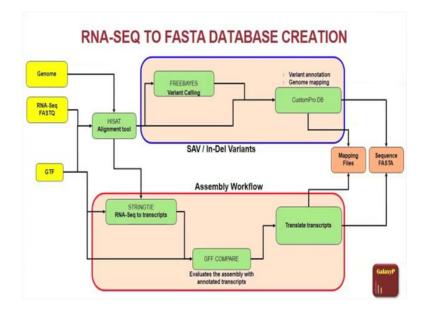
Introduction to Proteogenomics Dr. Sanjeeva Srivastava Dr. Pratik Jagtap Department of Biosciences and Bioengineering Indian Institute of Technology, Bombay University of Minnesota, Minneapolis (USA)

Lecture – 40 Bioinformatics solutions for 'Big Data' Analysis - II

(Refer Slide Time: 01:41)



Welcome to MOOC course on Introduction to Proteogenomics. In last lecture Dr. Pratik Jagtap talked to you about the workflow of RNA sequencing to fast a database creation. In today's lecture he is going to continue the discussion about Bioinformatics solutions for 'Big Data' Analysis. He will talk more about the output files and data analysis to understand the unraveling the questions in hand. He will also be talking about three workflows; RNA sequencing to variant FASTA database, database searching using MSMS data and identification and visualization of novel variants.

He will explain about search query a set of freely available algorithms for mapping the MSMS data to peptide sequences. He will also talk about Lorikeet viewer, a spectra of viewer of all the b and y ions identified in the mass spectrum. I hope it will also refresh you about previous lecture by Dr. Karl Clauser, where he gave you the understanding of manual interpretation of data sets. So, let us welcome Dr. Pratik Jagtap for a today's lecture.

So, these two tools both of these workflows, this one as well as this one generates a protein sequence FASTA file as well as a mapping file, right.

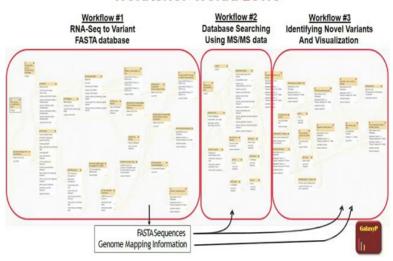
(Refer Slide Time: 01:50)

# OUTPUTS

>generic ENSMUSP000	00107433	Erp29 ER pr	otein 29				
MAAAAGVSGAASLSPLLSV	LLGLLLLF	APHGGSGLHTKG.	ALPLDTVTFYKS	RLLL	SP		
>generic ENSMUSP000	00120715	Rps2   riboso	mal protein :	52			
MADDAGAAGGPGGPGGPGL	GGRGGFRG	GFGSGLRGRGRG	RGRGRGRGRGAR	GGKA	EDKEWI	PVTKLGRLVI	KDMKIKSLEEIY
LFSLPIKESEIIDFFLGAS:	LKDEVLKIN	MPVQKQTRAGQR					
Genomic Mappir	ng File						
Genomic Mappir	ng File	121452190	121452340	_	0	150	
	-	121452190 121449139	121452340 121449163	-	0 150	150 174	
ENSMUSP00000107433 ENSMUSP00000107433	chr5 chr5	121449139	121449163	-	150	174	
ENSMUSP00000107433 ENSMUSP00000107433 ENSMUSP00000120715	chr5 chr5 chr17				150 0	174	
ENSMUSP00000107433 ENSMUSP00000107433 ENSMUSP00000120715	chr5 chr5	121449139	121449163	++	150	174	
ENSMUSP00000107433 ENSMUSP00000107433 ENSMUSP00000120715 ENSMUSP00000120715	chr5 chr5 chr17	121449139 24720275	121449163 24720452		150 0	174	
ENSMUSP00000107433	chr5 chr5 chr17 chr17	121449139 24720275 24720533	121449163 24720452 24720731	+	150 0 177	174 177 375	

So, this is a protein sequence FASTA file as you can see there is the ensemble identify here along with the name of the protein. If it is, you know if the function is known and this is another one and the genome mapping file basically gives you again the identifier of the or the accession number of the protein, which chromosome it comes from, what is the start side, what is the end side and it also gives you the length of the exon.

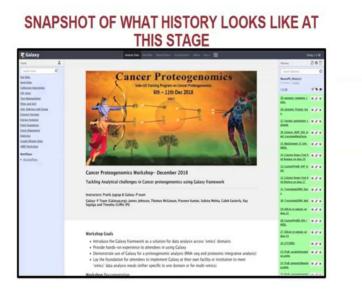
Now one of the things you will observe here is this is basically saying that this is the 174 base pair sequence or base sequence and then you can see that there is some gap between this and this, which basically shows that there is an intron present there, right. So, this kind of helps you to map your regions of the protein. And later it actually also helps you because if you just identify a peptide from this protein, one can use these coordinates to go back and find out what are the coordinates of the peptide as well, ok.



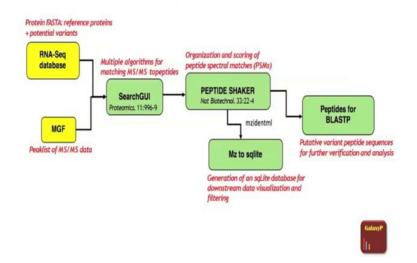
### WORKSHOP WORKFLOWS

So, let us say; let us imagine that you have run that workflow right this workflow which started with the FASTQ files and the FASTA file as well as the GTF file ran through these all of this and you got a protein sequence file.

(Refer Slide Time: 03:19)



And so, this is how your history looks like once you go through the documentation, once you use go and access the galaxy instance, go through documentation after the first work flow you will have this has a history. And one of these here would be a protein FASTA file that you can use your for your analysis.



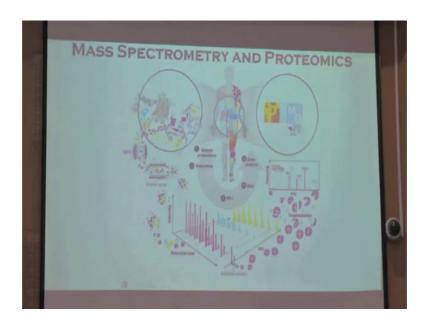
# PROTEOMICS DATA ANALYSIS USING GALAXY

So, the second part here is you have your RNA-Seq database or protein FASTA generated from the RNA-Seq database, you have your MGF files which you have acquired from the same data. And now you search it with we have got a tool called SearchGUI which basically has at least nine different search algorithms in it

So, it helps you to identify the peptides spectral matches and then it uses peptide shaker to perform both FDR analysis as well as protein grouping and then there is another tool called as Mz to sqlite that is used to perform further analysis. And it also generates peptides so, that you can perform BLASTP analysis. And BLASTP analysis is important because you want to go back to the NCBI and our database and ensure that, this does not these peptides do not match to what you have. And there have been some instances where in something that we found to be a novel peptide just a month before is no longer novel because somebody annotated it, right.

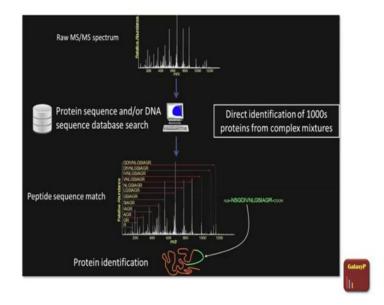
So, you have to ensure and even report in your manuscript that December of 2018, they still was an all peptide; now in January maybe it is not, but at least you have covered your requirements by saying this is the database as was against on this date, right.

# (Refer Slide Time: 04:56)



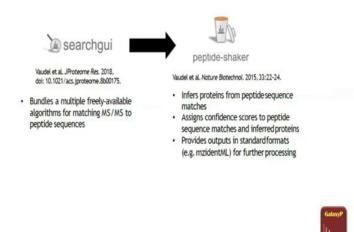
So, I might actually skip this slide, but this is really; I mean, I use this slide a lot to explain the complexity or with the process of how you start with how we start with proteins digest the peptides, put it through LCMS ms/ms and identify them. But you know this basically captures everything and you know you have perhaps seen this many many times so, I will not go through this, but this is a really good article this comes from one of the from Jürgencox lab in Germany and it covers most of the contemporary proteomics research that is going on. So, something I would suggest to go and read if you are really interested in basics or even contemporary status of proteomics.

(Refer Slide Time: 05:43)



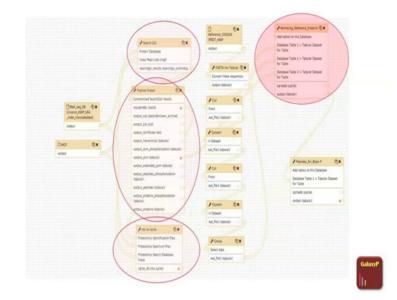
So, again you take your spectra you match against your protein database and if you know you can identify your peptides.

(Refer Slide Time: 05:50)



searchgui as I said is a freely available software you can also run it as GUI, but we have got it galaxy. So, it kind of runs in the workflow that I mentioned and then there is peptide shaker which basically does protein inference does FDR analysis, and identifies not only peptides PSMs, but also proteins from your sample. It also generates a mzidentml file which you can use it for further processing.

(Refer Slide Time: 06:15)

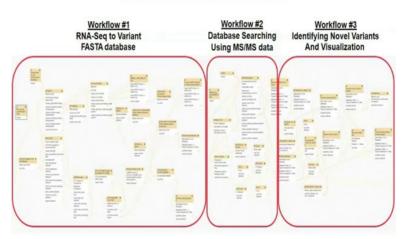


So, in this second workflow, right; so, we are at the second workflow now you have we have got your protein FASTA file from RNA seq data we have our MGF files and you want to search that.

So, what it does is it takes those two inputs does searches the data uses peptide shaker to generate multiple outputs and then this goes into a tool what we called as a mz to sqlite tool. So, what it does is it takes any tabular outputs and generates as sqlite database out of it. What it helps you to do is, it helps you to perform some more complex analysis on your tabular inputs and get outputs from that. So, instead of you processing these tabular files in multiple processes, you can have it in one place and then perform more complex queries on that to get answers, right.

And eventually what comes out of this is you identify peptides that we think a novel in the sense they do are not present in your reference proteome and then you can take those peptides and subject it to BLASTp analysis.

(Refer Slide Time: 07:17)



# WORKSHOP WORKFLOWS

So, that kind of covers these two workflows, the first one which generates a protein fast a file the second which gives you two outputs. One is your peptides that are sub going to be subjected to BLASTp analysis and it also gives you a mz to sqlite file and I will talk about it what it means, right. So, the third workflow which I think is the most interesting because you kind of you know this, these two workflows you can in some way maybe the second one you can use it, you know you can use any other software to do it right

you have a MGF file you have a RNA-Seq generated protein FASTA file can use any search algorithm to get your outputs. But these two workflows are something that you can you know with us the work that we have done hopefully helps you to get it done easier on your data sets.

(Refer Slide Time: 08:11)

1	2	History CO
#0	Sequence	Canach dataiens
GABLAK PSM=3081 length=7	CARLAK	March Distance
UNMELSEK PSM=623 length=8	UNMELSEK	MousePG_History2
LSIVPVRR PSM=2071,2128 length=8	LSAPARE	55 shows, 4 tables
ETHQQVVSR.P5M=1264.1585 length=9	(THQQVISR	2.04.08
LLATCSQDR PSM=1149,4918 length=9	LLATCSOOR	Loss and the second second
SUNITAR PSM=4948 length=9	SUNCTION	S4. Protides for Blast: @ / x
ELCSSON TAR PSM=4554 length=10	ELESSDETAR	27 Invi. 1 comments
F01QVPYCEX PSM=1325, 1728 length=10	FOTOMPYCEK.	27 tres, 1 comments format tabular, database mes50
LVEPCSPAEX PSM=3462 length=30	LVEPGSPAEK	terter treater sector les pe
QGLDGLLSVK PSM=2783,3670 length=10	QCLDGLISVK	SILECT Sequence () *PSM+* ()
CTGNSAPVIK PSM=3271,4265,4777,4799 length=11	CTONSAPVPK	group_concattid(',)    ' length='
NY/TLLSCFK P5M-2849.3395.4155 length=11	NYITUSCH,	InngthSequence) as "ID".Sequence FROM pum
TYSYLTPOLINK PSM-2634,2814 length=11	TYSYLTPOCIEK	WHERE length/Sequencel >6
CTGIVSAPVPIX PSM=2179.2332 length=12	CTOVSAPVPKK	AND length(Sequence) <= 30
SPYQ07TDHLVK PSM=399,529,2097 length=12	SPYGETTDHLVK	CROUP BY Sequence ORDER BY
SPVREFTDHLVK PSM+787 longth=12	SPYREFTCHEVK	GROEK BY Jength/Sequencel.Sequence
LLWDKETDUFK PSM-290,872 longth=13	LUWDKETDEFFK	
SLEEVIJSLPIK PSM=5147,5715 length=13	SLEEPILFSLPIK	5.0 D M 7 9.0
AVDPDSSAEASCLR PSM=5730 tempth=14	AVDPDSSAEASCLR	1
DGDUINPYLYSGAVK PSM+4725,5467 length+15	DCDLINPVL15GAVE	ata See
DSGASESREASAAR PSM=5469 length=15	DSCASCSLEASAAR	GALTLAN POR-MAIL Length-7 GAL
EVAED/INCHVGLGVK PSM-296 length-16	EVAEDENCHVGLGVK	UNELSEX PSH-623 Tangeli-d UNI
AVDPOSSAEASGLRAGER PSM+434 length+18	AVDPDSSAEASCURAQDR	LSDWWR PSH-2871,2528 Saught-8 LSD
SGLLAGORLVEVNGENVEK PSM+315,337,502 length=19	SOLLACORLVEVINGENVEK	ETHODAVSE FSH-1264, 1565 Langet-# ETH
ESSREALVEPTSESPREALAR PSM=2168 longth=21	ESSREAL VEPTSESPEPALAR	
SEQAEPWAAADTHEACEQNEAEX PSM=889 length=23	SEQAEPPAAADTHEACDONEAEK	53. Removing Reference
VPSQUILDOP.PDFSvCEQX PSM=3337 length=23	VPSQD4.DCPUPPPSACEQK	

# YOUR CURRENT HISTORY

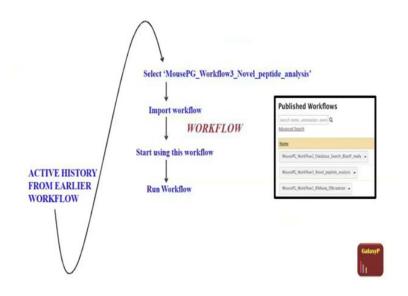
So, the third workflow; so, by the end of this your history would look like this, it will have 54 items and there are 54 of these because you have kind of added you know outputs from each of these tools.

(Refer Slide Time: 08:24)



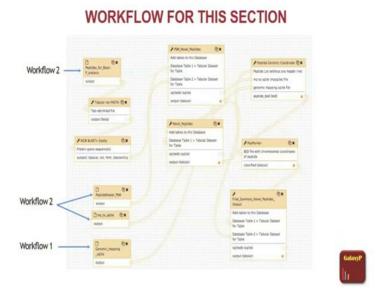
Again and this is just as a backup in case your search has not run because you know because it got stalled or because of some reason, you can always go back to that galaxy instance I talked about and download the history for the third workflow.

(Refer Slide Time: 08:46)



So, again, the same you know process you take your active history run it with the third workflow and what does the third workflow have?

(Refer Slide Time: 08:55)



The third workflow basically takes in inputs from workflow 2 which is this blastp peptides for blastp, it will also takes the PSM report.

So, the peptide spectral match report from PeptideShaker and this is generally at one person global FDR and then it also has this mz to sqlite files. So, this is basically all the information in a mz identml file right, the mz identml file is file that is generated by a protein or PSM search, that particular file all that mz identml file information gets into that. And this is really important because it also has information about continuous b ions continuous y ions all the spectral annotation information and that is something that you can use to process your data.

Student: If in MGF file. So, all the information is there. So, why we need a mzident file?

Right. So, MGF file basically is just your peak list right, it is it does not have any peptides annotations right, while the mz identml file has that and has got a lot more information.

Student: If my software is not giving me mzident xml file. So, from where I convert that my raw data file to mzident file.

I mean, if you want to use the you know the next tool which is the ability to look at the spectra right view your spectra, then you should have a mzidentml output in general and there are many software with I mean scaffold does that, protein pilot does that, I am sure proteome discoverer does that do I have not use that.

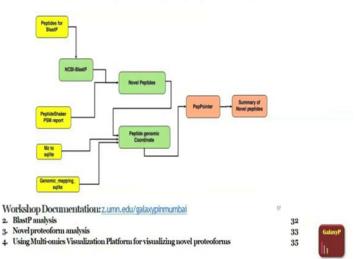
So, there are you know mzidentml is a quite a standard output nowadays that most of the software is generated. But that does not; I mean so, that will if your software does not do that, then it will avoid you to do the spectral visualization, but that does not mean you cannot do the rest of the things in this workflow.

But again if you run it in galaxy, since such go in peptides you can do that you can use that right, and then we also had this genomic mapping file from workflow one, right. So, all of these are actually used in the last workflow and what it does is you know; so, this tool here which is the query tabular tool. So, remember I talked about taking tabular files and generating sqlite database out of it. The sqlite database now can be queried and you can only say show me those peptides or peptides spectral matches which are novel right.

So, you matching it against the known database and try to find that. And these novel peptides or sorry these novel peptides would be a result of this NCBI blast that you have

done right and these novel peptides or novel PSMs now can be converted to just peptides. And these peptides can be subjected to looking you know, generating an output which can give you peptide it is genomic coordinates and other information. So, I will basically talk a little bit about that; so, there are two tools used here one is peptide genomic coordinate and peptide pointer.

(Refer Slide Time: 12:04)



WORKFLOW FOR THIS SECTION

So, look into the details right, we had this plus p peptide from workflow two this also from workflow 2, workflow 2 and this one from workflow 1 and then you know by NCBI have Blastp you use certain rules.

(Refer Slide Time: 12:20)

## **BLASTP ANALYSIS**

SELECTDISTINCT PSM.\* FROM PSM JOIN BLAST ON PSM.SEQUENCE = BLAST.QSEQD WHERE BLAST.PIDENT < 100 OR BLAST.GAPOPEN >= 1 OR BLAST.LENGTH < BLAST.QLEN ORDER BYPSM.SEQUENCE, PSM.ID

And these are the rules that we used, is if the peptide if it is BLASTp IDENT is less than 100 or there is at least one gap present or if the blast length is lesser than your query length, then you basically call it an all peptide. So, and we have kind of used this on multiple data sets and these three features seem to be enough to identify a novel peptide.

So, that is the information used here to identify novel peptides, and then it uses this tool called as pep pointer to generate a tabular output that you can use for further analysis. But before that I will like to maybe talk about this thing called is Mz to sqlite which is used as an input to visualize your data.

(Refer Slide Time: 13:13)

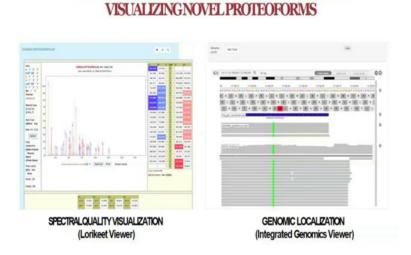
to sqlite on data 36,	data 7, and others			
eptide Overview			Fittering Sequences	
ed from Galilony - View in Protein		its Bequeroes for Fillering or Cour	Conception of the second se	
equence	1) Reacting Count	Protein Count		
NOPORSABASGLA	1	3		
ICPOSSAEASOLAAQOR	1	1		
COLENPACYSON O	2	1		
DISGASGISLEASAAN	1	3		
LOSSOLINA	1	1		
EDDREALVEPTDEDPHPALAR	1	1		
NHITLI &	1	A		
CHATTERING.	1	1		
oving 1 to 2 of 2 antins fillend from ov: 10 1 antins Instant Propositions (PSA) (PSA) (PS	ex.579 total anting and by lower +	Parka 1 Net		

MULTI-OMICSVISUALIZATION PLATFORM FOR

And I would really start if you are really interested in spectral visualization I will strongly encourage you to go to that you know the galaxy instance and the documentation. Because this really does not take time each workflow takes I think the first one takes 12 minutes the other take just 2 minutes each, but you know it will help you to go through that.

So, what the mz to sqlite data does this it takes this mz ident ml file that you have and generates a list of all the peptides that have been identified with associated information. Now you can select novel peptides; so, remember we actually identified it is a novel peptides here which we already in the history you can use that. So, you could say I just show me the subset of this novel peptides right and then if you do that, it this list goes down and then by using various tabs that are open here you can visualize the spectra using lorikeet, this is all within galaxy.

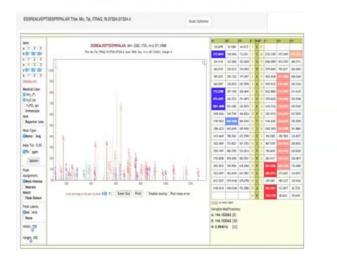
(Refer Slide Time: 14:03)



MULTI-OMICSVISUALIZATION PLATFORM FOR

# So, you can use a lorikeet visualization tool to look at which b and y ions are or which peaks are annotated with b and y ions, how many of these are continuous b and y ions and so, on. So, basically look at the spectral quality because even if you are done 1 percent FDR if you are getting really good great score for your peptide, it is not necessary that the spectral annotation is good. And I think it is important that if you are identifying a novel peptide you at least have the means to you know to show or convince yourself and then the reviewer that this is you know this is a peptide that indeed is novel and is definitely needs further interrogation.

(Refer Slide Time: 14:47)



# ESSREALVEPTSESPRPALAR

One can also perform genomic localization using that mvp tool that I mentioned. So, again in that tutorial there is this peptide that we looked at and you can see it is annotated quite well. Now the lorikeet viewer also is interactive in the sense you can check on b and y ions and so, on and so forth to select these.

So, this if you were to do this manually this would maybe take you know 20 minutes to half an hour for somebody to look at each of these while this takes a few minutes or even a few seconds to go through this. It also gives you an option to say this is a good one or bad one so, that you get you know it is kind of you can look at the list of all your novel peptides that you have here. Let us say if you have 200 of them, you can look at the spectral quality and then only select those that are that are important for you.

GENOMIC LOCAL ZATION (INTEGRATED GENOMICS VIEWER)

(Refer Slide Time: 15:38)

The next part is genomic localization. So, again by using various buttons or various features where you look at the protein view report in the mzsqlite file, you can open the integrated genomics viewer and then you can look at you know what are the variations that you see. And there are you know you can lay down tracks and look at the RNA-Seq data, you can look at the three frame translation of the DNA sequence the annotated peptide sequences and so, on and so, forth. So, it kind of gives you a pretty good way of looking at the peptide that you identified and you can also scroll and look at adjacent peptides and so on and so forth.

### (Refer Slide Time: 16:25)

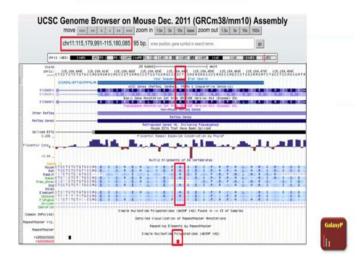
# NOVEL PROTEOFORM ANALYSIS

1	2	1		1		1			18
Prysideleouries	SpectralCourt	Proteins	Chromosome	Surt.	End	Stand	Antotation	CanomeCaordinate	VCSC, Genome, Browser
AUDPOSSAGASCUR	1	88886/590000021077_F1825.E186CA277T	(9/1)	115176449	115176491		C08	cm11115178449-115176491	https://penume.ucm.edu/rgi-bri/hgTracks?db=mm134pesition=chr12434(15126449-11512648)
AUDPOSSACAGO, RAQOR	1	ENGNEOPO0000021077_P1825-0196C-8272T	dv11	11517(44)	115174513		C25	091111517640-115176501	https://perame.acsc.edu/cpi-an/hg?racks?do-mm1Mapainton-chr11klA11512M449-11512M583
DCDUNPU.YSGAVE	1	ENIMUSP00000117347_N18%-Q187L3246C	05	171445444	121445481		05	09/5/121445444-121445488	https://genome.ucuc.edu/tgi-bin/%gTracks?db=mm104poultion=chr58384121445444-121445489
DSCASCSILIASAAR	1	578G.13.1_32,291	i9r17	22966997	22857042		heatingst	0v1722866997-22867042	https://genome.acia.edu/ogi-bin/hgTracks/db=mm1Mpusition=chr17X3A22866997+22867942
ELCORDL TAR	1	\$286,17.1,4,88,379	1942	91155262	91155292		1001	0929119262-9119292	https://genome.ucst.edu/rgi-bin/hg?sacis?db-mm304pesition-chr2K3A30155242-81155242
ESSREALVEPTSESPREALAR.	1	ENSMISPEDDDDD21077_P1835.E2%G.A2721	chr11	115180006	115180069		005	cw11115180006-115180069	https://genome.acxc.edu/rgi=bin/%gf1acks/dbwnm104pesition=chr11X34313140006-1153340069
NYTUSOK	- 1	\$786321_030,242	1914	58482150	58482383		interprot	01458482330-58482383	https://genome.uctor.edu/op-oin/kgTracks/db-mm104pesition-chv4k3A58482350-58482383
SPVBHTCHLSK	1	INVESTIGATION FOR THE PARTY PARTY OF THE PAR	19/17	34731702	24721828		Salustances	christ 24721702-24721828	https://percone.urs/.edu/cpi-bcr/hgTiacks/dbarmetHassidionach/17KIAN4721703-14771426

	chromStart	chromStop	name	score	strand	annotation
chr11	115176449	115176491	AVDPDSSAEASGLR	255	+	CDS
chr11	115176449	115176503	AVDPDSSAEASGLRAQDR	255	+	CDS
chr5	121445444	121445489	DGDLENPVLYSGAVK	255		CDS
chr17	22866997	22867042	DSGASGSILEASAAR	255	-	five_prime_ut
chr2	91155262	91155292	ELGSSDLTAR	255	-	intron
chr11	115180006	115180069	ESSREALVEPTSESPRPALAR	255	+	CDS
chr4	58482350	58482383	NIYITLLSCFK	255	+	intergene
chr17	24721702	24721826	SPYREFTDHLVK	255	+	SpliceJunction

So, the eventual output that comes out of this is it tells you a chromosome number what is the start and stop side and what is the peptide that came out of it and what is the annotation you know, where do these peptides lie. The ones that you found in CDS were basically the ones that were either single amino acid variants or there were some that we found in the five prime you have untranslated region and so, on and so, forth.

So, if you have let us say hundreds of them it helps you to identify or even classify these peptides, based on the very incentive form. It also gives you something more interesting, you can solved this by chromosome and then you might even see a pattern where in these peptides may be are lying around the particular region and that kind of gives you an idea about a mutation that could be occurring in that particular phenotype, right. You can use the information the output that you generated from this; you can just take that link and open it in UCSC genome browser.



# UCSC GENOME BROWSER

And then you can look at more details about you know how was the genome would homology of that particular gene in other organisms, and convince yourself that the mutation that you see is indeed novel or is similar to some other organisms right.

So, again there is information here that you can use it through UCSC browser and in this case for this that particular peptide, I talked about the same peptide right that is the localization we found that there was a mutation in this particular region, that G U had gotten converted into A.

(Refer Slide Time: 18:09)

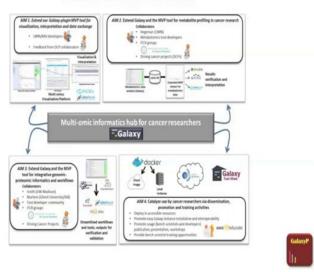
Query[8nbsp; Local query sequence) NP_038160,1 ka(+)H(+) exchange regulatory cofactor foal architectures: 3251	- PDZ PDZ PDZ	the second se
		1 EBP50_C superfamily
ezrin binding protein 50, partial tourionry spar: Blateria 3 Similarity score: 2 Total nr sequences: 899 Lookup sequences in Entrez	PD2 * # 'EBPSO_C	007599, EBP30, C-terminal; This C terminal domain allows Interaction of EBP30 w/m FERM (four-point one ERM) domains, resulting in the activation of Ezm-
PREDICTED: PDZ and LIM domain protein 5 Issonmy soar: Eutheria Similarity score: 2 Total nr sequences: 10 Lookup sequences in Entrez	• PD2 • • • • • • • • • • • • • • • • • • •	radixin-moesin (ERM), with subsequent cytoskeletal modulation and cellular growth control. It includes a disordered section between two
POZ and LIM domain protein 5 taxonomy span: Boreceutheria 4) Similarity score: 2 Total in secuences: 5	PD2 '	reasonably well conserved hydrophobic regions.

# CDART BLAST SEARCH



You can also and this is not even galaxy, but you can also go beyond and use some other tools to look at which you know what are the conserved domains that it occurs in.

(Refer Slide Time: 18:21)

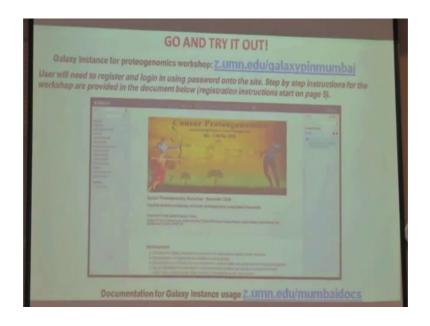


**PROJECT OVERVIEW** 

So, I mean basically I would really strongly encourage you to go ahead and use that tool because if you are interested in using proteogenomics or using galaxy for proteogenomics I am sure most of you are interested in proteogenomics. But and it kind of gives you a sense of you know how it can be used, now it is again set up for a small data set you know with something that works in maybe an hour or two. But you know you could also think about them upgrading that to you know using it for larger data set your own data sets for example.

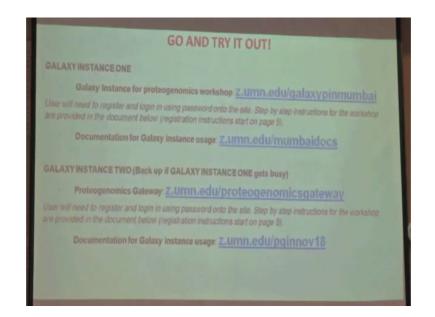
So, I just wanted to mention that galaxy p is not only about proteogenomics we also work in the area of meta proteomics and metabolomics. So, we are developing tools and workflows to enable that analysis as well.

# (Refer Slide Time: 19:12)



And then yeah; so, this is what I would encourage go try it out look at this site, there are instructions all you need is a registration and password and you know a login and password and if you use this documents which are really detailed and we have used it for at least three workshops last or this year you know feel free to try it out.

(Refer Slide Time: 19:33)



Again these are the same documents as I said as a backup if this fails you can always try this instance and as I was mentioning to somebody here, this would be instance that

would be there for longer. So, if you want to just skip that and start using that that is fine too.

(Refer Slide Time: 19:51)



I just wanted to acknowledge the people who were responsible to make this happen, Praveen Kumar in professor Timothy Griffin; professor Timothy Griffin is the PI of the galaxy grant and Subina Mehta were the people who actually; so, Praveen developed a few software along with James Johnson, but Subina Mehta actually tested them and also worked on the documentation that you can go and have a look at.

(Refer Slide Time: 20:20)



Lastly we work with multiple researchers around the world are not only users, but also developers, because you know the tools are getting developed in you know as the field is emerging not only in proteogenomics, but other fields as well. You see that many tools are getting developed. So, we try to work with the best tools that are available try to package it in galaxy and integrate into workflows, links to you know you can actually go to galaxy p dot org and find more information there is also galaxy p dot org slash contact if you want the contact us, but with that I think I will be happy to take any questions.

Student: Does this map the database containing novel junctions and splice junctions database?

Yeah; so, the mapping file that is generated from the second workflow would have that information.

Student: Sir, does it innovates currently have the list of novel and the current splice junction sequences?

It is so, yes for a particular data set it will have it right. So, you need to run your data set your RNA-Seq data to get that.

So, I mean for this example yes we do have it, if you run some other data sets we have which we have done we have that; yeah, so it is possible. So, that is why we have those two different workflows one which is single amino acid variants which uses a different set of tools and junctions which is a slightly more a larger genomic reorganization right So, yeah, and it is not only captured in your genome mapping file, but it is also captured in a protein FASTA file. You will have your genomic coordinates and you know the fact that it is a junction a novel junction.

Student: Are we going to do hands-on sessions for this?

I would encourage you to do the hands on session later with the documentation, but I can show you what I was talking about in the sense I can show you this in action or where things are and again if there are more questions I can take that.

# (Refer Slide Time: 22:32)



So, this is the galaxy instance right this is one.

Student: can we use it for work from home.

So, you know for you to log in and register this would be you know. So, I am already registered logged in into this right this is the tool format and this is the history and this is the pane the central pane. So, the way you started you go to shared data and then there are histories click on history.

(Refer Slide Time: 23:10)

laxy				1000	CON- COARDON	a many stars a	11	
	1	Published Hist	ories					History
taan O	search name, americation, o Advanced Search						Unsamed history (amply)	
Operations		Rame	Annotation	Owner	Community Rating	Community Tags	Last Updated;	
Operations		HousePG_History3		smehte	OANNA!		~14 hours ago	This reatory is empty. I load your and data or
ipulation Surl tract and Genup	HousePG_History2		smetta	0+++++		Nov 02, 2018	from an externel source	
	HousePG_Hatary1		smelte	000000		Nov 02, 2018		
armata atures		Housefil Joout History		smette	Q		Nov 02, 2018	
persons and								
play Data inhop								
-								

So, as you can see there is a input history here right, and if I click on that.

(Refer Slide Time: 23:23)

аху	acation frame morphism	Shared Data-	About this History
Histories   southte   HeuverPG_Ing	ut_History	Integral Industry Note a case of the heater set 2 to 0	
		0	Related Histories
datadh		Annotation	All published histories Published histories by a
nusculus.GRCm38.bed	•		Rating Community
d_peptides.pro.lam			() stop, 55 sempt
MGEs			Tags
g_uniprot_cRAP.lasta	•		Community: none Yours:
acutus. GRCm28.86.gtf	•		4
in .	•		

It shows these are the 5 files or 10 files in fact, that I have here and I can explain why there are 10 files and not 5 or and we can see only 6 of them sorry and. Then I go to this place called import history and you can name this anything.

(Refer Slide Time: 23:36)

Challenson   percebba   MonocolPE	Importing history "MousePG_Input_History"	history	Author
M_Input_History	Enter a title for the new history: Imported: MouseRG, Joput, History, Muttal		enalts Related Histories At admined Interes
annatus Alice Victori	6	Cancel Import	Published Automatics Review Rating Community (3) wereauting
MBD ma			Taga Community: name
No. All Concession	:		4

So, I can call this Mumbai and then imported once I import it, it basically goes here.

# (Refer Slide Time: 23:50)



So, this is just one way this is for the tutorial, but you could also have all your data sets on your the somewhere else right on the FTP site or on your computer and you can upload that as well. So, there is a upload function that is also available, but you know for the sake of this So, here as you can see there is a fastq file, there is a gtf file, fasta file, MGF files and so on, and you know there are again icons here, I can try to look at the data.

(Refer Slide Time: 24:22)

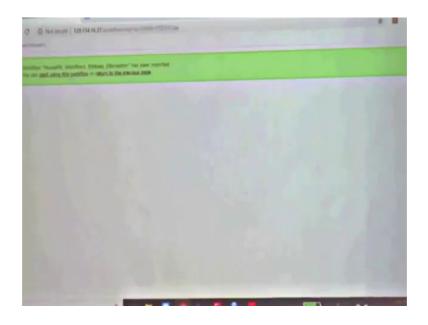
			Ane	iyze Deta			aller a	wite line -	Ш		
	1	-	Start	End	Name	Score	Strand	ThickStart	Thickind	RemRGE	History
		Chrom	3214401	3671490	ENSHUST0000070533	1000		3216021	3671340	0,0,0	search datamets
	0	chr1	3999556	4409241	ENSMUST00000208660	1000		3999556	4409187	0,0,0	
		chr1	1999556	4409241	ENSMUST00000208660	1000		3999556	4409187	0,0,0	imported: HousePG_Input_Hist
		chr1	4292980	4409187	ENSMUST00000194992	1000	*	4292980	4409187	0,0,0	Housers Inder Chan
tions		chr1	4344145	4360314	ENSHUST0000027032	1000	*	4344599	4352025	0,0,0	
		chr1	4490930	4496413	ENSMUST0000027035	1000		4491715	4493406	0,0,0	913.15 MB
		chr1	4491249	4496757	ENSHUST00000195555	1000		4491715	4492591	0,0,0	10: Mus_musculus.
		chr1	4491309	4497354	ENSMUST00000192650	1000		4491715	4495155	0,0,0	GRCm38.bed
d Group		chr1	4491712	4496363	ENSMUST00000116652	1000		4491715	4493406	0,0,0	A CONTRACTOR OF
		chr1	4492457	4496330	ENSMUST00000191647	1000		4492457	4495155	0,0,0	9: identified_peptid vs.pro.bam
		chr1	4492464	4493735	ENSMUST00000191939	1000		4492464	4493406	0,0,0	H.J. L. L.
		CV1	4492466	4496395	ENSHUST00000192913	1000		4492466	4493406	0.0.0	8: Mo_Tal_MGFs
		chr1	4773205	4785710	ENSMUST00000130201	1000		4774451	4785677	0,0,0	A.214
		chr1	4773210	4705739	ENSHUST00000156816	1000		4776463	4785677	0,0,0	3: Ref_P D0_unipre
		chr1	4781220	4785739	ENSMUST00000146665	1000		4782220	4785677	0,0,0	L_cRAP.fasta
ta .		CHI	4807822	4546739	ENSMUST0000027036	1000		4807913	4045016	0,0,0	2: Mus. musculus.G
		011	4807829	4841286	ENSMUST00000150971	1000		4807913	4041159	0.0.0	RCm38.86.gtf
		ches	4007095	4045174	ENSHUST00000119412	1000		4007913	4041112	0.0.0	State of the second second
		(2/1	4807597	4840969	ENSHUST00000137687	1000		4807913	4040969	0,0,0	1: ProB.faste
		evi	4807910	4845352	ENSHUST00000115529	1000		4807913	4845016	0.0.0	4.3.18
		01	4608236	4841093	ENSHUST00000131119	1000		4830274	4841093	0.0.0	farmet fastgnanger. i
		ev1	4057013	4297905	ENSMUET00000081551	1000		4857913	4295364	0.0.0	mm10
		del.	4858037	4897909	ENSMUST00000165720	1000		4858407	4896364	0,0,0	uploaded faituranger I
		01	4909375	5019539	ENSMUET0000002533	1000		4910473	\$019379	0.0.0 .	

So, if I click on this, the central pane shows me which chromosome you know what is the format of that particular file and that is true for your fastq file as well and so on and so, forth right bam file; so, one can look at that, on the left side here are tools available. So, you can also search tools for example, if I were to look at SearchGUI you know, it shows me that there is this tool called SearchGUI and then associated peptide checker too. But again for the sake of workshop if I go; so, you can actually you know spend a lot of time looking at this and then if you go to the shared data, you can go to workflows now and what we are trying to do is convert this fastq file you know protein fasta file.

(Refer Slide Time: 25:10)

ixy	Analyza Data	an passing and the H	
1	Published Workflows		History
- 0	jeenth name, annotation, per Q. Advanced Search		imported: HousePG_Input_History
Iperations	Rame	Annetation Owner Community Rating Community Tage	
the scores	HousePo Workfow2 Detabase Search BrastP rendy .	seets 02020	913-18 148
dation et	Housefil, Workford, Novel, paptide, analysis +	anata ARREN	10: Mus_musculus. a GRCm38.bed
it.amil.Genntp math	HousePG moniford #NAsee Obvieton .	andra 22222	9: identified_poptid a es.pro.batt.
urms nema	importadi ABAF_Norisfox1_ASAlaas_DBcreation .	forme and all the	B. Mu_Tal_MGFs
sents			3: Ref_5000_smiprs @
ny Data Imp			2: Mus.menculus.G a RCm28.86.ptf
			L: Proll.fastu
			4.3 103 former fastquarger, com sem10 uptented fastquarger file

So, go to a workflow and I am going to select the first workflow which is of the second one here which is a workflow one RNAseq database constraint generation right. So, I go here and I say import. (Refer Slide Time: 25:25)



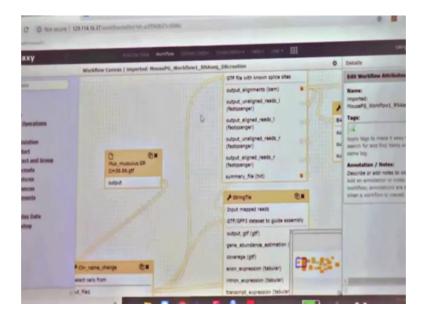
And once I import I can say start using this workflow.

(Refer Slide Time: 25:30)

аху	Anaiyze	Dete	distant in		14 M	11. INT.		History	-
7	Your workflow	vs				marrie for source	fron + 1	Cannon defaults	
. 0	Rame		Tags O		r of iteps	Published	Show in tools panel	imported: MousePG_Input_Hist	berg
per atlenes		orted.	14 1	N 2	2	No		913.15 MB	
dation at	MousePG Workfront	Arolong, Dillowation +	1					10: Mus_musculus. GRCm38.bed	
t and Group nats	Rat .							9: identified_peptid es.pro.bam	
arms Acces	Descrited							B: Mo_Tal_HGFs	
enta	Ratatta							2. Ref_5000_semileren	
y Data mp	Delete							2. Mus. musculos.G RCm38.86.ptf.	
								1: Prob.fasta	
								4.3.108 former fastquanger	
								upteniad festeranger	

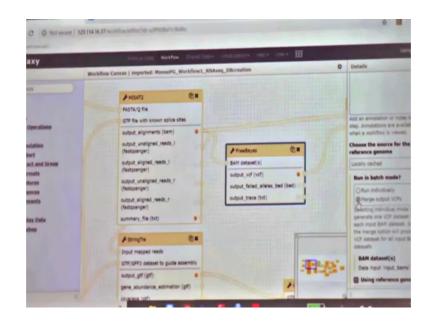
Right and there is one feature which I did not talk about, but this is the workflow now. So, if you look here if I click here, this workflow is in your, you know you have kind of transferred the workflow in your domain you can modify the workflow So, I can show you that. So, if you have this right I go to this, but. So, there are many things here, but I have and I would strongly encourage you to go and explore these. So, I go to this place called edit and then hopefully very soon it will open the workflow output that we are seeing earlier.

(Refer Slide Time: 26:10)



So, you can see here this basically shows the structure, and I think there is a way of reducing this to I do not know how to do this on this computer. But you know you can reduce the size and look at this, but here you can actually see that there is a fastq file, the gtf file as an input and there are these various tools that we talked about, right.

(Refer Slide Time: 26:35)



So, now for example, if I wanted to go into this tool right and if you for example, wanted to not have I mean I am just giving an example let us say this is run individually, but if I want to say no I want to change that to merge, I can do that and save this, but I might name this differently or I could I can change anything here and save it And so, you can modify these you can also add things to this workflows you can add inputs or outputs from these workflows add tools to these workflows. So, it is quite flexible that way right. And then so, I am not going to save this because maybe I will just have this yeah. So, and from here I can now run the workflow right. So, maybe I will save this and I will run the workflow, you can also run it from the earlier place that I showed you.

C @ Net tenare | 129.114.16. V Run w v: imported: Workflow1\_RNAseq\_DBcreation 1 Workfle 913.15 H us.GRCm38.86.98 00 20.04.01 NIPROT\_CRAP C 2 2 Raf \$000 unigrot cRAP/auto # 4: Chr\_name\_change (Galaxy Version 1.0.0) 1: ProB.fas # 5: HISAT2 (Galaxy Version 2.1.0) # 8: StringTie (Galaxy Version 1.3.3.1) ......

(Refer Slide Time: 27:33)

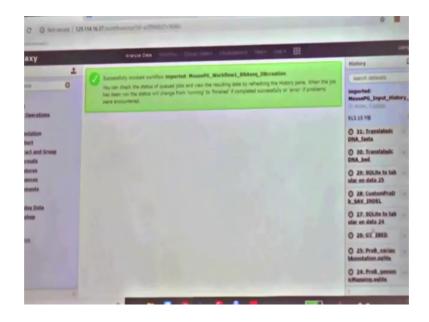
And when I am running the workflow now all I need to ensure is that I select the right input. So, here this says the first one should be fastq, but it showing fasta file. So, I might go to you know select the right one which is the fastq file, here I need to select gtf file. So, I select that and thirdly this is right. So, this is the protein fasta file that you have and then I go ahead and just click and you know.

# (Refer Slide Time: 27:57)

ixy	Average Data provides prevailable frequencies rates large II	1	
1	Markflow Imported	✓ Run workflow	History
. 0	Workflow: imported: MousePG_Workflow1_RNAseq_DBcreation		search datasets
	# 17: SearchGUI compatible Protein Names Variant Annotation (Galaxy Version 1.0.0)		imported: MousePG_Input_Histor
	# 18. bed in protein map (Galaxy Version 6.1.0)		Coloma, Coloma
perations	# 19. Uniprot+cBAP+Translatz_sDNA+CustumProDb (Gatasy Version 1.2.0)		913.15 HB
dation			10: Mus_musculus.
el	# 10. Variant_Annotations (Galaxy Version 2.0.0)		
ct and Group nots	Add tables to an existing database		9: identified_peptid
mats.	Database Table	-	
10.05	1: Database Table		& Ho_Tal_HGFs
ents	Tabular Dataset for Table		
	Clubber instant vol. Rec' New aller 17		3: Ref_5000_smipro
y Data	Filter Dataset Input		and the second
mp	Table Dutines		2: Mus_muscalus.6 RCm38.86.gtf
	If have the splite database in your history		1: Profi.faste
	tu		And a state of the
	I SQL Query to generate tabular output		4.3 1/8 turner fastganger om
	Inter		mm10
	include goary result related beaders		upnaded Networps Fil

So, each one of these as you can see our tools in that workflow right and there are 20 such tools. So, or oh sorry there are 22 such tools right and then I go ahead and just say well run workflow.

(Refer Slide Time: 28:10)



And then what will start happening is and you should should be able to see this very soon is that the outputs as you can see have started piling up.

So, our in initial inputs were here and the first file I started getting converted from the first tool and this will start happening for rest of them and all these proteins are all these

in inputs will start getting generated. So, this is how you are generating history right and maybe they just wanted to have a more dramatic way of saying I generated history, but I do not know I mean that is why it is called history. So, at the end of it we should have these outputs that are coming up right. So, I am going to kind of just jump and skip to the next history which is.

(Refer Slide Time: 28:53)

аху	Analyze Det	and the second	and the second	· rate part	1	
1	Published Histor	ies				History
- 0	Advanced Search	٩				imported: MausePG_Input_History
operations	Kama A	nnotation Owner	Community Rating	Community Tags	Last Opdated;	Martine, Children
Det ations	HousePG_History3	smetta	0.00000		~14 hours ago	913.15 HB
dation of	MousePO_History2	amahta			Nev 02, 2018	O 39. genomic_ma
ct and Group	MausePG_History1	smehta			Nov 02, 2018	O 38. Genemic, Pr
mata	HouseFG_Input_History	smelte	0+++++		Nov 02, 2018	otein_map
10.00 10.00	House's Logo Lossay					O 37: Variant_ann - etation_splitedb
eenta re Data						O 36: Uniprist_cRA P_SAV_indel_transl atedbed_fasta
						0 35: Bed Jarotein
						O 34: Column Reg ex Find And Replace on data 29
						O 33: CustomProd

So, let us just imagine that you know that history is run right which it showed in 12 minutes we have checked that. So, I am go to history two and import that history, ok.

(Refer Slide Time: 29:06)

аху	in the state State Data -	and the last II	
Histories   smelta   MoorePii_History2			Abread then Heatery
PG_History2		Roke A mark of this backer and 3 to 4	hor
		0	Related Historie
		Annotation	All published instances Published histories by
			Rating
ten_for_Blast-P_analysia			Community
comp. Matherance. Produitik	•		(Longs, 1.1 sample)
un data 31			Tags
rt. un data 30.	•		Community: none
dada AP	•		There is
t on data 47			-
option um data 30. data 7. and athers			
data 41.	•		
stinior_profess			
diaber_pepties			

And sometimes I actually put in dates. So, that you know which of these workloads were on which dates, but in this case I am just saying Mumbai because I know if I just store the same thing it might conflict with my not conflict, but I might get confused later if I have to select it.

Student: (Refer Time: 29:26) workflow 2.

What is that?

Student: History 2 is a workflow 2.

History 2 is basically a, no history 2 is an output from the first workflow.

Student: Ok.

So, the ones that were grey right now, once you run them you will have history 2; so, I am kind of doing like a cooking show right I mean just imagine this is cook now or let look at the second yeah.

Student: History 1 will be the raw file then.

This will be the input files yes, yeah.

Student: Input files, input with the output from the.

First workflow and the third one will be from the second workflow.

Student: Ok.

History 1 should be the result of our first workflow, ok.

# (Refer Slide Time: 30:00)

axy	and the stand Data-	
Histories   amelita   HousePG_History1	Insport Distance	Alarmit Data History
PG_History1	0	Author smalls Related Histories
	Ansotation	All published histories Published histories In I
the mapping aging	•	Rating
ne Protein map		(1 ostrop. 3.5 average) Yours
Lannatation_autitedb	• Balles in the second second second	Tags
A CRAP NAV under translatedbed Fasta	•	Community: none
utain SJ SAV INDEL	•	there is
Negez Find And Replace on data 29.		
Prodb_SAP_indel	•	
Rayna First And Raplace on data 27		
mutcOliA_fanta		
InduDNA_Beel		
in takolar an data 25		

So, I am just going to import it this is the way it is. So, that would have everything that is run and I will show you what we do next. I wanted to kind of just mentioned that if you have a completed history you can convert this into a workflow. So, I will I know this is little advanced, but I will just show you So, for example, this is your you know after running the first workflow and you generally do it this in a stepwise manner right you can go to this wheel and then say extract workflow.

(Refer Slide Time: 30:40)

аху	Analyze Data and Data Second	And I Real Property lies in the local data and the	
L Operations	The following list contains each tool that was run to create the you with to include in the workflow. Tools which cannot be run interactively and thus cannot be in Workflow name Workflow constructed from Today		Hotory bearth Galaxity imported: HousePG_Hist 7 march 1 million 1.3 GB
wistian	Create stort flow Check all Uncheck all	History items created	39: genomic_mappi a ng_sqlite
iort act and Group muts turns	University of the clear of a service of the service	1.FASTQ_ProB.fastquamper * Treat as input dataset FASTQ_ProB.fastquamper	38: Genomic_Pratei a n_map. 37: Variant_annotat a ion_sqlitedb.
ences menta	Unknown This fami cannot be come in work-frame	2 Hus_musculus.GRCm38.86.gtf * Treat as input dataset Hus_musculus.GRCm38.1	36: Uniprot_cRAP_5  AV_indel_translated bed.fasta
ley Data Jump	Unknown Pha tour cannot be used in sectificant	3 Ref_5000_umprot_cRAP.fasta	35: Bed2protein_63
•	Dataset Collection Creation	8 Mo, Tal, MGPs	34: Column Regen.F
	Dataset univolute created in a way not compatible with monthing	P Treat as input dataset (No. Tol. (HDPs	33: CustomProlib_B
	Unknown	9 identified_peptides.pro.ham	32: Cohema Bauera #

And so, now, what happens is, if I have run a history right I can call it something right. Right, and then I say create workflow and it will generate a workflow now I can share this workflow with you can take your input files on it, I can share the workflow with anybody.

C A histomer 1231141527

Avery Cold

Aver

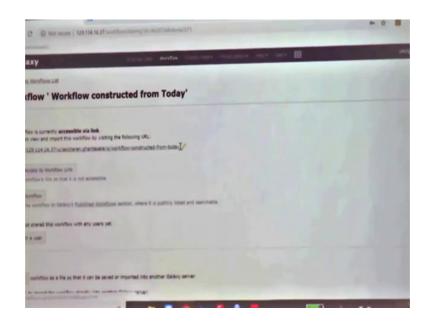
(Refer Slide Time: 30:54)

So, the ability to share so, let us say if we are on the same galaxy instance, I should be able to take this workflow, right and then I can share it.

(Refer Slide Time: 31:05)

жу	Acaryne Data	and lights		-	and lines	18	History
	Your workflows				minist for and	· ·	
. 0	Rame	Tags	Owner	e of Steps	Published	Show in tools panel	imported HousePG_His
per ations	workflow constructed from Today +	4	TEN	28	No	0	1.9.08
ation	0	14	You	22	No		29. yunamir_mappi eq_aqiite
s and Group	mported Mausefü Visrafiors Rickand Disreators -						38: Ganamit_Profei
ata 115							37. Variant_arrestat
na nta							38. Uniperal_siBAP_8 AV_inder_transisted best.fasta
Data							25: BedZarstein_SJ
							34. Column Report? and And Replace on data 23
							22. ContemPradb_8
							32 Column Bours P

# (Refer Slide Time: 31:13)

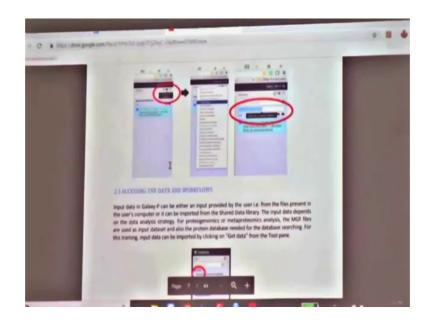


So, all I have to do is generate a link and then send this to you by email, if you click on this you will log in on this you should be able to open that workflow. So, that kind of helps you to not generate.

Yeah, we will use the same parameters. So, coming back to that yeah; so, again we are once you do this, what I will do is maybe just show you. So, we did the same thing now we search the database and then we generate outputs, what I will show you the last output that is generated, which is basically everything together right your input files, output from the first workflow, second workflow and third workflow and which is one here.

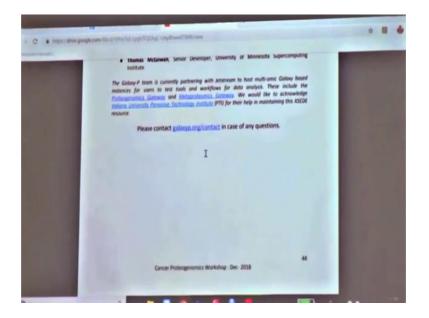
And then we can very quickly see if we can look at some of the features, and galaxy has been in using genomic studies for quite a few years now we basically just adapting it and using the ability that the fact that you know it is really strong in genomics and transcriptomics to use it proteogenomics because then it is easier to merge these.

# (Refer Slide Time: 32:19)



So, you can basically go through this really step by step manner, it also gives you information of basics of galaxy what are histories and all that. So, it is really is a very easy way of learning galaxy if you are interested. So, those who are interested please go and use this and then later once you get a good feel for it, you can also contact us at galaxyp dot org I think which is at the end we have contact.

(Refer Slide Time: 32:47)



Yeah, so, you can contact as galaxyp dot org n we can suggest, so, there are some sites in Europe right now which I have got a lot more you know infrastructure. So, you can run your data sets on that or we could also contact some researchers who are locally running galaxy as well to run some of your data sets there.

So, this is the final summary or the, you know after running all the three workflows.

Galaxy		Ameryze D	eta -	Quality manager and in	- III		History C
anime y	1		1	1		3	History 6
entime ata Data Data Sun Operations ent Languation entrant Enatures Enatures Enatures Enatures Display Data Enclose Support Data Enatures Support Enatures Enatures Enatures	0	и плановецина нарадования и россияния проставили нарак оранования валарования вала валарования валарования валаро	SpetteCourt	Protein Experimentation Experi	Connessme der13 der13 der1 der1 der12 der12 der12 der13 der2 der13 der13	0244 118176449 121485846 220403846 220403846 2118156006 8118156006 80482330 249721702	Exercit accounts Impacted: HouseRD, Math 21 Annuel 1 Comments 51 Annuel 1 Comments 51 Annuel 1 Comments 51 Annuel 1 Comments 51 Annuel 1 Comments 52 Annuel 1 Comments 52 Annuel 1 Comments 53 Annuel 1 Comments 54 Annuel 2 Comments 55 Annuel 2 Comments 5

(Refer Slide Time: 33:19)

As you can see here it generates this output with the peptide, the number of spectra that it was identified with what was the localization which chromosome what was the start site.

(Refer Slide Time: 33:29)

Galaxy		A.w	and a sector base here merenes and and II	ting)
en house anta Data Data Sun Operations est anopartations and Surt autorect and Secure (Aramata Inguestions) Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Autorect Aramata Inguestions Arama	± 0	maCourthAM 11517444-11517460 1217444-11517400 1214454-12147400 1216174-2201740 1216100-11111000 10161200-11111000 1016200-11111000 101700-0472100	94 UCCC, General, Brower Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.34115174444-1131450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3111514444-1131450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3111514444-1131450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3111514444-1131450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3111514444-1131450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.31115124444121450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.31115124444121450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.31115124444121450 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.31115124444121420 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.31115124444121420 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3111512444412442130 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.31115124444121420 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.31115124444121420 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3111512444412170 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.311151244472170 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3142472170 Rate //general acts data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3142472170 Rate //general acts data rate data rap-bin/hg/facis/fdawmi108.pp0faces/fdavefi/19.3142472170 Rate //general acts data rate data rate data rate //general acts data rate rate rate rate rate rate rate r	History C Tempor descents Imported Henselfs, History (1999) Imported Henselfs, History (1999) Im

It also gives you this link that you can use to open in a UCSC browser to look at, you know which is kind of an sorry an alternative way of looking at then IGV browser So, this is the mz to sqlite like mvp output that I talked about. And if I click on this visualize in mvp application it opens this application.

(Refer Slide Time: 33:51)

Now so, these ones here are not novel these are all the peptide sequences from mz identml file, but in your history you have a place where it says novel peptides yeah. So, this 58 history is a list of novel peptides.

(Refer Slide Time: 33:51)

FOR Data Visible Scores Vi	eter Sizes 🔕 Case				
Legentes	11	Spectra Count	U	Protein Count	
LINGONEBININE		2		1	
BAAAAE CAARPLER		1		1	
BAAADDISEEF BIEFETB		1		1	
BAARETER		1		1	
KAALAAAAAAAAAA		1		1	
AAAOLLONGAGOEGAGOT		3		1	
ALANTAR O		1		1	
NACHEAL OR		1		1	
AADUEIDODIER		1		1	
MARIELOW VEIL		1		1	

So, these are only you know since these are small dataset there are what eight of them here right and then the mvp application basically helps you to load that from galaxy.

(Refer Slide Time: 34:18)

And then here you would see only those eight peptides.

(Refer Slide Time: 34:21)

	FOR Data	Voitin Larren	Verified Scotts () Close al Bran					1			
TOM Detail for 8									×		
			spectrumTitle	their	entical	PeptidaShaker PSM score	PeptideBhaker PSM confidence				13
		enter-018	Ma Tau (TRAD IS 07324 07324			#7 2125	100		100832	71875	1.04

If I click on it, it opens this oh, and then yeah.

# (Refer Slide Time: 34:52)



So, it gives you this peptide here and now I can select you know whatever hidden lines.

Student: names are not visible

Yeah, but it is not so, it is not been annotated right now. So, I am if I add this and you are right I mean you know if you think it is not a you know you know if I start doing this you know you might agree with it, but it is a very subjective thing right I might think three ions continuous is good, but you might not and at least it gives you an ability to look into that. But, yes I mean it helps you to do that, but you can also do it for other spectra and so, on and so, forth. So, right now oh yeah so, and then; so, if I go from load from galaxy if I do that, I should be able to. I can get those eight now and I can also go to view and protein which I might not cover.

# (Refer Slide Time: 35:44)

FDR Data Valles 1	term terms from () Der et bars		
	View Peptides in Protein		1.
nz to sqlite on data 36, d	AMOPUSTALASOLI		
	AVDPDSSAEABGLRADDR		-
Pepilde Overview	DODLENPVLYBOAVK	equences	-
Lond Irom Galaxy Varie in Protect	DISGASIOSILEASAAR		-
Contraction of the	ELEMETICIAN		
and the second second	ESSPEALVEPTIENPROVALAR	-	10.00
Sequence	NEWTLLEFR		
BNDPOSSAEASGLA	BETREFTINGER	100 C	
BN/DFDEDAEASICLRADOR	Can		
DODLENPALYSIAN®		_	
COSASSIE EASAM	1 1		
DILOUISCETAR	1 1		

Right now I can then select this peptide.

Student: How can we say that its novel peptide?

(Refer Slide Time: 35:50)

· C A fait accer   128 114 14.37 physics cound-adverse	Transpiration Statest, S-a23110.1	1234631	*	
PDR Data Visible Boover Vertical Sec	O Char of Scene			
ENSMUSP00000021077_P182S.E196G A	272T : 17707.0535jENSM	UST00000021077 ENSMUSG0000020733 :		
M S A D A A A G E P LIP R L C C L E KI	G P N G Y G F H L H G E	KGKVGQFIRLVEPGSPAEKS		
nz to sqlite on data 36, data 7, and ot	hers			
Peptide Overview User ton December 4	NEGDREALN	Filtering Sequences	2	J
	The Own		_	-

Because we went from a RNA-Seq data. So, it is from your protein FASTA file.

# (Refer Slide Time: 35:57)

THE Data Maller	Scores Vertical Scores 🖸	Dear at Sears		
CEVIPSQEHLDO	PLPEPFSNGEI	QKESSREALVEP	SESPRPALARSASSI	
VIPSQEHLDG	PLPEPPSNGEI	QUESSREALVEP	TSESPRPALAR	
			-	
		N		
r to solite on data 36. d	lata 7 and others	b		
nz to sqlite on data 36, d	lata 7, and others	6		
	lata 7, and others	6	Filtering Sequences	×
Peptide Overview		SIMEALS	Filtering Sequences	
Peptide Overview	Restor 1 M		Filtering Sequences	
Peptide Overview	Restor 1 M	STREALS	Filtering Sequences	
nz to sqlite on data 36, d Paptae Overveer Lord fore Cleans View & Police Register	Restor 1 M	STREALS	Fittering Sequences	
Peptide Overview	Rester 1 M		Fittering Sequences	

But then it gives you ability to look at look at.

Student: exons

Yeah, so, this is the genome centric view.

(Refer Slide Time: 36:14)

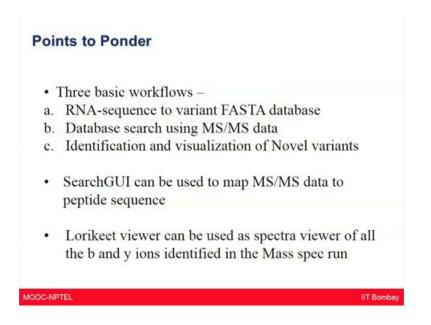
	If Data Visible Scores	Verliet Game 🕗	Chier all Scarit					
-								
E A D					* * *			
		-						
			b			THEAT IS C	00000007338	
MUSP000	00021077_P182S.	E196G.A272T :	17707.0535					
	Chr chr11		Chr. ch					
VIPSQ	EHLDGPLP	EPFSNGE	IQKESS	REALVI	PESES	PRPAL	ARSASSI	
		EPFSNGE				PRPAL		

And now if I go here you know. So, there are ways that you can look into this or you can expand this and look at the three frame translation of that particular peptide, we can see,

that peptide is here the third frame and you know. So, anyway I mean again this is for one peptide, but if had 200 peptides or many peptides you should be able to see that, ok.

So, I think with that I will end this.

(Refer Slide Time: 36:48)



In today's lecture I hope you have learnt that two output files from RNA-Seq to FASTA database creation workflow, which means the FASTA file and genome mapping files. You also got a glimpse of how one can search for novel variants using three basic steps shown by Dr. Pratik Jagtap. We also learnt that is how one can use galaxy output files and use it in different online softwares like Integrated Genomics Viewer; IGV to understand the mutations. In a gene in the next lecture you will listen another speaker Dr. Ratna Thangadu who will talk about large scale data science.

Thank you.