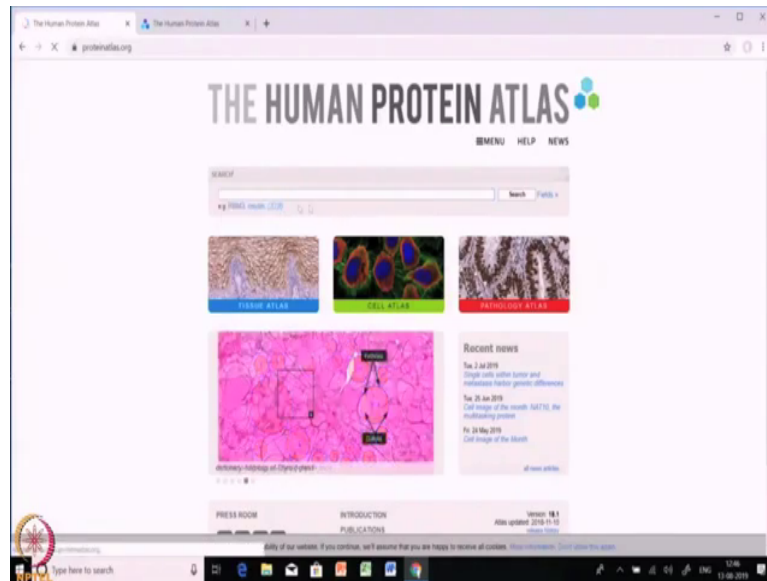


Even you can read each of these tabs tissue atlas, cell atlas and pathology atlas to get more information.

(Refer Slide Time: 22:33)



(Refer Slide Time: 22:39)



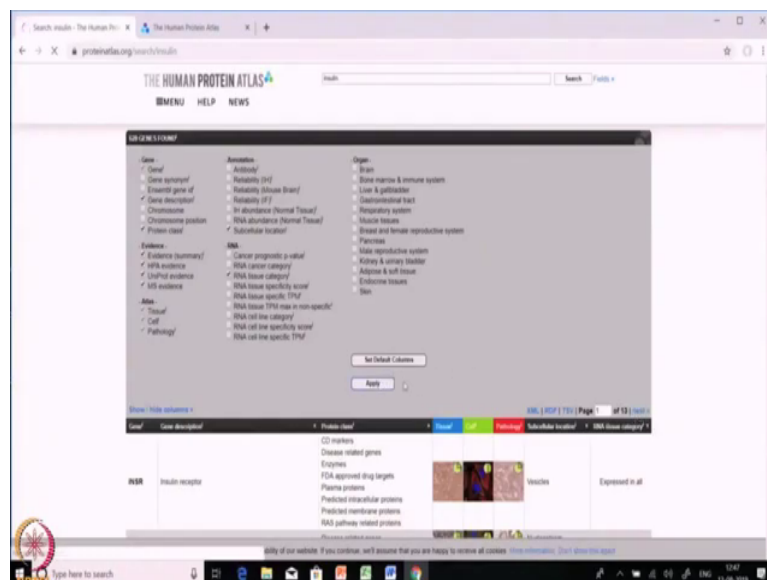
So, let us try with one example insulin.

(Refer Slide Time: 22:40)

| Gene | Gene description | Protein class | CD markers | Disease related genes | Enzymes | FTSA approved drug targets | Plasma proteins | Predicted intracellular proteins | Predicted membrane proteins | RAS pathway related proteins | RNA | Neutrophils | Macrophages | Plasma membrane | Epithelium | Expressed in all |
|--------|--|--|------------|-----------------------|---------|----------------------------|-----------------|----------------------------------|-----------------------------|------------------------------|-----|-------------|-------------|-----------------|------------|------------------|
| INS | Insulin receptor | CD markers Disease related genes Enzymes FTSA approved drug targets Plasma proteins Predicted intracellular proteins Predicted membrane proteins RAS pathway related proteins | | | | | | | | | | | | | | |
| INSR | Insulin receptor substrate 1 | Disease related genes Plasma proteins Predicted intracellular proteins | | | | | | | | | | | | | | |
| INS | Insulin | Cancer related genes Disease related genes Plasma proteins Predicted intracellular proteins RAS pathway related proteins | | | | | | | | | | | | | | |
| IGF1R | Insulin like growth factor 1 receptor | Cancer related genes CD markers Disease related genes Enzymes FTSA approved drug targets Plasma proteins Predicted intracellular proteins Predicted membrane proteins RAS pathway related proteins | | | | | | | | | | | | | | |
| IGFBP3 | Insulin like growth factor binding protein 3 | Cancer related genes Plasma proteins Predicted intracellular proteins | | | | | | | | | | | | | | |

So, as you can see when I am searching insulin all the proteins, which are all the genes which are available for insulins are present here. So, now, you can see that there are multiple tabs which are present and which are giving you multiple columns which are present which are giving you lots of information. Like gene description protein classes whether it is what is the information available in tissue atlas, cell atlas and pathology atlas.

(Refer Slide Time: 23:12)

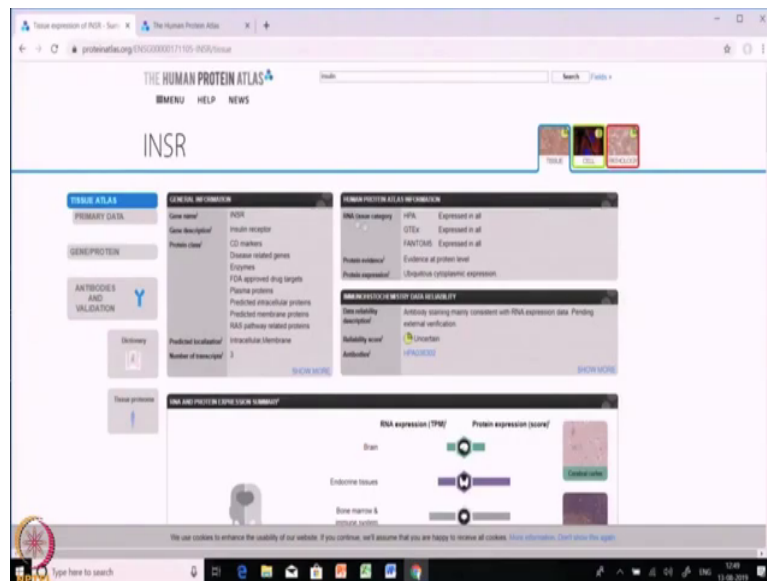


Now, let us try to incorporate some more information. If you are clicking the advanced option, you will find that there are multiple tabs available over here. So, if I want to check whether the proteins in the evidence level whether it is having the evidence level. Whether it is having any HP evidence, whether it is present in UniProt; that means, it is having an UniProt evidence or whether the protein is having any MS evidence.

So, by clicking these four things so, by clicking these four options we can select apply and you can see that all this information has been incorporated in the columns. So, now, if I am choosing the first one INSR, which is insulin receptor the insulin receptor is having evidence level both in case of HPA UniProt and also it is having a MS evidence.

And if you just put your cursor in the green colour box, you will find that what is the evidence level it is present. Like if I am putting in MS evidence it is showing that it is present in evidence at protein level. So, now, let us click INSR insulin receptor.

(Refer Slide Time: 24:20)



And let us try to explore more what are the different information available for this particular protein. So, the first thing here we can see the there is a general information of this protein and it has given gene name protein class and even localization. So, next thing is given for the human protein atlas information and RNA tissue category protein evidence and protein expression.

So, protein evidence is saying that it is evidence is present in a protein level. Another thing I want to tell the final annotated protein expression, which is given here is estimated on the basis of antibody data RNA seq data and available protein and gene characterization data.

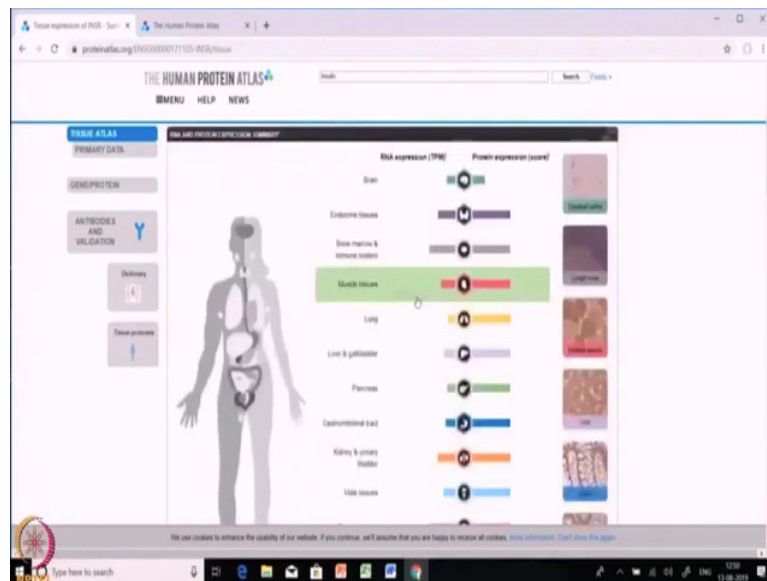
So, if we move forward we found that there is a reliability score present here and it is telling uncertain. So, what does this mean? Actually on the basis of the reliability score that proteins has been divided into four category; first is the enhanced, second is supported, third is approved and fourth is uncertain.

So, enhanced means if one or several antibodies with non overlapping epitopes targeting the same gene have obtained; based on orthogonal or independent antibody validation method.

Then they give this code as enhanced second is supported; that means, the consistency with RNA sequence data or protein characterization data. In combination with similar staining pattern, if independent antibodies are available, then they will give this score as supported. Approved says consistency with RNA sequence data, in combination with inconsistency or lack of protein gene characterization data. Or alternatively the same thing there is an inconsistency with RNA sequencing data.

Then this data is telled as approved. Finally, uncertain that inconsistency with or lack of RNA sequencing or protein characterization data, in combination with dissimilar staining pattern then they give this reliability score as uncertain.

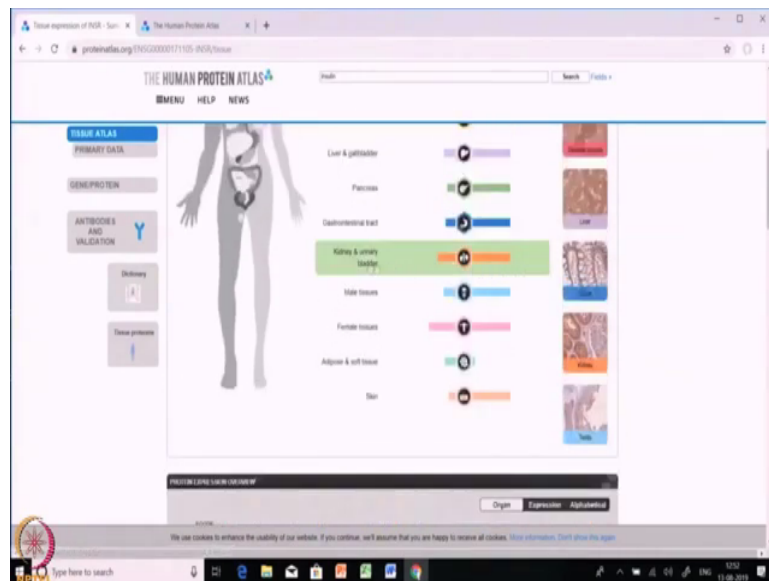
(Refer Slide Time: 26:38)



So, finally, if we move down, we will find that there is a RNA expression; RNA expression profile and protein expression score. So, as we can see what this profile has given so, much information that what is the particular level of expression in RNA and protein for a particular protein.

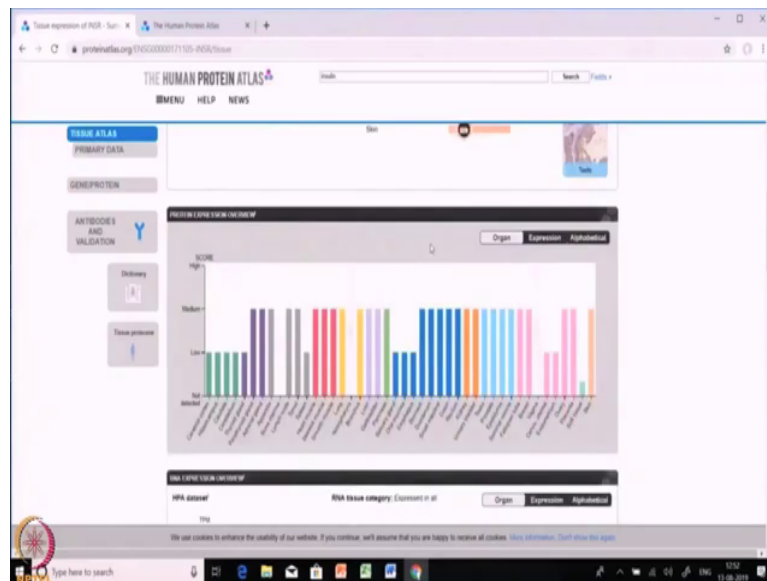
Same thing if we go down, we will find that the RNA expression profile and the protein expression profile is given for the particular protein in different tissues over here.

(Refer Slide Time: 27:08)



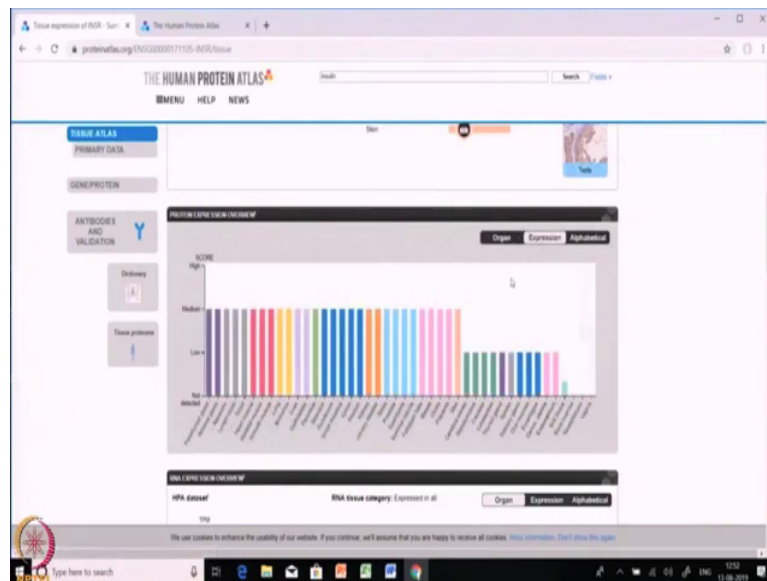
And to get more information, we can click each of these tabs over here and we will get more information.

(Refer Slide Time: 27:14)

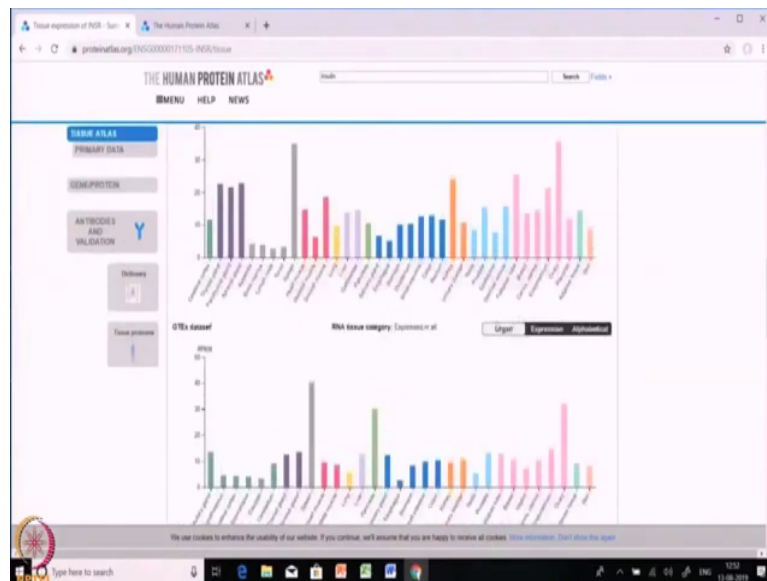


Next the protein expression overview and the RNA expression overview; this has given a complete. So, this has given the complete view of how what is the level of protein.

(Refer Slide Time: 27:28)



(Refer Slide Time: 27:36)



What is the expression of protein in 44 different tissues; in protein in terms of protein and in terms of RNA.

(Refer Slide Time: 27:40)

The screenshot displays the 'The Human Protein Atlas' website. The main content area shows the profile for the CD203 gene. The left sidebar contains navigation links: 'Tissue Atlas', 'Primary Data', 'Gene/Protein', 'Antibodies and Validation', 'Delivery', and 'Tissue presence'. The main content area is titled 'CD203 (HNC, human)' and includes a 'Gene description' section. The description states: 'The gene encodes a member of the receptor tyrosine kinase family of proteins. The encoded preproprotein is proteolytically processed to generate alpha and beta subunits that form a heterodimeric receptor. Binding of insulin or other ligands to this receptor activates the insulin signaling pathway, which regulates glucose uptake and release, as well as the synthesis and storage of carbohydrates, lipids and protein. Mutations in this gene underlie the inherited insulin resistance syndromes including type A insulin resistance syndrome, Donohue syndrome and Rabson-Mendenhall syndrome. Alternative splicing results in multiple transcript variants. (provided by RefSeq, Oct 2016)'. Below the description, there is a 'Protein domain' section showing a bar chart of protein domains. The bottom of the page features a search bar and a footer with a copyright notice.

THE HUMAN PROTEIN ATLAS
HOME HELP NEWS

Tissue Atlas
PRIMARY DATA

Gene/Protein

ANTIBODIES AND VALIDATION

Delivery

Tissue presence

CD203 (HNC, human)

Gene description

The gene encodes a member of the receptor tyrosine kinase family of proteins. The encoded preproprotein is proteolytically processed to generate alpha and beta subunits that form a heterodimeric receptor. Binding of insulin or other ligands to this receptor activates the insulin signaling pathway, which regulates glucose uptake and release, as well as the synthesis and storage of carbohydrates, lipids and protein. Mutations in this gene underlie the inherited insulin resistance syndromes including type A insulin resistance syndrome, Donohue syndrome and Rabson-Mendenhall syndrome. Alternative splicing results in multiple transcript variants. (provided by RefSeq, Oct 2016)

Protein domain

CD203-001 2638-002 2638-004

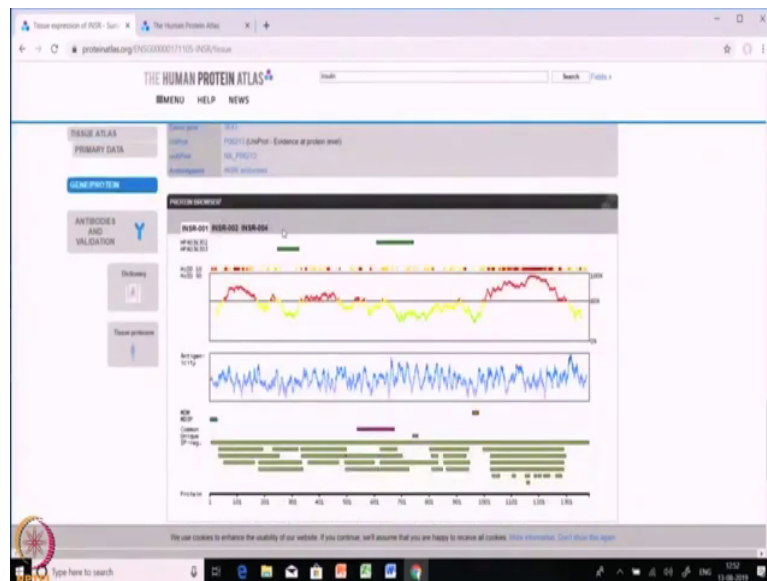
49435113
49435113

We use cookies to enhance the usability of our website. If you continue, we'll assume that you are happy to receive all cookies. [More information](#) [Don't show this again](#)

Type here to search

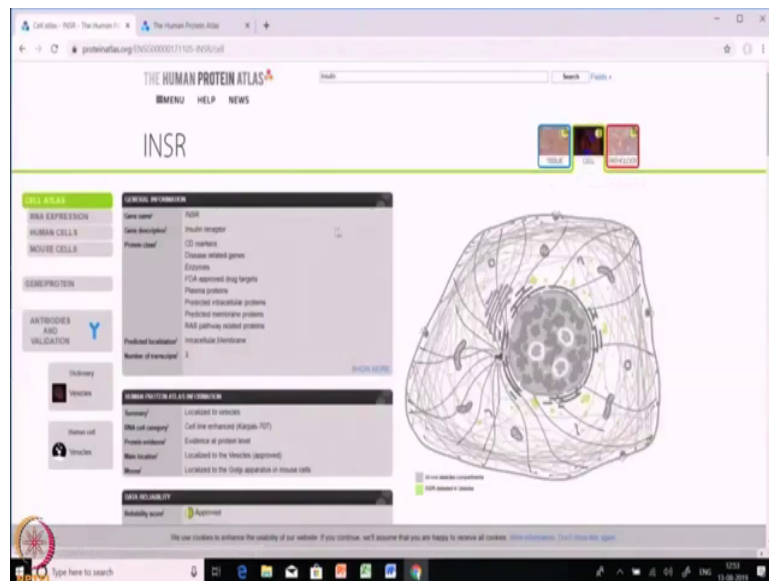
And finally, if we come down we will get more information regarding the gene.

(Refer Slide Time: 27:44)



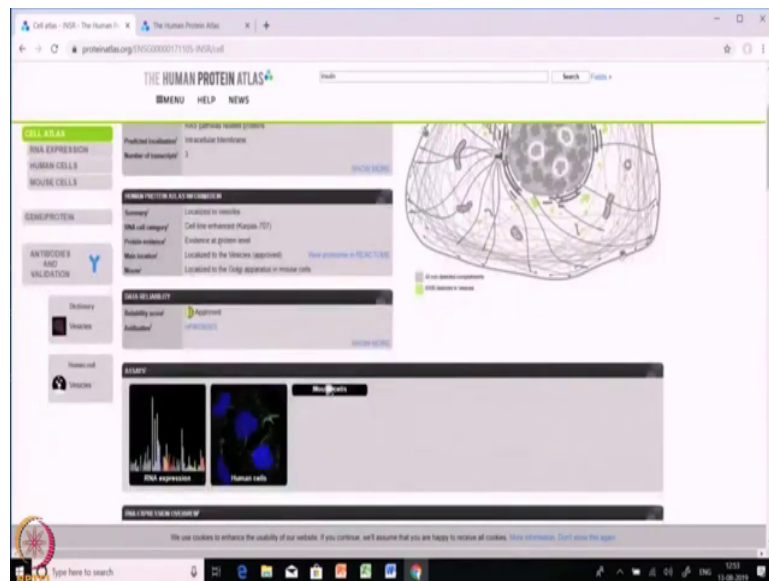
So, now what is protein browser? So, the protein browser displays the antigen location on target proteins and the features of the target protein; that tabs at the top of the protein view section can be used to switch between the different splice variant to which an antigen has been mapped. So, as you can see there are of information's available even in only a single tab that is tissue atlas of HPA.

(Refer Slide Time: 28:11)



So, now let us move to the next one that is the cell atlas. Cell atlas will give you more information regarding the localization of the protein. So, as you can see in the first tab it is given that a predicted localization of this protein is intracellular comma membrane.

(Refer Slide Time: 28:28)

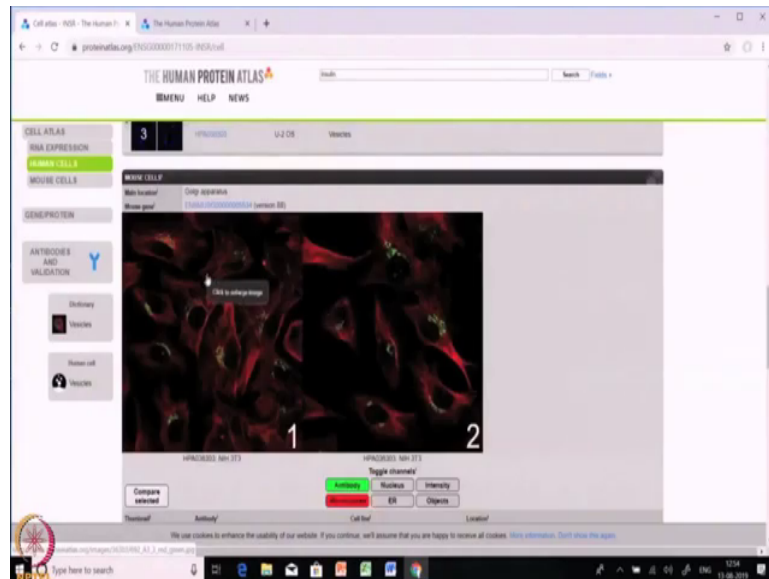


And apart from that, it has also given that what are the main location of the protein. And it is given that the location of the protein is in the vesicles which has already been approved; with the help of the indirect immunofluorescence microscopy image.

(Refer Slide Time: 28:40)

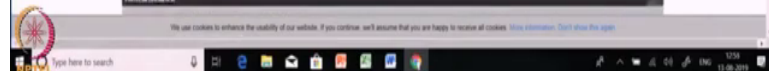


(Refer Slide Time: 28:45)

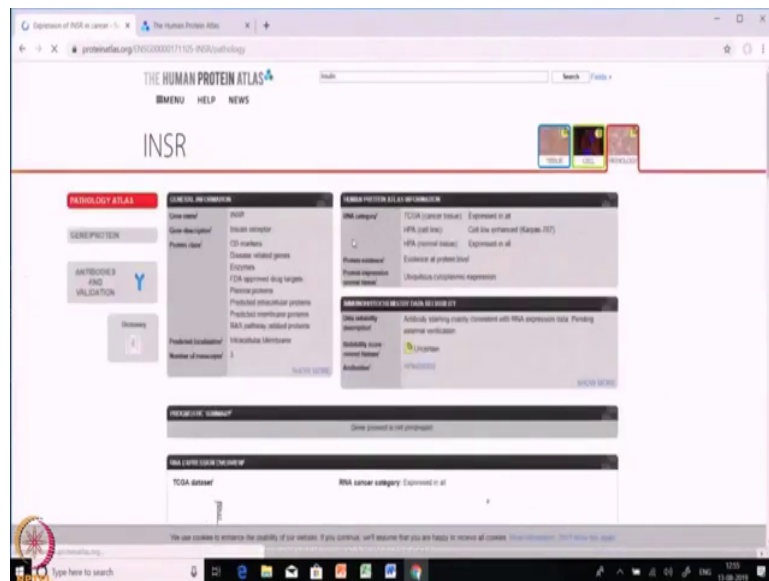


So, the same thing is also present in case of the mouse cell. And here the protein is found to be located in the Golgi apparatus.

So, like this you will if you explore you will understand and you will find more information regarding the protein in the cell atlas.

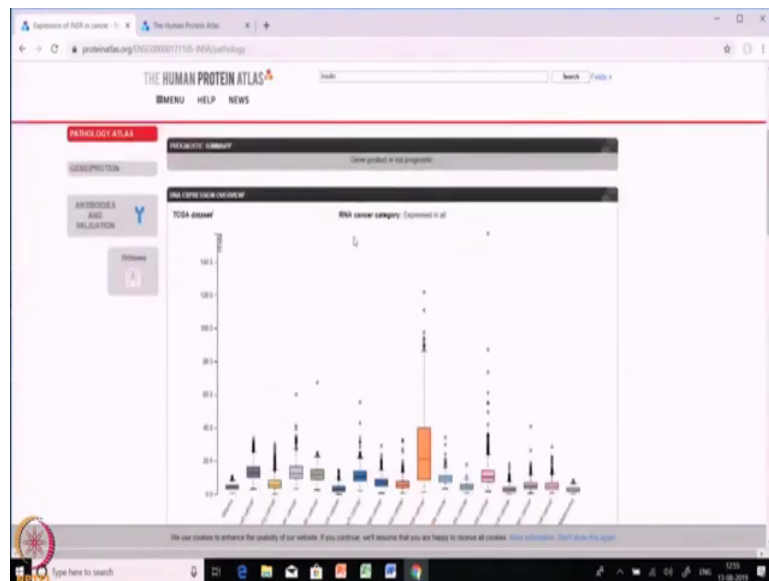


(Refer Slide Time: 29:06)



So, now the third one is the pathology atlas. The pathology atlas itself defines that it is based on the different diseases.

(Refer Slide Time: 29:12)



And as you can see, what is the status of this protein in different diseases in terms of RNA expression has been given over here.

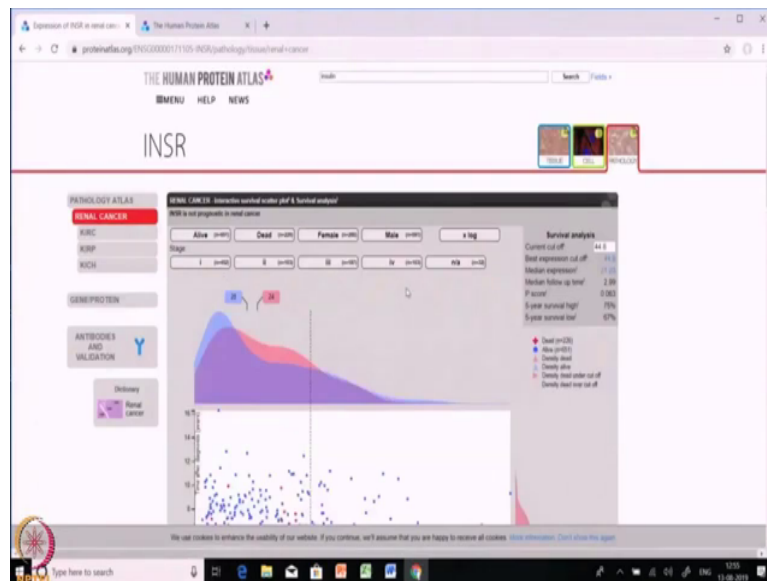
(Refer Slide Time: 29:26)

The screenshot displays the 'INSUR' (Integrative Network for Survival and Utility Research) database interface on The Human Protein Atlas website. The page is titled 'INSUR' and 'RENA CANCER: Genetically modified cancer (GEM) & Survival analysis'. It features a sidebar with navigation options: 'PROTEOMIC ATLAS', 'RENA CANCER', 'GENEPROTEIN', 'ANTIBODIES AND VALIDATION', and 'Dictionary'. The main content area shows a table of samples with columns for 'Sample', 'Description', and 'DPH'. The table lists 10 samples, including TCGA-B0-5980-01A, TCGA-B0-5710-01A, TCGA-CJ-4885-01A, TCGA-BP-4881-01A, TCGA-BP-4881-01A, TCGA-BP-4881-01A, TCGA-BP-4881-01A, TCGA-BP-4881-01A, TCGA-BP-4881-01A, and TCGA-BP-4881-01A. The table also includes a 'Survival analysis' section with a 'Survival analysis' button and a 'Survival analysis' table showing 'Survival analysis' results for 'Survival analysis'.

| Sample | Description | DPH |
|------------------|--|-------|
| TCGA-B0-5980-01A | 61 years, female, white, stage I, alive, 1380 days | 121.8 |
| TCGA-B0-5710-01A | 61 years, male, white, stage I, alive, 2430 days | 110.8 |
| TCGA-CJ-4885-01A | 62 years, female, white, stage I, alive, 1480 days | 97.7 |
| TCGA-BP-4881-01A | 64 years, male, white, stage I, alive, 1413 days | 84.2 |
| TCGA-BP-4881-01A | 61 years, female, white, stage I, alive, 7 days | 30.2 |
| TCGA-CJ-4885-01A | 38 years, male, white, stage I, alive, 1531 days | 80.6 |
| TCGA-BP-4881-01A | 47 years, male, white, stage I, alive, 1655 days | 80.6 |
| TCGA-BP-4881-01A | 50 years, male, white, stage I, alive, 953 days | 80.1 |
| TCGA-BP-4881-01A | 50 years, male, white, stage I, alive, 1329 days | 80.9 |
| TCGA-BP-4881-01A | 59 years, female, white, stage I, dead, 361 days | 87.2 |
| TCGA-CJ-4885-01A | 51 years, female, white, stage I, alive, 2480 days | 80.1 |
| TCGA-BP-4881-01A | 60 years, male, white, stage I, alive, 1000 days | 81.3 |

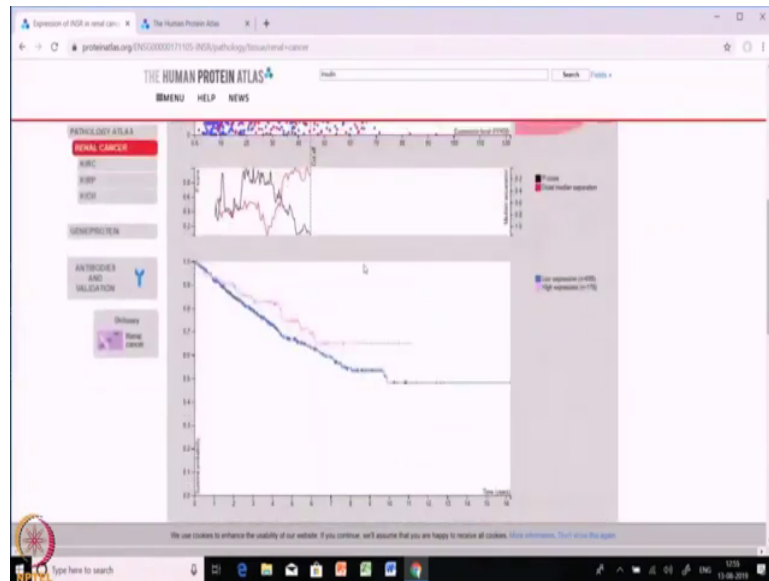
So, now if we click renal cancer, we will find that there is a lots of information available about the disease.

(Refer Slide Time: 29:29)

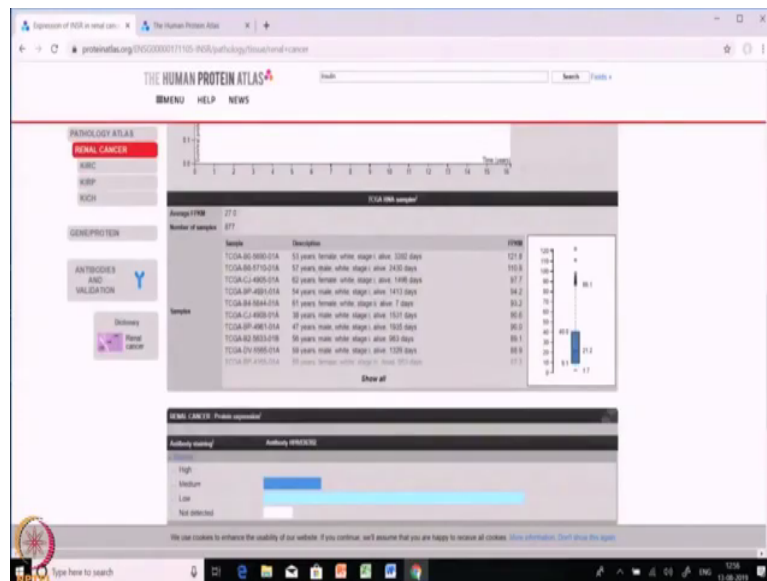


Here we can see that what is the number of patients that are available what is the number of patients that are alive. And what is the number of patients that is dead what is the sex ratio of the patient.

(Refer Slide Time: 29:47)

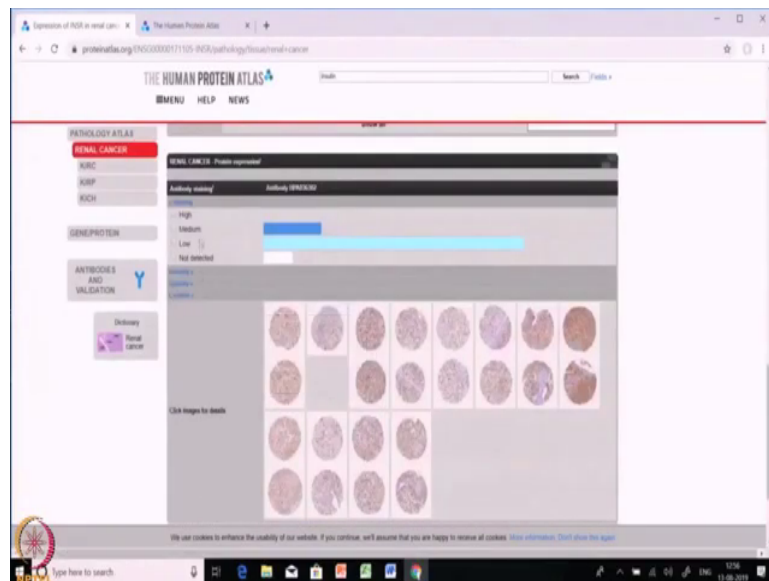


(Refer Slide Time: 29:50)



And everything over here is given even the survival rate curve is available. And the information of the patients in terms of their age, in terms of their survival rate and what is the status of the patient is also available.

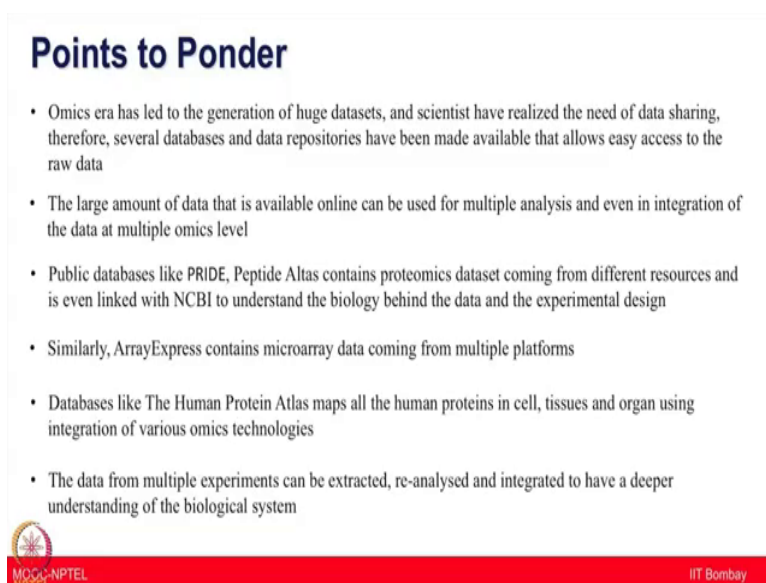
(Refer Slide Time: 30:07)



One of the important features of a human proteome atlas is most of the information will be we are self explanatory. If you just prove your put your cursor over the I present in the data set. So, as you can see that, if you want to know about this part just put your cursor over here and it will give you a explanation about that part.

So, as you can see that human proteome atlas contains so, much information of only one protein. So, now, as we talk a lot about different data and databases and there are many databases that we can explore and that contains different information.

(Refer Slide Time: 30:41)



Points to Ponder

- Omics era has led to the generation of huge datasets, and scientist have realized the need of data sharing, therefore, several databases and data repositories have been made available that allows easy access to the raw data
- The large amount of data that is available online can be used for multiple analysis and even in integration of the data at multiple omics level
- Public databases like PRIDE, Peptide Atlas contains proteomics dataset coming from different resources and is even linked with NCBI to understand the biology behind the data and the experimental design
- Similarly, ArrayExpress contains microarray data coming from multiple platforms
- Databases like The Human Protein Atlas maps all the human proteins in cell, tissues and organ using integration of various omics technologies
- The data from multiple experiments can be extracted, re-analysed and integrated to have a deeper understanding of the biological system

MOOC-NPTEL IIT Bombay

So, by now you know that there is huge amount of data, which is available in the public repositories and databases, which could be extracted and used for the analysis. There are many big research groups.

And a large funded programs, like human protein atlas The Cancer Genome Atlas or TCGA and there are various labs from the broad institute of MIT and Harvard. They are sharing their entire raw data into different databases. Also as they as they are imentioned all the journals are making it mandatory now to provide the raw data files. So, you have access to large number of you know very good quality data sets available. All you have to do is to extract the data and perform different type of analysis.

So, if you are familiar that you know for which type of data set what kind of analysis to be done, then you do not have to rely on donating your own data all the time. You are you can

always start even by sitting in a small college somewhere in the remote part of India or anywhere in the world, you can start your own experiments on your own computer. And then you can start coming up with a very very fascinating hypothesis which should be really you know transformational in nature.

Because now, you can see that you are working on the raw data and you are looking at some hypothesis which nobody has tested. Think about you know the cancer genome atlas what they do they provide almost thousands of cancer patient data set. They also provide lot of clinical data set. They also provide a lot of follow ups of these patients, which kind of you know drugs were given to these patients; what was their response which patient you know, showed the recurrence of the tumour.

So, now there are many questions which one could start looking at that what was the effect of a given drug? And now you are analysing the data based on just effect of a given drug, on one subtype of the cancer population. Or you are looking at which percentage of the patient you know they got they showed the recurrence of tumour you know recurred again.

And then which subtype they belong to, what kind of you know the genes were expressing that; can you associate anything linked to that with their patient survival or recurrence of tumour. So, many interesting thing can be done which is actual real research project, just on your own platform on your own laptops on your own system which does not require generation of data set.

So, I hope this kind of you know the information regarding the availability of databases and resources, as well as different tools available for doing analysis is going to be very practical and very useful for your own research. Another important point for obtaining these data sets are, you can always do metadata analysis you can start comparing the datasets obtained from Indian population versus Caucasian population within Indian population.

You know maybe something from the eastern region versus a western region, northern region. You can start looking at the demographic based analysis as well, which is otherwise not

possible you know meaning investigators always look for in certain samples of their region. But you can go trans-Atlantic you can go much you know pan Atlantic.

And you can start looking at data in a very comprehensive way right. And you can also start integrating data set you need not to only look at proteomics data from different regions. But why not you know it start integrating proteomics, metabolomics and transcriptomics data, along with genomic information to really you know get the systems information, which is otherwise not possible from the individual investigators.

So, I hope a lot exciting thing can be done, by you know in this day and age of you know computational driven field lot of you know things can be done without having access to the instrumentation and technology. So, I hope your understanding obtained from this course is not going to limit you, just to rely on the you know instruments and generate your own data.

But rather its not looking at the data in a bigger context it start looking at metadata analysis. And make biological sense of interesting question which you have always wanted to address. I hope you will be able to use some of these repositories in the databases in your own research.

Thank you.