An Integrated Systems Biology Platform for Complete Proteogenomic Analysis.

Pratik Jagtap[1]; John Chilton[1]; Ebbing de Jong[2]; James Johnson[1]; Joel Kooren[2]; Getiria Onsongo[1]; Sricharan Bandhakavi[3]; Timothy Griffin[2]

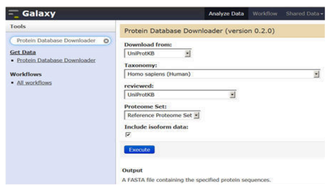
1. Minnesota Supercomputing Institute 2. Department of Biochemistry, Biophysics and Molecular Biology, University of Minnesota, Minneapolis, MN 55455 3. Bio-Rad Laboratories, Richmond, CA

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.**

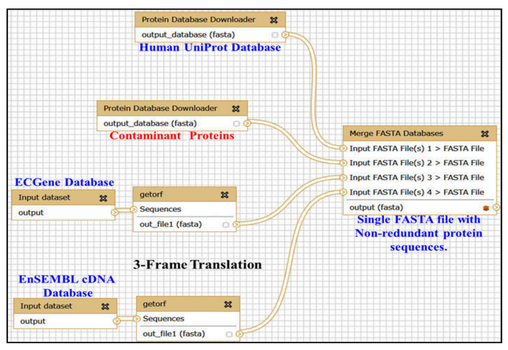
TOOLS, WORKFLOWS and COMPONENTS.

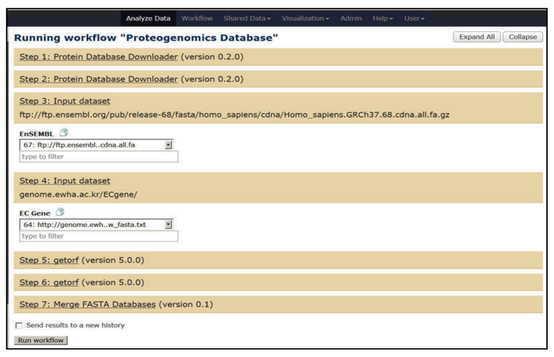
- ***Galaxy-P Tools : Protein Database Downloader.***



Galaxy-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.***

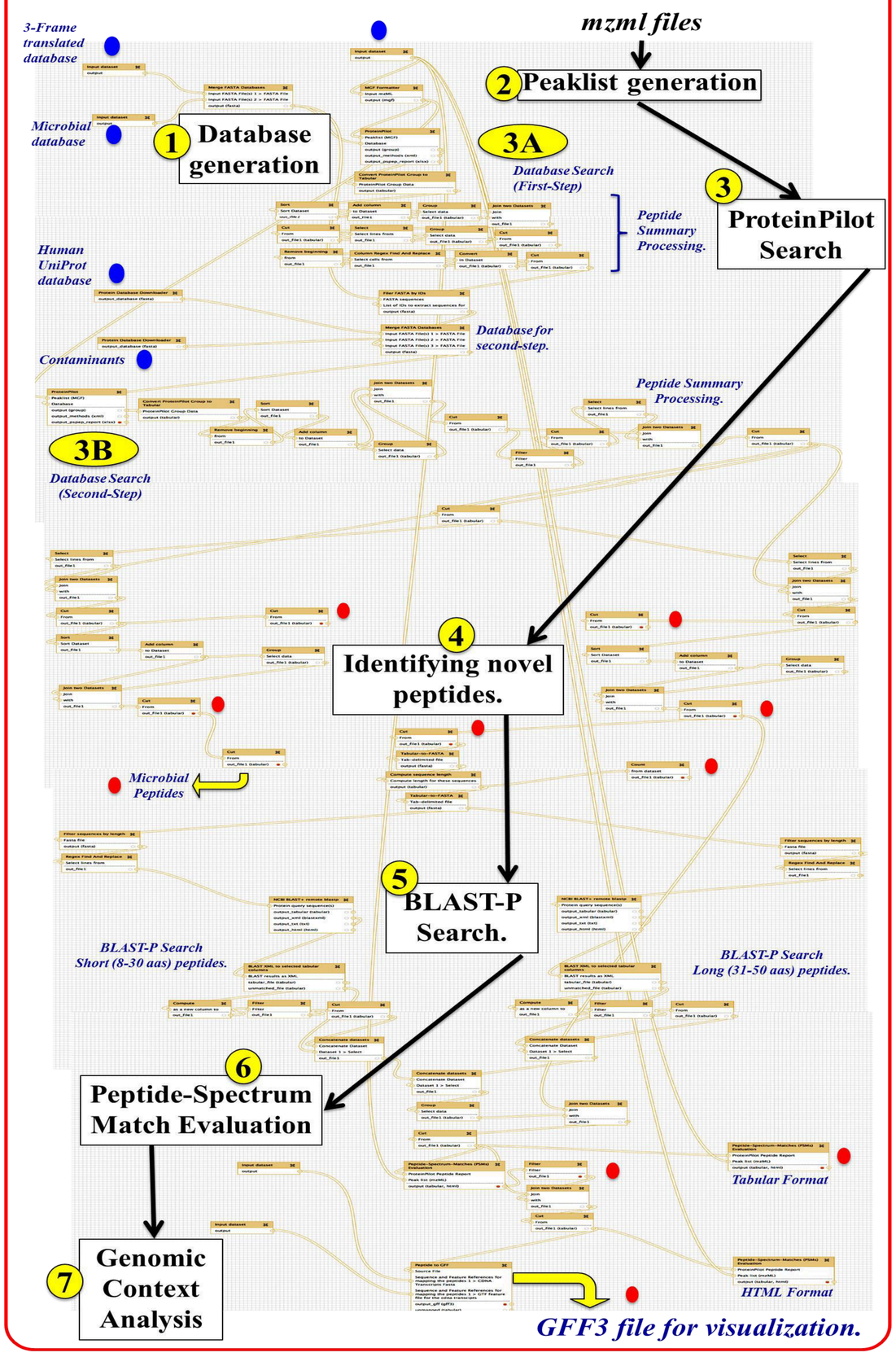




- ***Components of proteogenomics workflow.***

	STEP	INPUT	TOOL	OUTPUT
1	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.
A	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.
4	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.

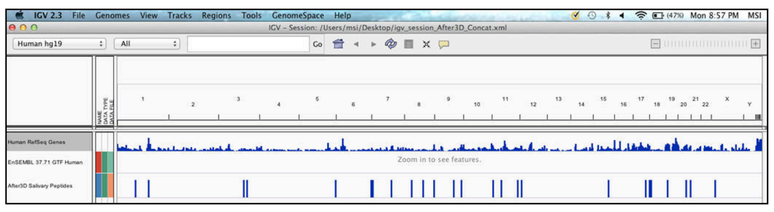
INTEGRATED WORKFLOW



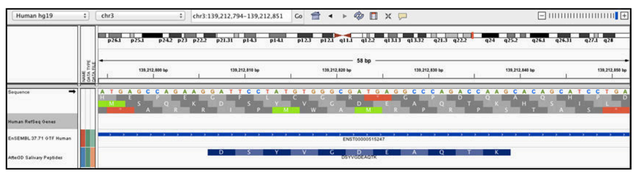
GENOMIC CONTEXT ANALYSIS

• DATASET 1: 3D-fractionated salivary dataset. ProteoMiner ™ 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.

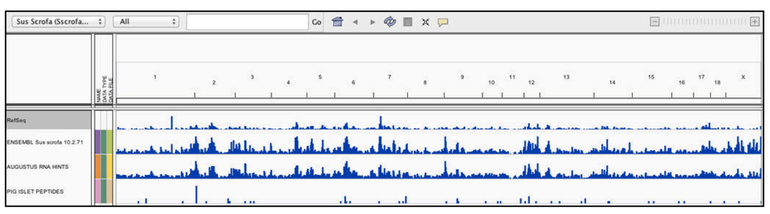


DSYVGDEAQTK : 67 spectra. Pseudogene. ENST00000415794_7

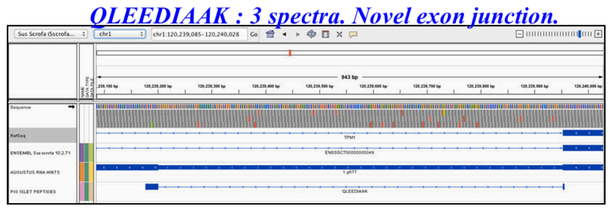


• DATASET 2: iTRAQ-labeled Pig Islet dataset. Three Replicates. 45 RAW Files. Orbitrap Velos (HCD) instrument. (de Jong *Unpublished*).

28 novel peptides. (See Poster 583 MP29)

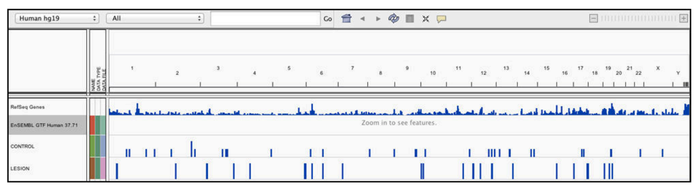


QLEEDIAAK : 3 spectra. Novel exon junction.

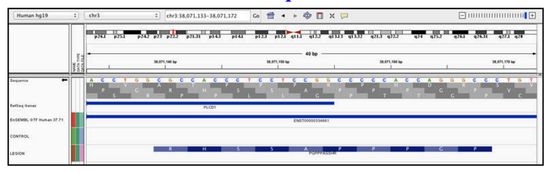


• DATASET 3: Oral pre-malignant lesion exudate dataset. 6 matched pairs (with controls). 84 RAW Files. LTQ-Orbitrap. (Kooren *Unpublished*)

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.

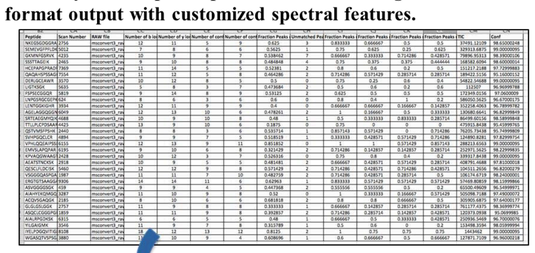


PGPPPASSHR : 1 spectrum. UTR.

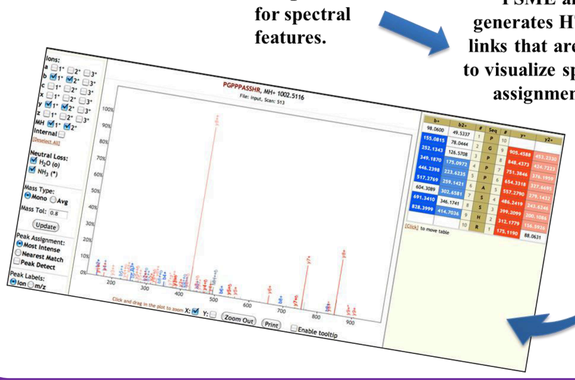


SPECTRAL ASSIGNMENT EVALUATION

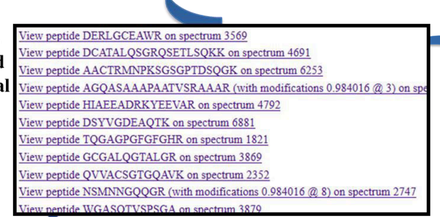
PSME (Peptide-Spectral-Match Evaluator) uses spectral summary as an input to parse mzml data and generates a tabular format output with customized spectral features.

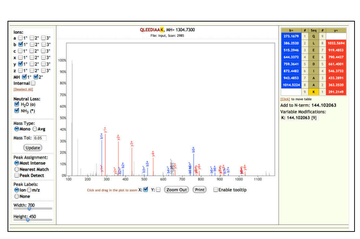


Tabular format can be filtered using thresholds for spectral features.



PSME also generates HTML links that are used to visualize spectral assignments.





• Interactive boxes can be used to change ion assignments and other parameters.

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.*