Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 39
Bioinformatics solutions for 'Big Data' Analysis - I

Welcome to MOOC course on Introduction to Proteogenomics. Today's lecture will be delivered by Dr. Pratik Jagtap. He is Research Assistant Professor at Department of Biochemistry Molecular Biology and Biophysics at University of Minnesota, Minneapolis in USA. His current research interest includes developing analytical workflows using galaxy platform in the areas of proteomics, metaproteomics, proteogenomics and data independent acquisition data analysis. Dr. Jagtap is going to talk about Bioinformatics Solutions for Big Data Analysis.

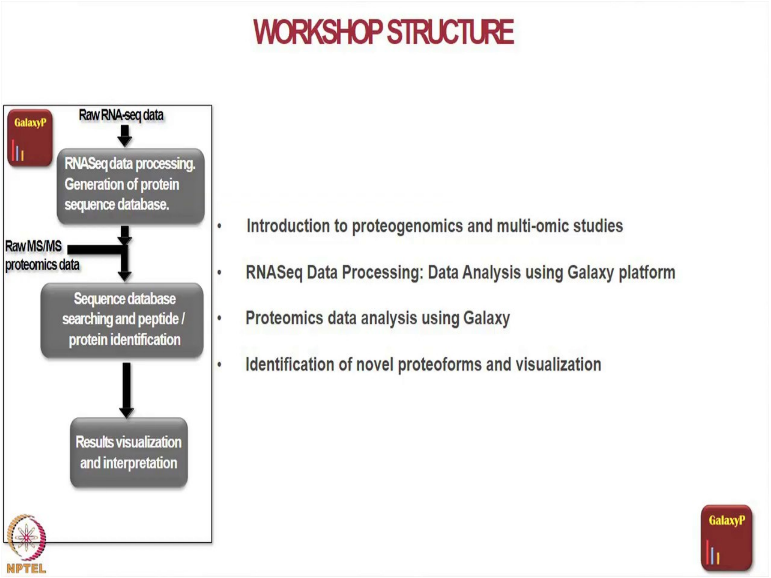
As you are aware big data being generated by variety of technology platforms such as MGS, Mass Spectrometry, RNA-Sequencing and many other different technology platforms are contributing towards big data analysis.

Dr. Jagtap will talk to us about how to analyze RNA-Seq files by first converting them into the protein FASTA files followed by locating it onto the genome. He will also provide a demo of galaxy software and tell us about its application to understand the multi-omics data set. So, let us welcome Doctor Pratik Jagtap to tell us about Bioinformatics Solutions for the Big Data Analysis.

I am going to talk about using galaxy platform for proteogenomics. So, we have been working in this field for last 5 years wherein the researchers at University of Minnesota some of them that you see here we have been trying to put in proteomics tools within galaxy framework and the idea was to not only have the galaxy framework or the tools to perform you know standard mass spectrometry searches but be able to do something more complicated like perform proteogenomics analysis.

We also work in the area of metaproteomics but that is not what we are going to cover here and one of the things that we have realized is as you are working in this field, it is very important to work with the user or a project to make these possible. So, the structure is going to be more like a demonstration of using galaxy for proteogenomics and I will obviously be talking about a few concepts as we go along.

(Refer Slide Time: 03:05)



WORKSHOP STRUCTURE

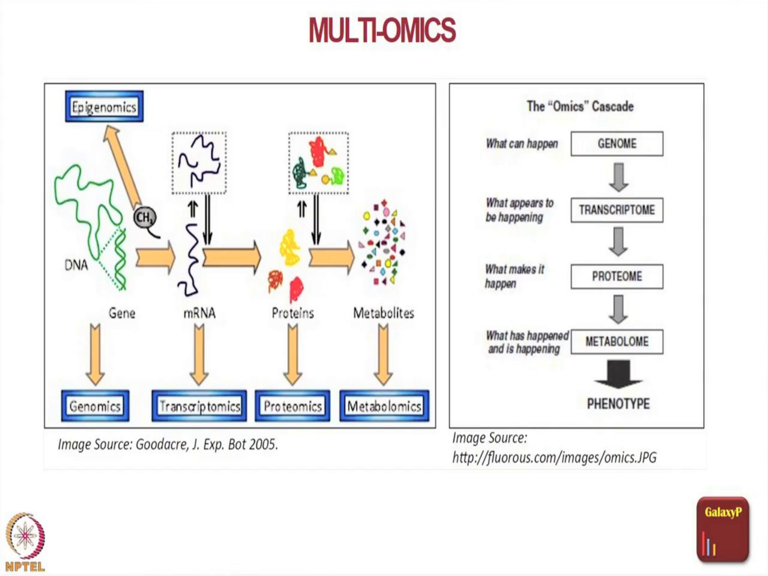
GalaxyP
Raw RNA-seq data
RNASeq data processing. Generation of protein sequence database.
Raw MS/MS proteomics data
Sequence database searching and peptide / protein identification
Results visualization and interpretation
NPTEL
GalaxyP

- Introduction to proteogenomics and multi-omic studies
- RNASeq Data Processing: Data Analysis using Galaxy platform
- Proteomics data analysis using Galaxy
- Identification of novel proteoforms and visualization

So, I am going to talk about Introduction to proteogenomics and multi-omic studies and I might actually now spend much time what I will mostly focus on the next three bullet points is if you have RNA-Seq data, how do you use the RNA-Seq data to convert into protein FASTA file and then use that protein FASTA file to search against your mass spectrometric data and then, eventually once you have peptides identified from that how do you make use of that to one look at the spectral quality.

Secondly, to localize that on a genome and then basically you know try to make some biological conclusions out of it.

(Refer Slide Time: 03:43)



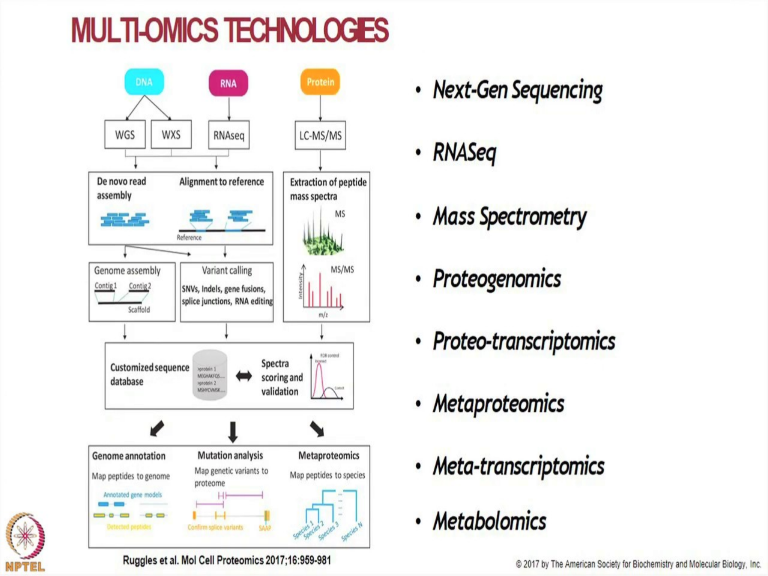
MULTI-OMICS

Image Source: Goodacre, J. Exp. Bot 2005.

Image Source: http://fluorous.com/images/omics.JPG

So, going through multi-omics you know each of these field has its own strengths transcriptomics, proteomics, genomics, metabolomics but the strength actually lies in making the best use of the features that are available in each of these and help you to answer the questions that you as a researcher I have put forth.

(Refer Slide Time: 04:09)



MULTI-OMICS TECHNOLOGIES

- Next-Gen Sequencing
- RNASeq
- Mass Spectrometry
- Proteogenomics
- Proteo-transcriptomics
- Metaproteomics
- Meta-transcriptomics
- Metabolomics

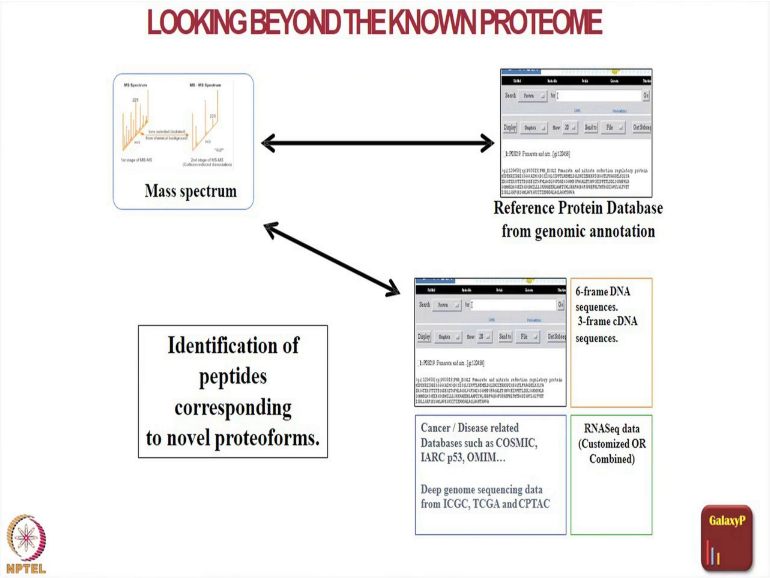
Ruggles et al. Mol Cell Proteomics 2017;16:959-981

© 2017 by The American Society for Biochemistry and Molecular Biology, Inc.

Again there are many technologies available, many newer coming up given the fact that there are lot more sensitive instruments as well as the instruments that can also have got really fast scan speeds. So, they are not only able to go deeper but also much more complex data sets can be handled by newer mass spec instruments which kind of helps you to approach the transcriptomics sensitivity for most of the analysis.

So, I will not cover much of this aspect except to say that you know because of the ability to not only have tools that work really well in each of these domains, we are also developing now tools, so that you can make correlations amongst various disciplines.

(Refer Slide Time: 04:57)



LOOKING BEYOND THE KNOWN PROTEOME

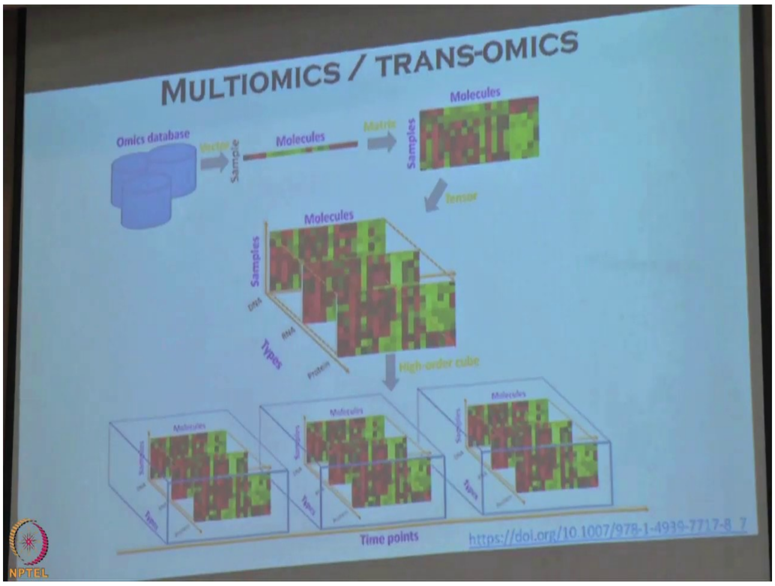
Mass spectrum
Reference Protein Database from genomic annotation
6-frame DNA sequences.
3-frame cDNA sequences.
Cancer / Disease related Databases such as COSMIC, IARC p53, OMIM…
Deep genome sequencing data from ICGC, TCGA and CPTAC
RNASeq data (Customized OR Combined)
Identification of peptides corresponding to novel proteoforms.
GalaxyP
NPTEL

So, this is I am sure this has been covered but I just wanted to reiterate the fact that if you have a mass spectrum and you have many mass spectra, generally if you have a reference protein database you end up identifying proteins which are annotated or of known proteins identified.

However, you can actually expand your number of identifications by if you are let us say searching it against you know what was earlier used as a 6-frame DNA sequence genomic DNA Sequence or even 3-frame cDNA Sequence but nowadays with the amount of RNA-Seq data that is available and ability to generate both RNA-Seq data as well as mass spectrometer data for the same sample one can actually use RNA-Seq data.

I will talk a little bit about that. The researchers have also used repositories like there is repository like cosmic and others which actually help you to just go and get data from somebody else’s research. So, you know it could be a protein FASTA file or it could be a RNA-Seq data from representative you know clinical samples and you could use that to search against your mass spectrometry data with the understanding that that particular data might actually not match completely because you might have some unique you know sequences that have been expressed in your sample and this all leads basically to identification of peptides which are corresponding to novel proteoforms which is what we think is proteogenomics is all about.

(Refer Slide Time: 06:31)

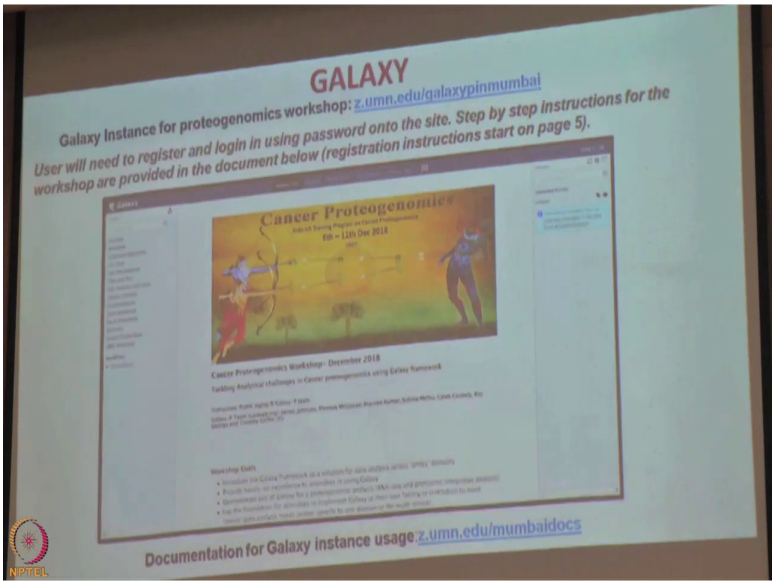


MULTIOMICS / TRANS-OMICS

https://doi.org/10.1007/978-1-4939-7717-8_7

So, again we talked about one data point comparison and that is obviously some is challenging but there are methods in place and people are doing that but then eventually the field is going to move and I am sure has already started moving in looking at time points of you know RNA-Seq data as well as proteomics data and then try to merge that, so that you can make a time dependent or temporal analysis of the expression of these RNA and proteins and that also means there is going to be a lot of data and a lot of analytical part that would be required.

(Refer Slide Time: 07:05)



GALAXY

Galaxy Instance for proteogenomics workshop: z.umn.edu/galaxypinmumbai

User will need to register and login in using password onto the site. Step by step instructions for the workshop are provided in the document below (registration instructions start on page 5).

Cancer Proteogenomics

6th – 11th Dec 2018

Cancer Proteogenomics Workshop- December 2018

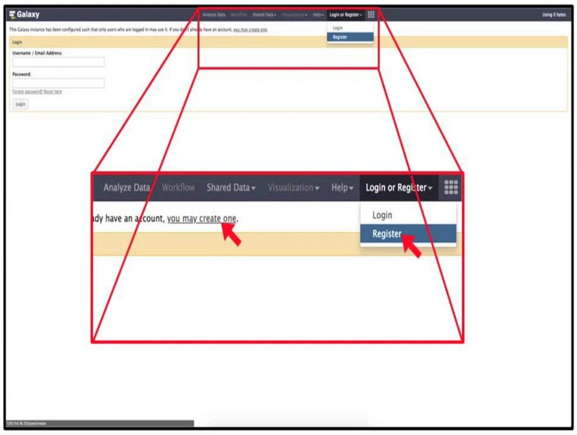
Documentation for Galaxy instance usage:z.umn.edu/mumbaidocs

So, what I am going to start with is there is this instance that we have set up. It is a galaxy instance and I will come back to what is galaxy and I will not encourage you to go on it right now because I am going to demonstrate on that particular instance, but I will definitely encourage you to go to that instance.

It is dot umn dot du slash galaxypinmumbai and it basically is a galaxy instance on which you can use step by step directions which have been provided here in this in this documentation and that you should be able to use you know use the instance like and I am going to use right now. So, let me take you through this.

(Refer Slide Time: 07:51)

REGISTER



This Galaxy instance has been configured such that only users who are logged in may use it. If you don't already have an account, you may create one.

Login

Username / Email Address:

Password:

Forgot password? Reset here

So, all you need to do is first go on to that website and register and all you need is a login and a password and you know and once you register, you basically would have to go on to this in this place called as histories; yeah.

Student: If I want to if I want the cDNA.

Yeah.

Student: Like a 3 frame cDNA sequence of whole genome. So, is any tool available on this?

Yes, so, there is a tool called GetORF that is available that can do that.

Student: We can give the whole genome.

You can take the cDNA, yes. So, you can go to ensemble for example ensemble has links to genomic DNA and cDNA.

Student: cDNA.

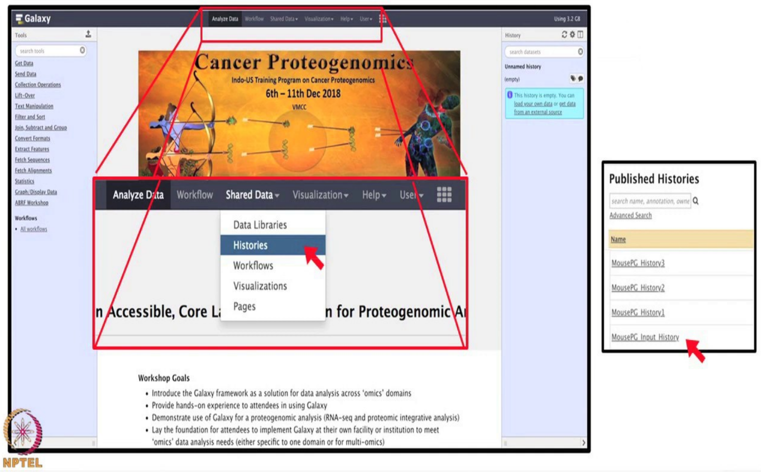
Take this download the cDNA on to galaxy use this GetORF tool and then, you should be able to use it ok.

So, and that if that fails and am just keeping this as a backup you can also go to this site it is z dot umn dot edu slash proteoenomics gateway and this is the document that goes along with it z dot umn dot edu slash pginnovember18. So, this was done last month. So, that is why that.

Anyway this if you go to Mumbai slides, you should be able to see all of that. So, you have to go to the site and get registered and anybody can register to this. This is again on a cloud instance in Indiana university and then what you have in this is A.

(Refer Slide Time: 09:23)

IMPORT HISTORY



This history is empty. You can load your own data or get data from an external source

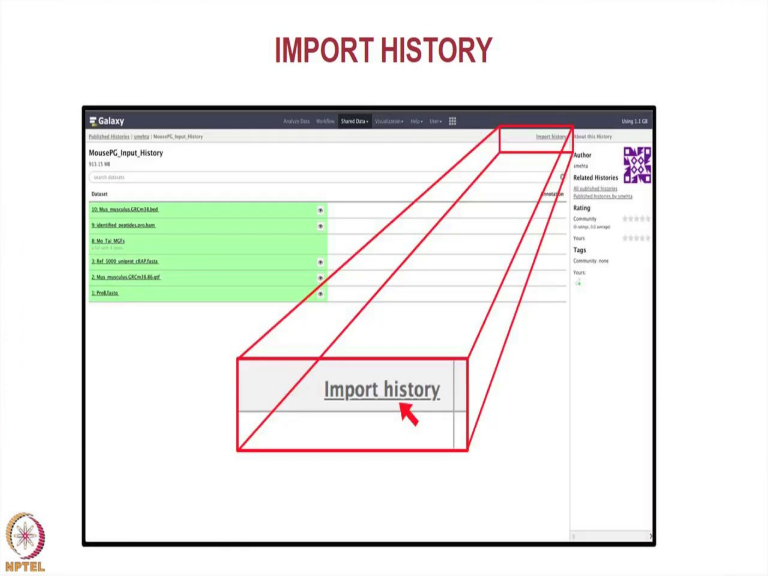
Workshop Goals

- Introduce the Galaxy framework as a solution for data analysis across 'omics' domains
- Provide hands-on experience to attendees in using Galaxy
- Demonstrate use of Galaxy for a proteogenomic analysis (RNA-seq and proteomic integrative analysis)
- Lay the foundation for attendees to implement Galaxy at their own facility or institution to meet 'omics' data analysis needs (either specific to one domain or for multi-omics)

So, galaxy is basically made up of you know there are tabs available at the it is; it is a web based platform. So, there are tabs available and you can go to what is called as a history and you can import the history there, ok. A history is basically a collection of your inputs or your any data that you have would have processed and I will talk a little bit about that a little later, ok.

So, just remember the word histories and workflows at this point and so in that you would actually have quite a few published histories and the first one would be a Mouse Proteogenomics Input History Data.

(Refer Slide Time: 10:03)



IMPORT HISTORY

Published Histories | smehta | MousePG_Input_History

MousePG_Input_History

913.15 MB

Dataset

10: Mus_musculus.GRCm38.bed

9: identified_peptides.pro.bam

8: Mo_Tai_MGFs

3: Ref_5000_uniprot_cRAP.fasta

2: Mus_musculus.GRCm38.86.gtf

1: ProB.fasta

Import history

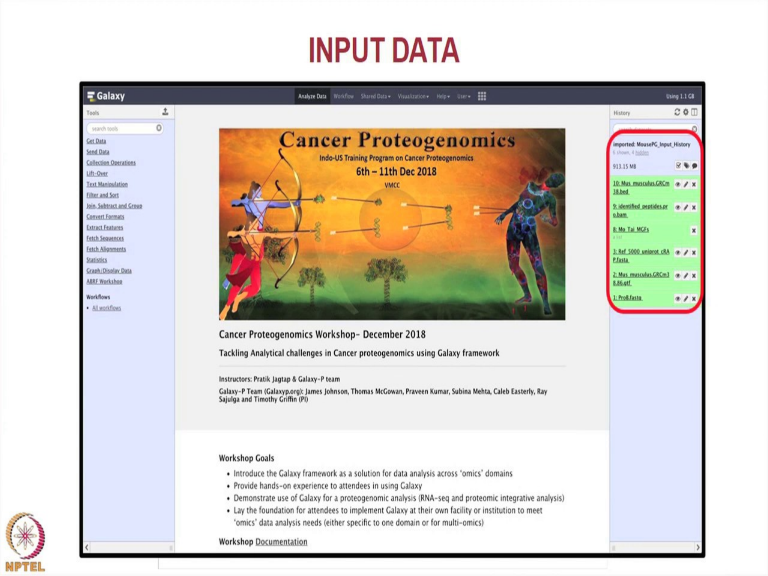
Author: smehta

Rating: Community (0 ratings, 0.0 average)

Tags: Community: none

And then once you go there, you should be able to import the history. So, you are basically bringing that history into your browser and what it does is actually gets you know uploaded onto your what is called as a history panel.

(Refer Slide Time: 10:21)



INPUT DATA

Cancer Proteogenomics Workshop- December 2018

Tackling Analytical challenges in Cancer proteogenomics using Galaxy framework

Instructors: Pratik Jagtap & Galaxy-P team

Galaxy-P Team (Galaxyp.org): James Johnson, Thomas McGowan, Praveen Kumar, Subina Mehta, Caleb Easterly, Ray Sajulga and Timothy Griffin (PI)

Workshop Goals

- Introduce the Galaxy framework as a solution for data analysis across ‘omics’ domains
- Provide hands-on experience to attendees in using Galaxy
- Demonstrate use of Galaxy for a proteogenomic analysis (RNA-seq and proteomic integrative analysis)
- Lay the foundation for attendees to implement Galaxy at their own facility or institution to meet ‘omics’ data analysis needs (either specific to one domain or for multi-omics)

Workshop Documentation

imported: MousePG_Input_History

6 shown, 4 hidden

913.15 MB

10: Mus_musculus.GRCm38.bed

9: identified_peptides.pro.bam

8: Mo_Tai_MGFs

3: Ref_5000_uniprot_cRAP.fasta

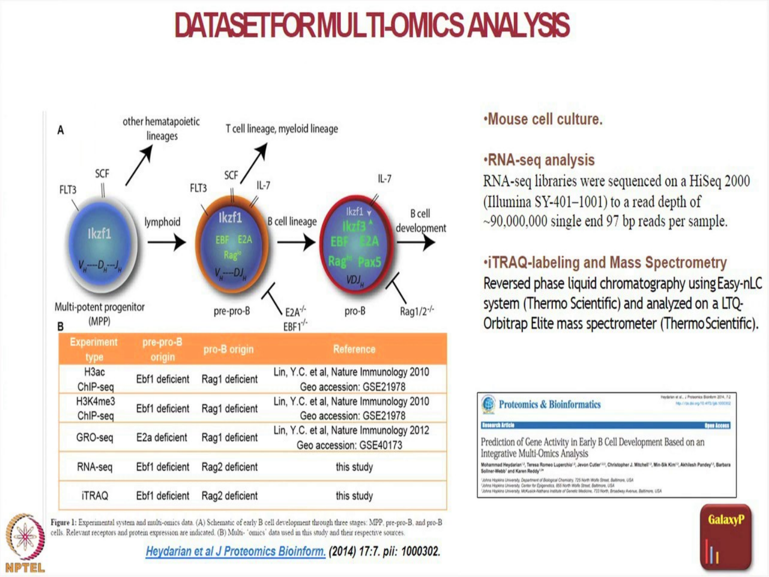
2: Mus_musculus.GRCm38.86.gtf

1: ProB.fasta

So, the panel on the right here is the history panel. On all you have there are these input files, there is a protein FASTA file which we are going to use for database search, there is a *Mus musculus* a GTF file, right and that is used for estimating genome coordinates. There is another protein FASTA file and then there are MGF files that you have generated from your from your mass spec data and sorry this is actually a FASTQ file, not FASTA file.

So, the FASTA file was somewhere here and then you also have a BAM file and a byte file and I will talk about that why we need that later, right. So, these are the inputs that one uses.

(Refer Slide Time: 11:07)



DATASET FOR MULTI-OMICS ANALYSIS

Experiment type	pre-pro-B origin	pro-B origin	Reference
H3ac ChIP-seq	Ebf1 deficient	Rag1 deficient	Lin, Y.C. et al, Nature Immunology 2010 Geo accession: GSE21978
H3K4me3 ChIP-seq	Ebf1 deficient	Rag1 deficient	Lin, Y.C. et al, Nature Immunology 2010 Geo accession: GSE21978
GRO-seq	E2a deficient	Rag1 deficient	Lin, Y.C. et al, Nature Immunology 2012 Geo accession: GSE40173
RNA-seq	Ebf1 deficient	Rag2 deficient	this study
iTRAQ	Ebf1 deficient	Rag2 deficient	this study

Figure 1: Experimental system and multi-omics data. (A) Schematic of early B cell development through three stages: MPP, pre-pro-B, and pro-B cells. Relevant receptors and protein expression are indicated. (B) Multi- 'omics' data used in this study and their respective sources.

Heydarian et al J Proteomics Bioinform. (2014) 17:7. pii: 1000302.

- Mouse cell culture.
- RNA-seq analysis
  RNA-seq libraries were sequenced on a HiSeq 2000 (Illumina SY-401–1001) to a read depth of ~90,000,000 single end 97 bp reads per sample.
- iTRAQ-labeling and Mass Spectrometry
  Reversed phase liquid chromatography using Easy-nLC system (Thermo Scientific) and analyzed on a LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific).

Proteomics & Bioinformatics

Heydarian et al., J Proteomics Bioinform 2014, 7:2
http://dx.doi.org/10.4172/jpb.1000302

Research Article Open Access

Prediction of Gene Activity in Early B Cell Development Based on an Integrative Multi-Omics Analysis

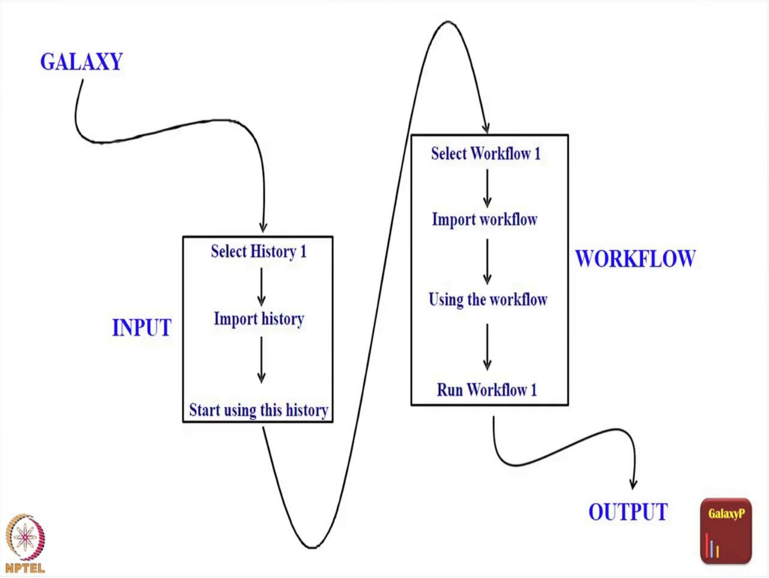
Mohammad Heydarian1,2, Teresa Romeo Luperchio1,2, Jevon Cutler1,2,3, Christopher J. Mitchell1,3, Min-Sik Kim1,3, Akhilesh Pandey1,3, Barbara Sollner-Webb1 and Karen Reddy1,2*

1Johns Hopkins University, Department of Biological Chemistry, 725 North Wolfe Street, Baltimore, USA
2Johns Hopkins University, Center for Epigenetics, 855 North Wolfe Street, Baltimore, USA
3Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, 733 North, Broadway Avenue, Baltimore, USA

And the data set used here and this is just to mention that this is representative data which was published in 2014.

It is a database about B-cell development and the researchers were interested in comparing two different types of B-cells in its development and we basically they had RNA-Seq data as well as proteomics data for that.

(Refer Slide Time: 11:29)



And we have basically taken a part of this to demonstrate the use of galaxy, ok. So, galaxy is the interface, the web based interface you can select the history. So, we are selecting the input file. So, remember the 5 files that I talked about. Earlier these 5 files that input files and then you can once you import the history, you can start using the history which means your history becomes active and then you select the workflow.

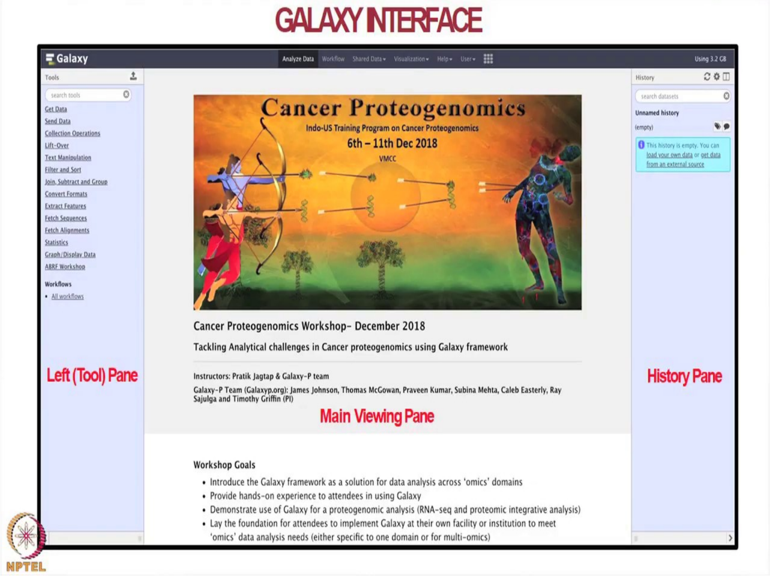
So, what is the work flow? Workflow is basically a set of tools that one would use in sequence, so that he can process the data. So, just imagine you have your input files, right and one of the input files needs to be converted from one format to another. So, it would use one tool.

So, you know you asked about converting cDNA into a three frame translation. So, there would be one tool but the process does not end there right. You want to use that in your next step and so maybe adding contaminant sequences to this would be the next step and so there could be a tool which could upload those contaminant sequences and then there will be tool which will merge them together and so on and so forth.

So, this basically becomes a workflow wherein you are not only taking one tool and running it and going to the other one, but you are taking the input and you can run this workflow of multiple tools, right.

So, you take select a workflow from one of the tabs, you import the workflow and you start using the workflow and once you run the Workflow, you get an output that you that you want. So, a workflow could be as small as you know two tools or it could be as large as 20 or 30 tools right depending upon how complex analysis you would like to do.

(Refer Slide Time: 13:15)



GALAXY INTERFACE

This history is empty. You can load your own data or get data from an external source

Cancer Proteogenomics Workshop- December 2018

Tackling Analytical challenges in Cancer proteogenomics using Galaxy framework

Instructors: Pratik Jagtap & Galaxy-P team

Galaxy-P Team (Galaxyp.org): James Johnson, Thomas McGowan, Praveen Kumar, Subina Mehta, Caleb Easterly, Ray Sajulga and Timothy Griffin (PI)

Workshop Goals

- Introduce the Galaxy framework as a solution for data analysis across 'omics' domains
- Provide hands-on experience to attendees in using Galaxy
- Demonstrate use of Galaxy for a proteogenomic analysis (RNA-seq and proteomic integrative analysis)
- Lay the foundation for attendees to implement Galaxy at their own facility or institution to meet 'omics' data analysis needs (either specific to one domain or for multi-omics)

Left (Tool) Pane

Main Viewing Pane

History Pane

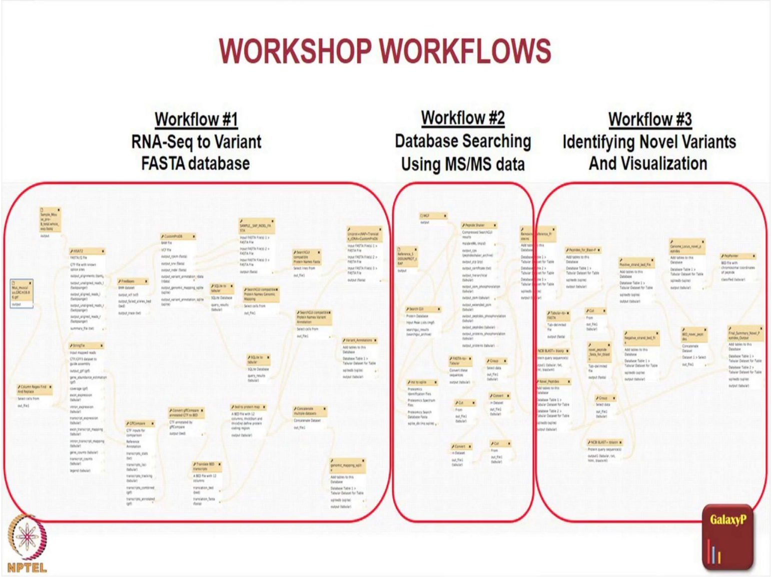
So, I will come back to the Galaxy Interface. So, this is the galaxy interface and if you go to z dot umn dot edu slash galaxy p in mumbai you should be able to come to this. So, a few things to kind of know about galaxy; one is on the left side we have what is called is a tool pane in the sense there are tools that somebody has developed and implemented in galaxy and that comes with a. So, let us say there is a tool like GetORF and it would have a place to show an input file.

So, it is basically a GUI interface for command line interface tool right. So, if you have let us say there is a tool, it gives you where to have an input, what are the parameters that need to be used. So, once let us say you have an input file here and you use that tool, the processed output will basically get added on in your history. So, input file data processed process data right.

So, you kind of start building a history and that is why this is called a history play pane when you starting with the input and generating input file and the central pane is a place where you can view your data. You can look at the parameters and you know there are other things that you can do in the central pane.

So, that is that is about galaxy interface.

(Refer Slide Time: 14:33)



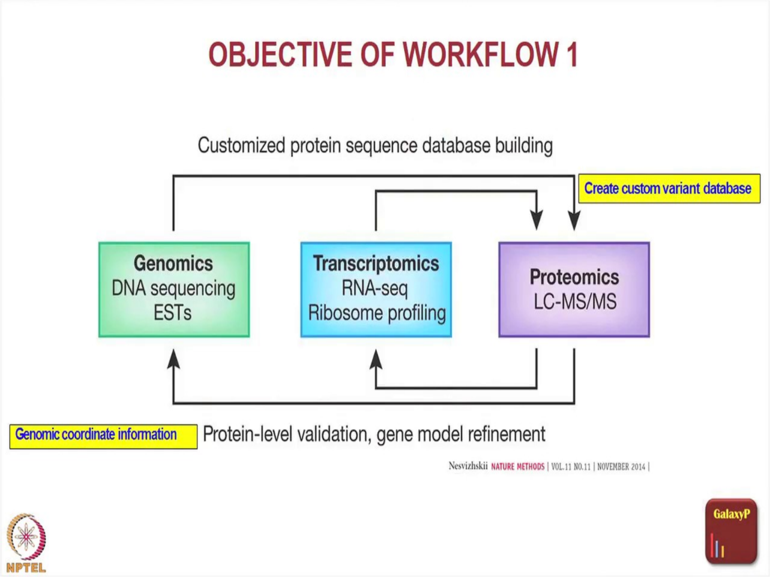
WORKSHOP WORKFLOWS

Workflow #1
RNA-Seq to Variant FASTA database
Workflow #2
Database Searching Using MS/MS data
Workflow #3
Identifying Novel Variants And Visualization
GalaxyP
NPTEL

In this particular demonstration that I wanted to show you we are going to cover RNA-Seq to variant FASTA database conversion. So, if you have RNA-Seq data how do you convert into a protein FASTA file, then take that protein FASTA file and search it against you know against that protein, against the mass spectrometry data and once you have identified your peptides, you can process them to either visualize this spectra or even identify.

The localization on the genome right and this all is possible because somebody has worked hard to put this tool into galaxy and also put it into work flows. So, each of this is a workflow. So, if you if you observe closely once you go and look at the slides, you will see each of this is a workflow which is kind of connected to each other right.

(Refer Slide Time: 15:25)



OBJECTIVE OF WORKFLOW 1

Customized protein sequence database building

Create custom variant database

Genomics
DNA sequencing
ESTs

Transcriptomics
RNA-seq
Ribosome profiling

Proteomics
LC-MS/MS

Genomic coordinate information

Protein-level validation, gene model refinement

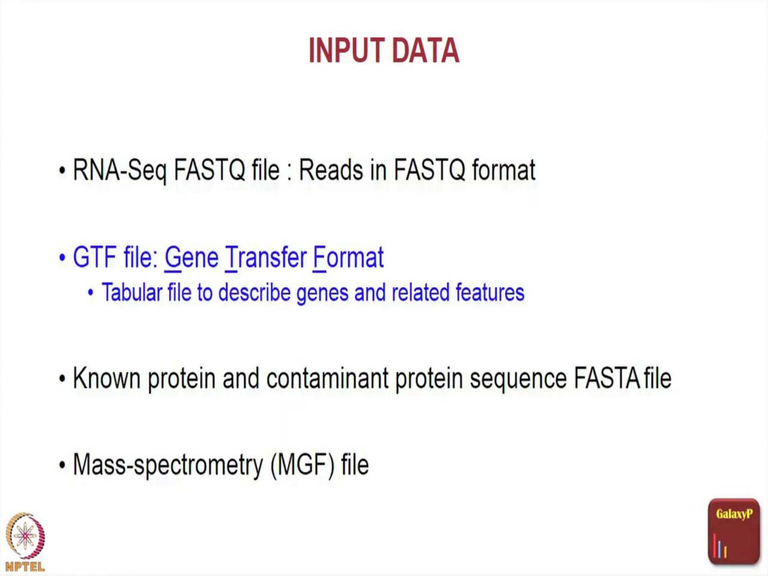
Nesvizhskii NATURE METHODS | VOL.11 NO.11 | NOVEMBER 2014 |

So, the objective of the workflow is to basically show that if you have genomics data, you can generate a customized database to generate a database that you can use it for your proteomics experiment or you can also use RNA-Seq data to do that and then once you get the data from that, you should be able to modify your gene mode.

You might actually find that there are some peptides that have been identified in regions that you know you would not have found earlier because of and because of some genomic rearrangement you have found them. As a result of even it could be just a single amino acid variant or it could be an insertion or deletion.

So, what is this work flow? First workflow which takes an RNA-Seq data and generates a protein FASTA file doses it generates a protein FASTA file, but it also generates a genomic mapping information, right. So, this is the input file that I described earlier.

(Refer Slide Time: 16:25)



INPUT DATA

- RNA-Seq FASTQ file : Reads in FASTQ format
- GTF file: Gene Transfer Format
  - Tabular file to describe genes and related features
- Known protein and contaminant protein sequence FASTA file
- Mass-spectrometry (MGF) file

And it has basically the FASTQ file which comes from your RNA-Seq data, right. There is a GTF file which is a Gene Transfer Format file which basically has genes as well as genome coordinates and which chromosome it comes from and so on and so forth. So, this basically helps you to connect your protein FASTA file or the accession protein accession numbers to your genome coordinates. We also have a known protein FASTA file generally from UniProt that we use and then we have the mass spectrometry files. So, the MGF files, right.

So, this GTA file is also available in on ensemble. Ensemble has a GTF file for that particular organism. So, you should be able to get that. So, this is something that you generate through experiment. This is something you generate through experiment. This is publicly available. UniProt and this is also publicly available. It is gets updated I think every 3 to 6 months if you are doing meta transcriptomics, if you are doing microbiome analysis or if you doing a two organism or you are talking about contamination.

Student: for some contaminants

Yeah.

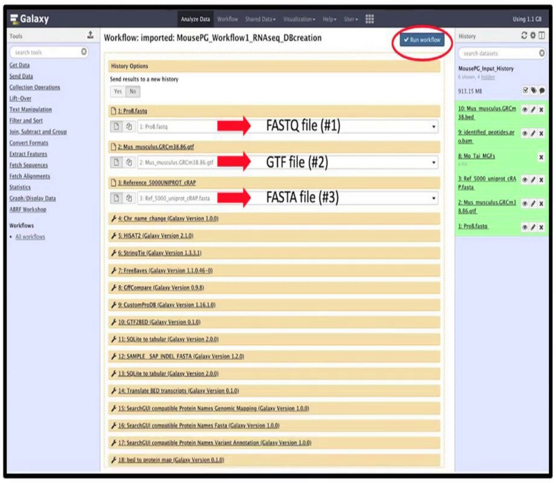
Student: We can use this pipeline or we can post convert it. So, no, you can still use it. You might need to do some QC filtering or you might you know. So, you are going to do some amount of mapping on to a genome to only select those sequences that are of interest to you but if you are doing a multi organism analysis, then obviously you want to retain all of that. Is does that answer the question.

Student: Yeah thank you

Sure. So, again we start with the sorry we start with the input files, the 5 files that I talked about and then we basically start using the first workflow which is the database creation workflow and I will perhaps use this time you know. So, this is where you can go to workflows you know once you go to that site and download this workflow again import the workflow.

(Refer Slide Time: 18:31)

SELECTING INPUT FILES TO RUN A WORKFLOW



Workflow: imported: MousePG_Workflow1_RNAseq_DBcreation

History Options

Send results to a new history

Yes No

1: ProB.fastq

1: ProB.fastq

FASTQ file (#1)

2: Mus_musculus.GRCm38.86.gtf

2: Mus_musculus.GRCm38.86.gtf

GTF file (#2)

3: Reference_5000UNIPROT_cRAP

3: Ref_5000_uniprot_cRAP.fasta

FASTA file (#3)

4: Chr_name_change (Galaxy Version 1.0.0)

5: HISAT2 (Galaxy Version 2.1.0)

6: StringTie (Galaxy Version 1.3.3.1)

7: FreeBayes (Galaxy Version 1.1.0.46-0)

8: GffCompare (Galaxy Version 0.9.8)

9: CustomProDB (Galaxy Version 1.16.1.0)

10: GTF2BED (Galaxy Version 0.1.0)

11: SQLite to tabular (Galaxy Version 2.0.0)

12: SAMPLE_SAP_INDEL_FASTA (Galaxy Version 1.2.0)

13: SQLite to tabular (Galaxy Version 2.0.0)

14: Translate BED transcripts (Galaxy Version 0.1.0)

15: SearchGUI compatible Protein Names Genomic Mapping (Galaxy Version 1.0.0)

16: SearchGUI compatible Protein Names Fasta (Galaxy Version 1.0.0)

17: SearchGUI compatible Protein Names Variant Annotation (Galaxy Version 1.0.0)

18: bed to protein map (Galaxy Version 0.1.0)

MousePG_Input_History

6 shown, 4 hidden

913.15 MB

10: Mus_musculus.GRCm38.bed

9: identified_peptides.pro.bam

8: Mo_Tai_MGFs

a list

3: Ref_5000_uniprot_cRAP.fasta

2: Mus_musculus.GRCm38.86.gtf

1: ProB.fastq

Start using the workflow in wherein you can say run.

So, once you import this work flow, it actually shows up in your workflow list of workflows and then what it shows you here and I really wish I could have shown you on this you know in the on the screen, but it basically gives you three input files to select to run your workflow.

So, what you see here is you know FASTQ file as the first input, GTF file is the second input, FASTA file as a third input and then there are these series of tools that are used ah, so that he can you can convert into a protein FASTA file eventually right.

Student: Sir, is this tool is freely available or it is for just sometime like because I am saying this is slash Mumbai.

Yeah.

Student: Like GTF.

Yeah.

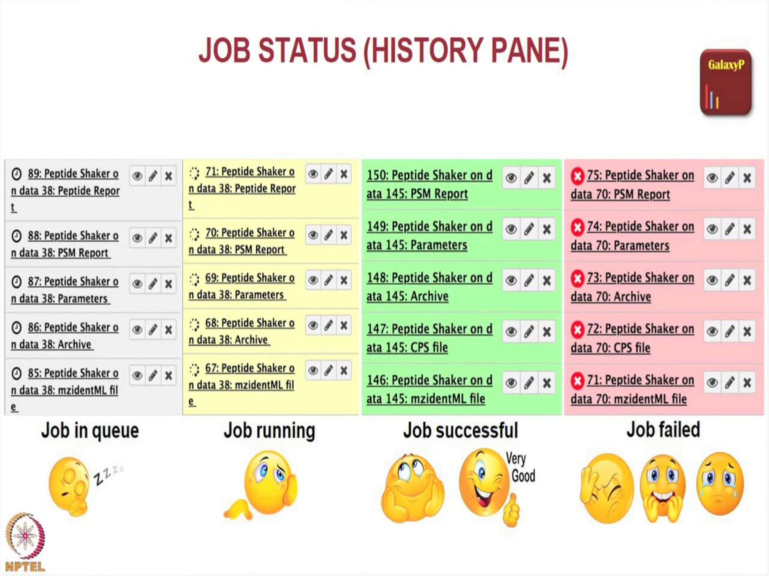
Student: So, is this the demo version for us or it's freely available and later on also we can use.

I mean we can keep this available for 3 months, but I mentioned there is another website called proteogenomics gateway, right z dot the slides as well that could be something you could use that would be available for much longer time.

Student: Ok not to the galaxy.

No both are galaxy instances. So one is a galaxy instance, for particular audience with the workshop and we will keep it going for 3 months but you can also try the other one, which will be available longer. And the reason we wanted to put in somethings more specific to this workshop but yeah both have got very similar content. once you have these inputs, you click run. So, imagine there is a button down here for people at the back maybe you cannot see, but if you click the button run here, it runs. It starts running.

(Refer Slide Time: 19:55)



JOB STATUS (HISTORY PANE)

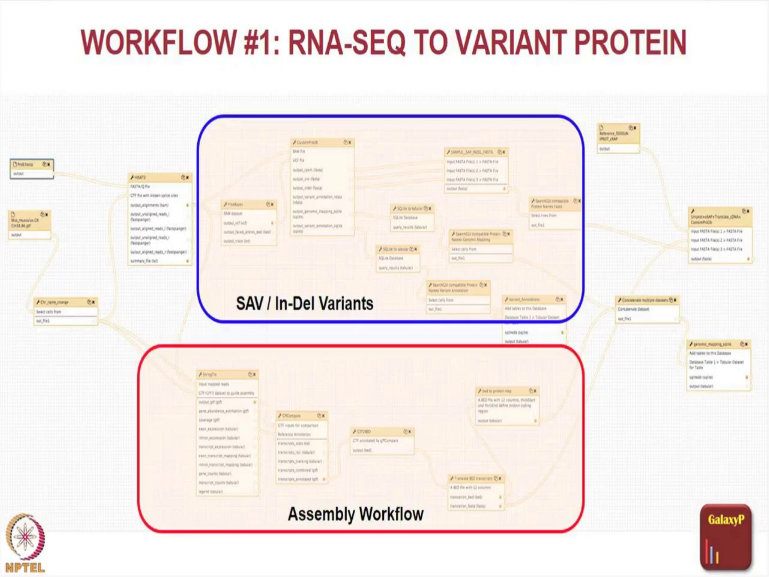
Job in queue	Job running	Job successful	Job failed
89: Peptide Shaker on data 38: Peptide Report	71: Peptide Shaker on data 38: Peptide Report	150: Peptide Shaker on data 145: PSM Report	75: Peptide Shaker on data 70: PSM Report
88: Peptide Shaker on data 38: PSM Report	70: Peptide Shaker on data 38: PSM Report	149: Peptide Shaker on data 145: Parameters	74: Peptide Shaker on data 70: Parameters
87: Peptide Shaker on data 38: Parameters	69: Peptide Shaker on data 38: Parameters	148: Peptide Shaker on data 145: Archive	73: Peptide Shaker on data 70: Archive
86: Peptide Shaker on data 38: Archive	68: Peptide Shaker on data 38: Archive	147: Peptide Shaker on data 145: CPS file	72: Peptide Shaker on data 70: CPS file
85: Peptide Shaker on data 38: mzidentML file	67: Peptide Shaker on data 38: mzidentML file	146: Peptide Shaker on data 145: mzidentML file	71: Peptide Shaker on data 70: mzidentML file

And what it does basically is it starts adding these outputs from each of these tools above this you know it starts building that history. Now when it starts basically it is grey in color. So, the job is in queue right. When the job starts running, it is it turns yellow and when the job is actually successful, it turns green.

So, each of these would have some output generated from that work flow. It is generally bad news. If it is red which means there was some error that was generated very early in your analysis and that is when you go back to your developer or your infrastructure specialist to ensure that you know something wrong did not go on the site of the network or the tool version and so on and so forth.

But just let us imagine that this is how it works, right. I mean in the sense let us let us hope that in our case we get these outputs and we are actually tested at least for this data set multiple times to ensure that that is how it works then we basically. So, let us talk about this particular workflow in general, ok.

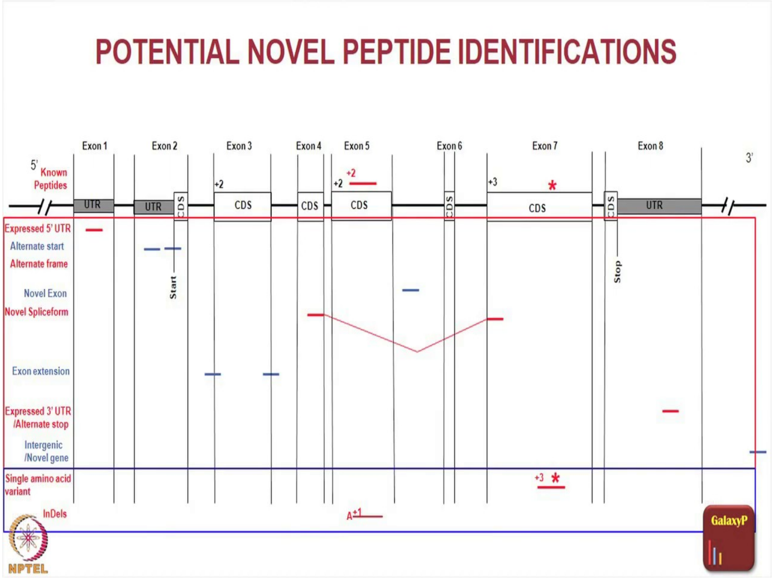
(Refer Slide Time: 21:05)



WORKFLOW #1: RNA-SEQ TO VARIANT PROTEIN

If you look at that workflow and expand it basically is made up of two parts. It starts with the FASTQ files that I talked about, but it also has GTF file and other files here. The third input that we use here the FASTA file here thank you and then it can be it is divided into two different groups and I will I will show you the details of this later, but the first one basically looks at single amino acid variants as well as In-Del variants. Well the second one basically looks at you know junctions novel junctions and so on and so forth.

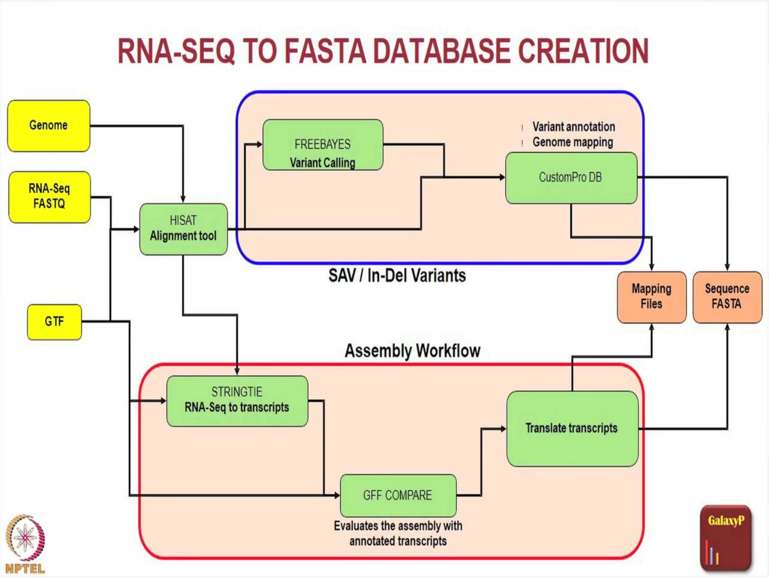
(Refer Slide Time: 21:41)



POTENTIAL NOVEL PEPTIDE IDENTIFICATIONS

So, well these are just different kind of variants that you can identify. For example, a single amino acid variants and In-Dels can be identified by this first workflow while the second workflow identifies rest of them here.

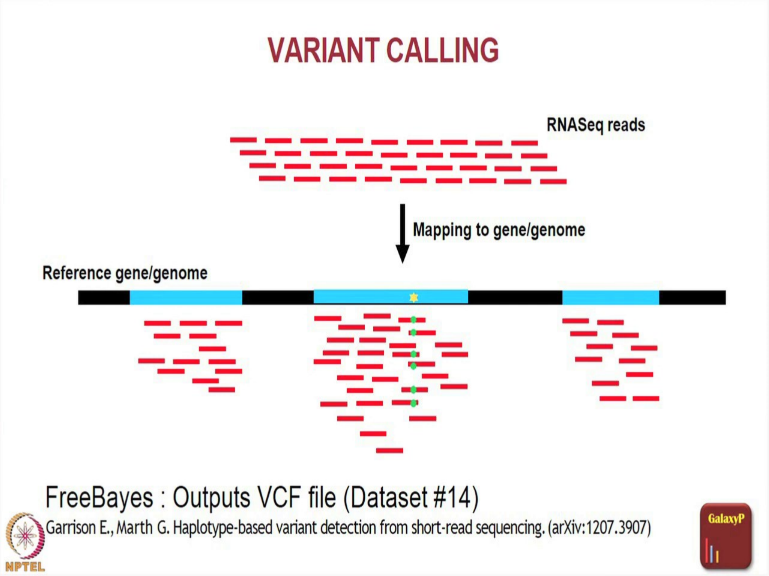
(Refer Slide Time: 21:57)



RNA-SEQ TO FASTA DATABASE CREATION

So, if you look at these details and hopefully this is a little more clearer, you have the FASTQ files, you have the genome coordinates and you have the GTF files, right.

(Refer Slide Time: 22:19)



VARIANT CALLING

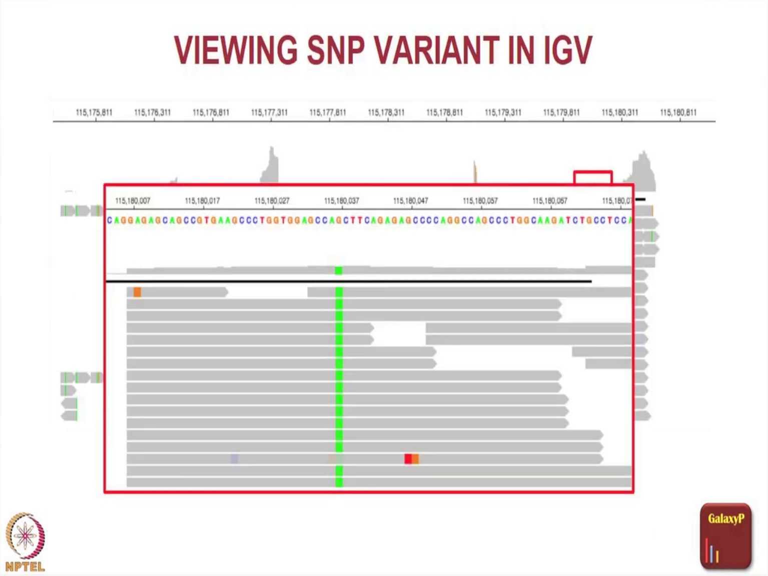
FreeBayes : Outputs VCF file (Dataset #14)

Garrison E., Marth G. Haplotype-based variant detection from short-read sequencing. (arXiv:1207.3907)

So, the initially what the first tool that is used is it uses the RNA-Seq data. HiSAT is the tool that is used to align your RNA-Seq data right. So, let us say these are all the RNA-Seq files, it maps to the gene or genome and then you can you know you can see where your RNA-Seq reads for and based on these now.

You can you can perform variant calling and that is where the second tool which is freebase. So, this tool here takes in all these aligned files and then starts finding something which is different than your reference genome, right. So, it is trying to find a variant in your sequences right, so that the green dots that you can see here that that is where it starts identifying those.

(Refer Slide Time: 23:03)

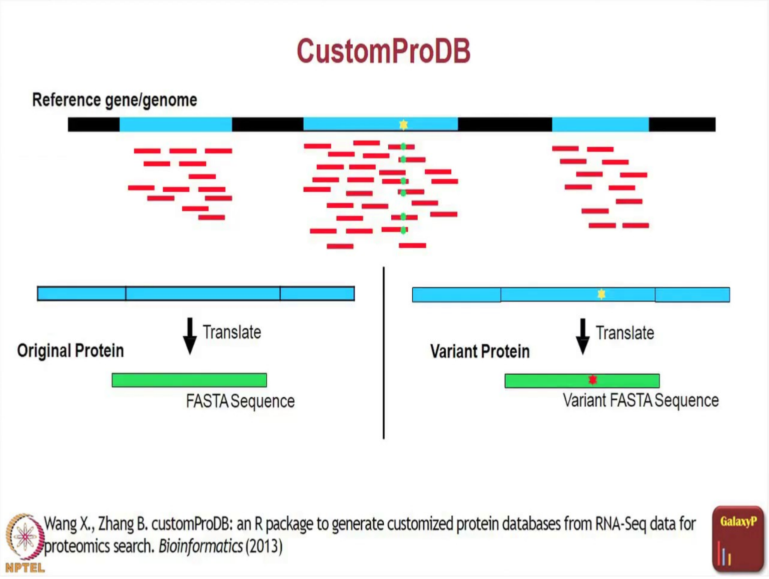


VIEWING SNP VARIANT IN IGV

And then so for example if you are looking at this I am basically going to focus on this region here and as you can see here this is your these are your RNA-Seq reads and you can see that the sequence in the reference genome has got a G in it right while in and you see a green Adenosine there, right. So, you can see that in the RNA-Seq data we actually finding a variant. Now that is what the tool does is it kind of captures this information, right.

So, the HiSAT aligns these RNA-Seq data freebase finds the variants in it.

(Refer Slide Time: 23:47)



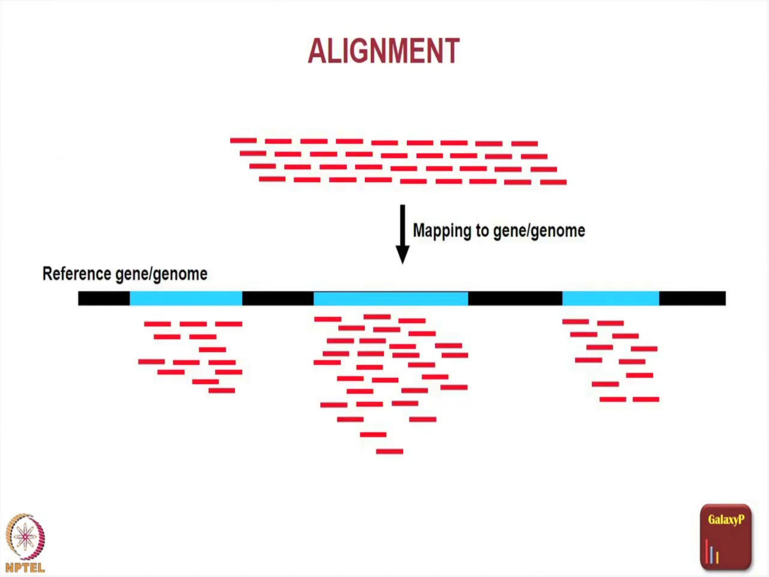
CustomProDB

Wang X., Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* (2013)

And then the next tool which is CustomProDB the one that and this one here basically looks at the reference genome these variants which have been identified by freebase and then it takes in this and then, translates not only the original sequence if it is you know if there is no variation, it just identifies translates that or it can also translate the variant sequence and it kind of puts that in the accession number that you know this is the variation along with the genomic coordinates.

So, you have all the information right there, right; and so, that is this part that we talked about right. So, you have it identifies your single amino acid variants as well as in Del variants. Now this part here is basically your part wherein you are identifying the changes in the junction and so on and in your assembly.

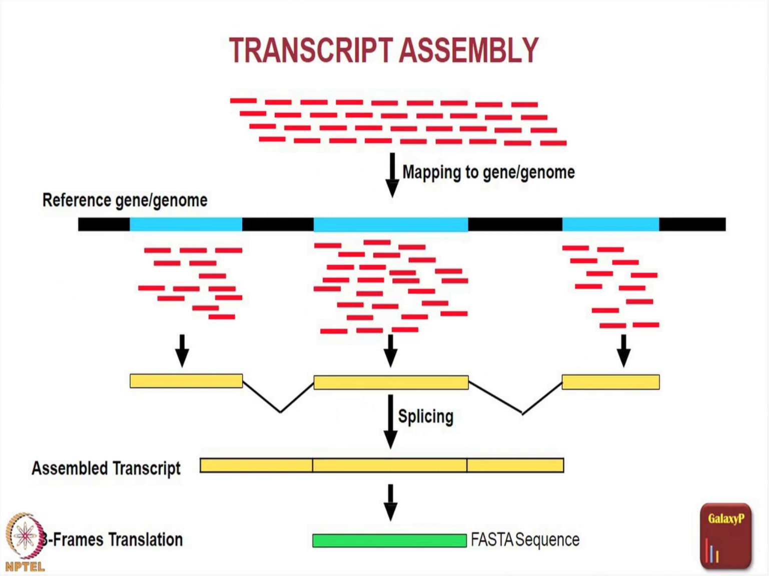
(Refer Slide Time: 24:39)



ALIGNMENT

So, for that again you start with alignment right. So, these are your aligned files.

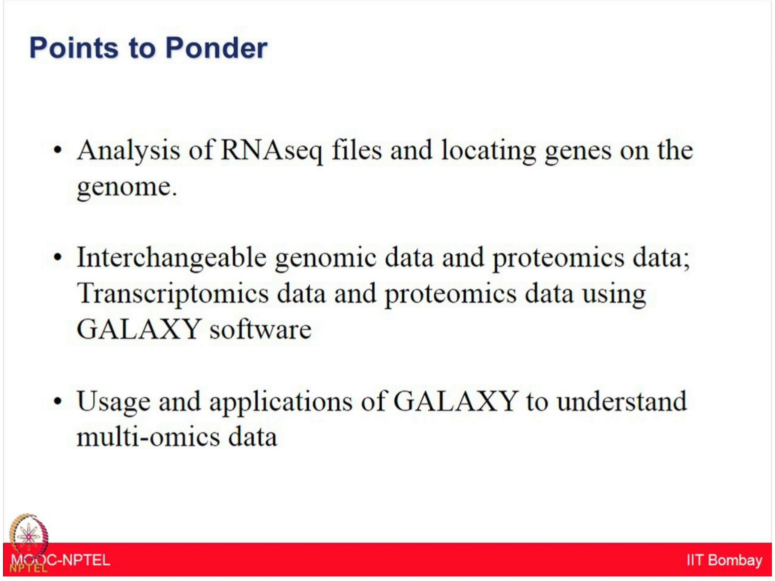
(Refer Slide Time: 24:45)



And then it uses a tool called as string tie which converts your you know you know your assemble transcripts into basically larger sequences and these are in this case this would be exons that you are looking at right and then, once you identify the exons it assembles this into an assembled transcript and that assembled transcript is generally you know it is subjected to 3-frames translation to give you a FASTA sequence right. Now one of the advantage of this workflow is you do not have to actually take all of those and then convert into protein FASTA sequence. It does a GFF compare.

So, it compares with the GTF file which is basically a list of all your known gene coordinates or assembled genes and compares it with what string tie has found out and if there you know if there is a variation from that, that is the only thing that gets you know retained and then converted into protein FASTA file, right. So, this output that you see here will basically have only novel transcripts from your assembly.

(Refer Slide Time: 26:05)



Points to Ponder

- Analysis of RNAseq files and locating genes on the genome.
- Interchangeable genomic data and proteomics data; Transcriptomics data and proteomics data using GALAXY software
- Usage and applications of GALAXY to understand multi-omics data

MOOC-NPTEL IIT Bombay

In conclusion, today you have learnt that genomic data and proteomics data as well as the transcriptomic and proteomics data are interchangeable and how one can make sense out of these multi-omics data requires a lot of skill sets, lot of experience and need for many softwares.

You got a glimpse of how one can process multi-omics data sets in today’s lecture, however in next lecture Dr. Jagtap will continue and he will talk more about the bioinformatics solutions for big data analysis in which way you can make a complete workflow and try to accomplish that.

So, next lecture will also be by Dr. Pratik Jagtap and he will finish this whole module of looking at Bioinformatics Solutions for Big Data Analysis.

Thank you.