AN INTEGRATED SYSTEMS BIOLOGY PLATFORM FOR COMPLETE PROTEOGENOMIC ANALYSIS.

Pratik Jagtap¹; John Chilton¹; Ebbing de Jong²; James Johnson¹; Joel Kooren²; Getiria Onsongo¹; Sricharan Bandhakavi³; Timothy Griffin² 1. Minnesota Supercomputing Institute 2. Department of Biochemistry, Biophysics and Molecular Biology, University of Minnesota, Minneapolis, MN 55455 3. Bio-Rad Laboratories, Richmond, CA

INTRODUCTION

- Proteogenomic studies use large-scale MS-based proteomics data to identify novel gene products and genomic reorganizations, thereby revealing new insights into genome biology.
- Proteogenomic analysis presents challenges such as large database searches, disparate tools for analysis and quality control and lack of a resource for integration of large-scale proteomic and genomic information.
- As a solution, using the Galaxy-P framework, we have developed a complete pipeline seamlessly integrating protein identification tools with genome mapping and visualization tools.

TOOLS, WORKFLOWS and COMPONENTS.

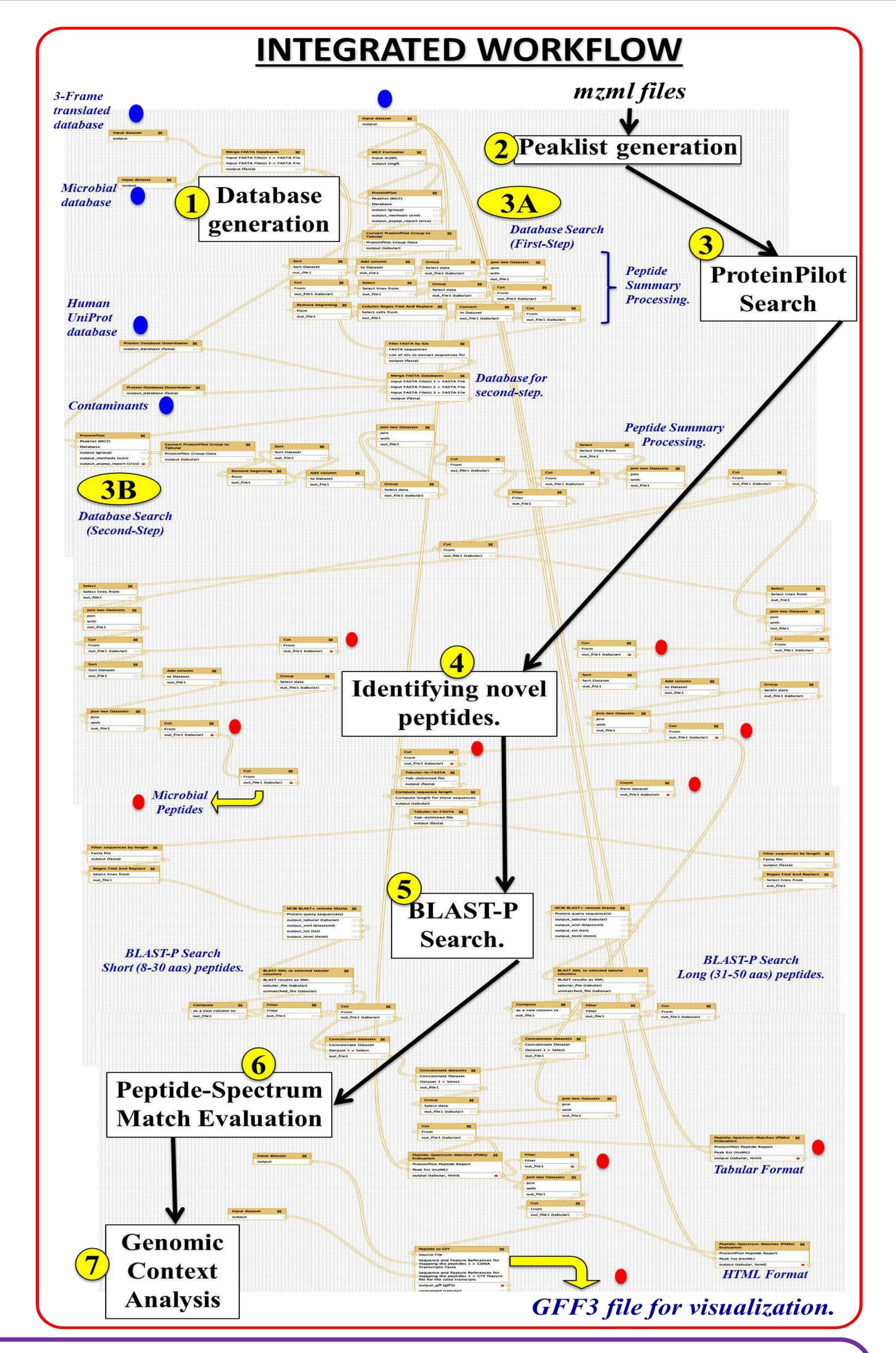
Galaxy-P Tools : Protein Database Downloader.

- Galaxy	Analyze Data Workflow Shared	p.a									
Tools	Protein Database Downloader (version 0.2.0)										
Protein Database Downloader	Download from:										
Get Data • Protein Database Downloader	UniProtKB										
- Protent Database Dominoader	Taxonomy:	_									
Workflows	Homo saplens (Human)	Dial									
 All_workflows 	reviewed:										
	UniProtKB										
	Proteome Set:										
	Reference Proteome Set 🔳										
	Include isoform data:										
	3										
	(Execute)										
	Output										
	A FASTA file containing the specified protein sequences.										

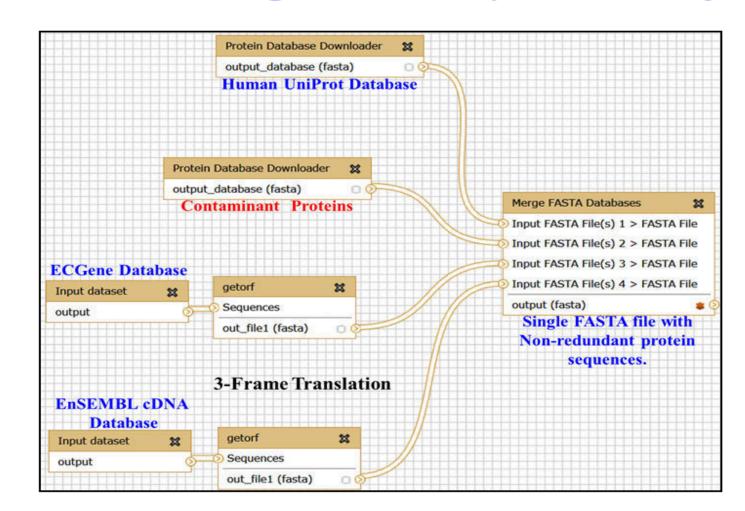
۲

•

Galaxy-P has multiple tools – some that are proteomics application specific and others from the genomics Galaxy framework. For example, Protein Database Downloader downloads UniProt protein FASTA databases of various organisms.



Running a Galaxy-P Workflow : Database generation.



Analyze Data	Workflow	Shared Data -	Visualization -	Admin	Help +	User +		
Running workflow "Proteoger	nomics	Database'	•				Expand All	Collapse
Step 1: Protein Database Downloader	(version	0.2.0)						
Step 2: Protein Database Downloader	(version (0.2.0)						
Step 3: Input dataset ftp://ftp.ensembl.org/pub/release-68/fa	asta/homo	sapiens/cdn	ia/Homo_saple	ens.GRCh	37.68.c	:dna.all.fa.g	jz	
EnSEMBL 67: ftp://ftp.ensemblcdna.all.fa type to filter Step 4: Input dataset								
genome.ewha.ac.kr/ECgene/								
64: http://genome.ewhw_fasta.txt								
Step 5: getorf (version 5.0.0)								
Step 6: getorf (version 5.0.0)								
Step 7: Merge FASTA Databases (vers	ion 0.1)							
Send results to a new history								
Run workflow								

Components of proteogenomics workflow.

STEP	INPUT	TOOL	OUTPUT
Database generation	cDNA databases, Protein FASTA files.	getORF, get data, merge FASTA.	Merged FASTA file.
2 Peaklist generation	Thermo RAW Files.	msconvert, MGF Formatter	mzml and MGF files.
3 Database Search	MGF Files, Search database.	ProteinPilot	.group file, peptide summary and PSPEP FDR report.
First-Step	**	Workflow with text manipulation tools.	.group file, peptide summary.
B Two-Step	MGF Files, Modified search database.	Workflow with text manipulation tools.	.group file, peptide summary and PSPEP FDR report.
Identifying peptides from translated nucleotide database.	Peptide Summary	**	Peptide List.
5 BLAST-P Search	Peptide List	BLAST-P and short BLAST-P.	List of peptides.
6 PSM Evaluation	Peptide Summary, mzml files.	PSM Evaluator, ProtVis.	Tabular or HTML Report.
7 Genomic Context Analysis	Peptide Summary, cDNA database, GTF file	Peptides to GFF3	GFF3 file.

• DATASET 1: 3D-fractionated salivary dataset. ProteoMiner TM treated. 52 RAW Files. LTQ-Orbitrap instrument. (Bandhakavi *et al* J Proteome Res., 2009, 8: 5590).

24 novel peptides. Previously annotated as introns (10); UTRs (5); different frame (4); pseudogenes (2) and unannotated (1) and 2 novel exon junctions.

00								IGV -	- Sessio	n: /U	sers/r	nsi/[Deskto	op/ig	v_ses	sion_	After3	D_Con	icat.xn	nl											
Human hg19		All		\$						Go		4	Þ	Ø		×										E	111	u))II	(1)(1)	111111	+
	NAME DATA TYPE DATA FILE		1	1	2	1	3	4	1	5	1	6	1	7	1	1	9	10 I		11	12	13	14 I	15 I	16 I I	17 1	19 18	20 ²¹	22 	x	¥ L III
luman RefSeq Genes InSEMBL 37.71 GTF Human			. A.	ملية		ام است							ام ام	يە د 2	loom	in to	see fe	atures.	aller	سة	. . .		ىسا .	<u>مدا ا</u>	Цu				4 101	لىلە خا	. 1
After3D Salivary Peptides			1				Т				Ť		T	T	1	11	1	f -	11	I				1		I	Ĩ	П	Ĩ		

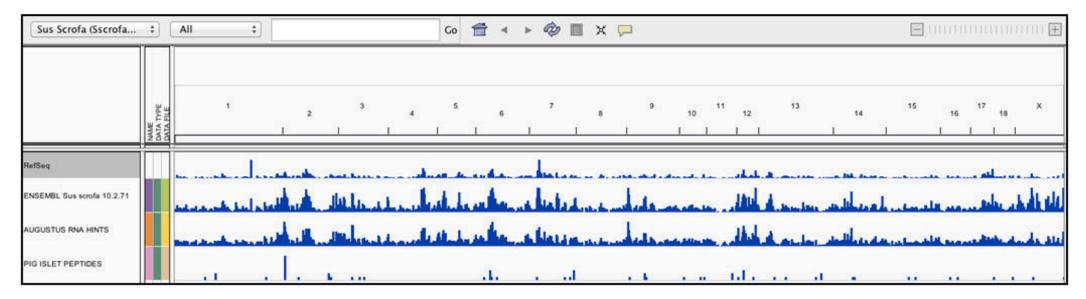
DSYVGDEAQTK : 67 spectra. Pseudogene. ENST00000415794_7

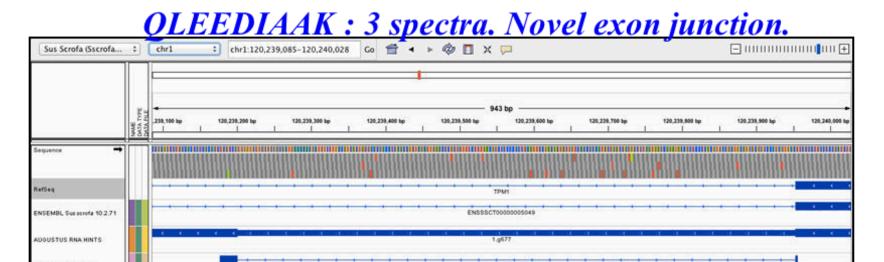
Human hg19	•	chr3			*	chr3:1	39,212,	794-139,	212,851	Go	Ê	٠	▶ (ф 🛙	X	7						E] 11111	шшп	
¥2		p26	1 p25	.1 p.	94.2 p2	p22.2	p21.31	p14.3	p141	pl	23	12.1	411.1	q12.2	q1 3.1	3 q13.3	2 q2	1.3 q	22.2	q24	q25.2	q26.	1 q26.3	1 q2)	.1 q28
	NAME DATA TYPE DATA FILE	•		138,212,600 I	bp	Ţ.	1	19,212,810 bp 	8	1		139,212,8 		58 bp -	0	138	,212,830 b 	p			139,212	1,840 bp			139,212,850 E
equence 🔫		A T H M	G A G	C C	A G A	K R	GAT CD	T C C	T A		r g q v	<mark>с с</mark>	R	A T C	GAG E R	G C	c c	A G	A C D T	C A Q P	A G C X S	A C A	G C	A T C	C T G
uman RefSeq Genes			-10 - 51					 																	
nSEMBL 37.71 OTF Human											-		ENST	00000515	247			-		-				• •	

GENOMIC CONTEXT ANALYSIS

• DATASET 2: iTRAQ-labeled Pig Islet dataset. Three Replicates. 45 RAW Files. Orbitrap Velos (HCD) instrument. (de Jong *Unpublished*).

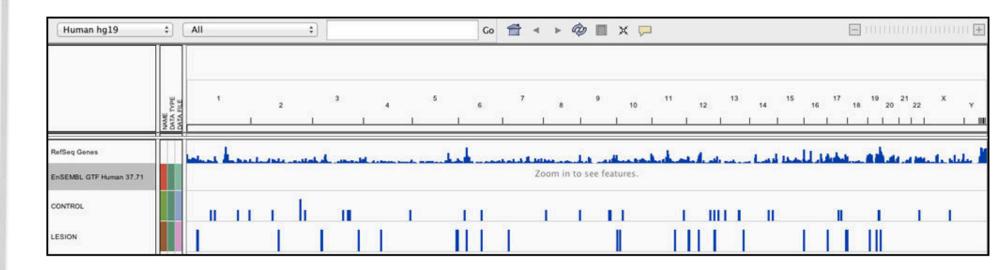
28 novel peptides. (See Poster 583 MP29)





 DATASET 3: Oral pre-malignant lesion exudate dataset.
 6 matched pairs (with controls). 84 RAW Files. LTQ-Orbitrap. (Kooren Unpublished)

38 novel peptides (21 lesion and 17 control). Previously annotated as introns (16); UTRs (8); different frame (7); pseudogenes (1) and unannotated (4) and 2 novel exon junctions.

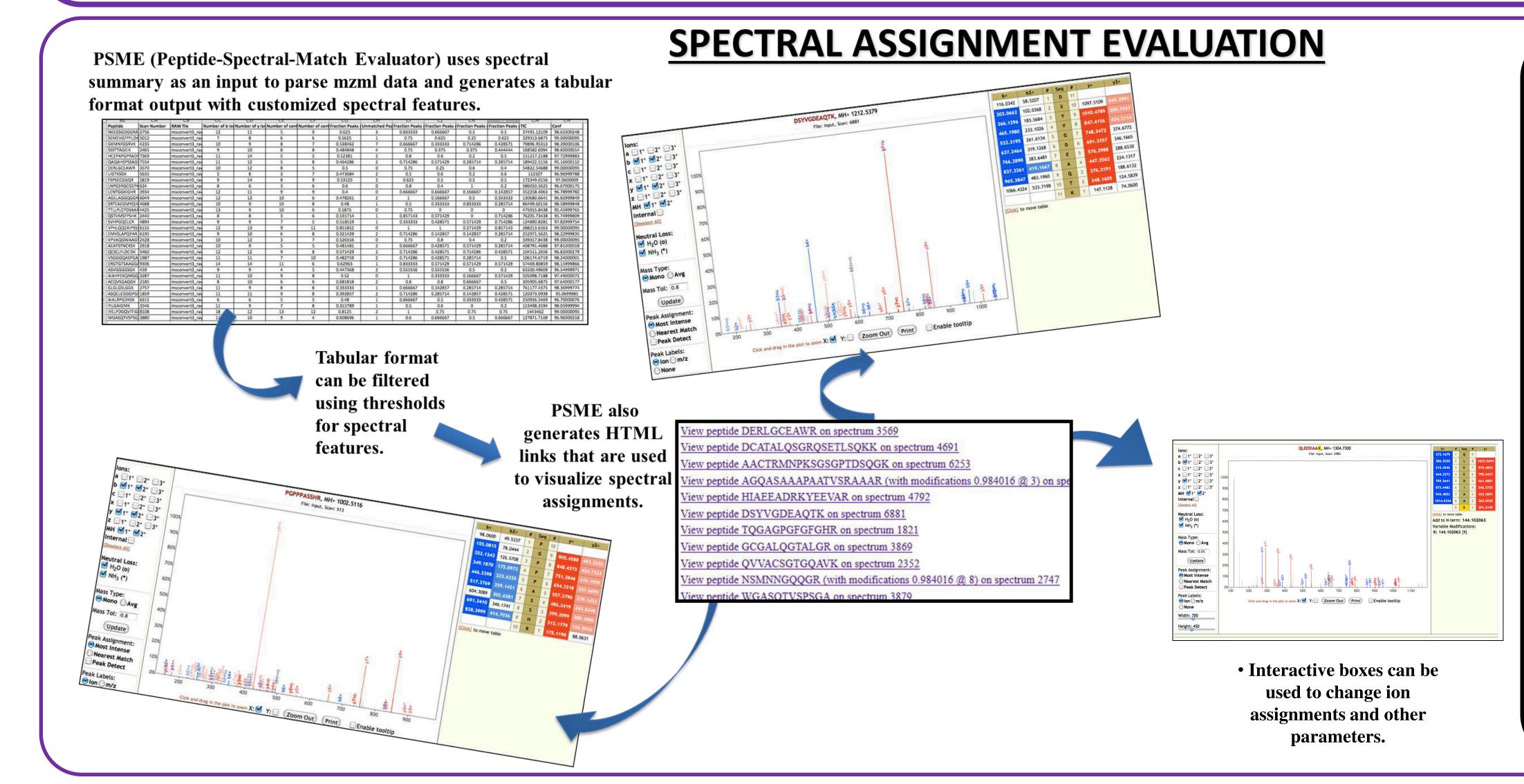


PGPPPASSHR : 1 spectrum. UTR.









RESULTS AND CONCLUSION

• We demonstrate the use of a complete platform for routine proteogenomic analysis, and highlight the potential for Galaxy-P as a solution for systems biology. At each step, we have used abundance of caution for selecting spectra, so that only meaningful results are analyzed and reported in the subsequent step. Using this platform, we identified 24, 28 and 38 potential novel peptides from three large datasets.

• It is also noteworthy that these **workflows are versatile** such that appropriate modifications can lead to use in metaproteomics or other systems biology applications.

• This complete, versatile, seamless and collaborative platform for systems-biology applications can be used for simultaneous proteogenomic and metaproteomic analysis using MS-based proteomics data.

• ACKNOWLEDGEMENTS: NSF grant 1147079; Minnesota Partnership for Medical Genomics & Biotechnology and The Center for Mass Spectrometry and Proteomics, University of Minnesota.