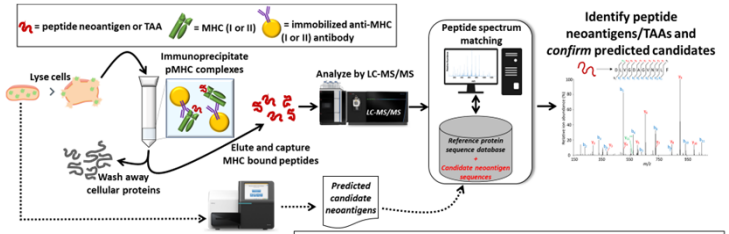
A COMPLETE, MODULAR AND FLEXIBLE PROTEOGENOMIC PIPELINE FOR PEPTIDE NEOANTIGEN DISCOVERY AND VERIFICATION

Tyler Jubenville[1], James E. Johnson[2], Sue Rathe[1], Flavia Popescu[1], Kyle Richards[1], Suzanne Coleman[1], Reid Wagner[2], Fengchao Yu[3], Alexey Nesvizhskii[3], Subina Mehta[4], Pratik Jagtap[4], David A. Largaespada[1], Timothy J. Griffin[4]

[1]Department of Pediatrics, U. of Minnesota; [2]Minnesota Supercomputing Institute, U. of Minnesota; [3]Department of Pathology and Department of Computational Medicine and Bioinformatics, U. of Michigan; [4]Department of Biochemistry, Molecular Biology and Biophysics, U. of Minnesota

BACKGROUND AND RATIONALE

- Immunotherapy studies in cancer seek to identify non-normal peptide sequences (**neoantigens**) derived from tumor-specific proteoforms, and also aberrantly expressed tumor-associated antigens (**TAAs**) from reference protein sequences, which are bound to the class I or II major histocompatibility complex (**MHC**) and activate the immune system
- Identification of neoantigens relies on next-generation sequencing (**NGS**) to identify genomic or transcriptomic alterations that may encode non-normal, candidate neoantigen peptides, coupled with algorithms that predict their binding to the MHC
- Mass spectrometry (**MS**)-based immunopeptidomics enrich the MHC and utilize LC-tandem mass spectrometry (**LC-MS/MS**) analysis to detect and identify MHC-bound peptides, directly confirming predicted candidates with potential for immunotherapy, including those carrying post-translational modifications (**PTMs**)



Overview of MS-based immunopeptidomics workflow

NEOANTIGEN DISCOVERY AND VERIFICATION: REQUIREMENTS AND SOLUTIONS

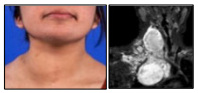
- Multi-omic integration of neoantigen prediction tools from NGS data with MS-based peptidomic analysis tools is a necessity
- Sequence database searching of immunopeptidomic MS/MS data to efficiently identify MHC-bound peptides using "no enzyme" constraints, identifying PTMs, and quantifying peptide abundance using label-free quantification (**LFQ**)
- Ideally these tools would be contained in a single environment, accessible by bench researchers complete with training resources to empower their adoption

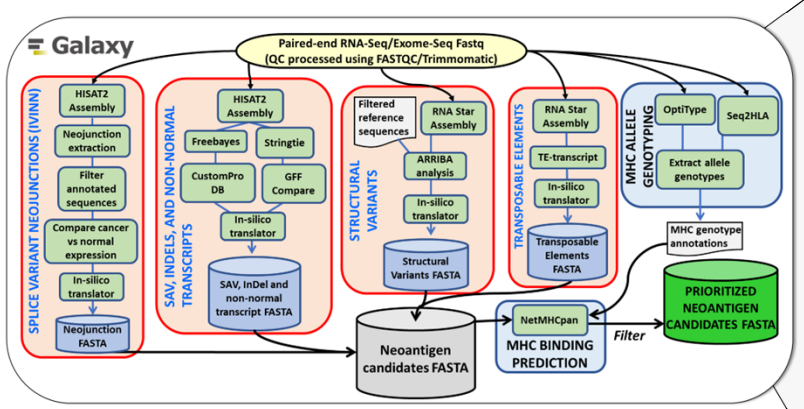
A solution: Integrated multi-omic analysis within the Galaxy ecosystem

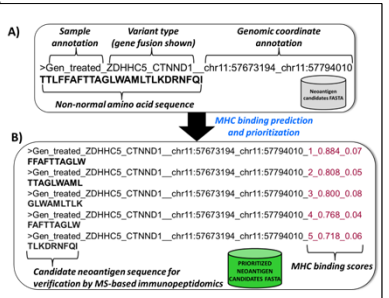
- The Galaxy ecosystem offers a platform to build multi-omic workflows integrating NGS analysis tools, MS/MS analysis tools, complete with access on publicly available instances, and training via the Galaxy Training Network
- Building on the Galaxy for proteomics (**Galaxy-P**) software suite, we are developing workflows for multi-omic neoantigen discovery and verification using state-of-the-art tools, including NGS analysis tools, the FragPipe tool suite (**fragpipe.nesvilab.org**) and PepQuery (**www.pepquery.org**)

RESULTS: NEOANTIGEN PREDICTION FROM NGS DATA

- **Developments guided by ongoing multi-omic study of neoantigens/TAAs from malignant peripheral nerve sheath tumors (MPNST) cells**

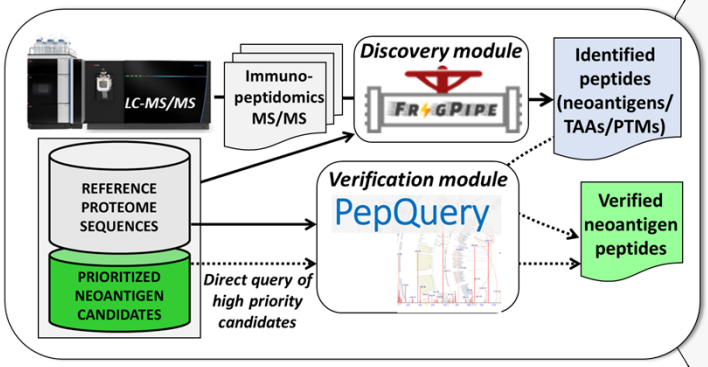




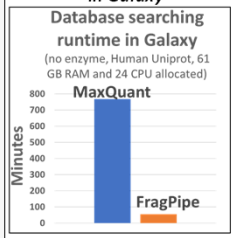


Example FASTA output of candidate neoantigen producing structural variant (A) and prioritization by MHC-binding prediction to generate an annotated FASTA database of prioritized candidates (B).

RESULTS: PEPTIDE NEOANTIGEN DISCOVERY AND VERIFICATION

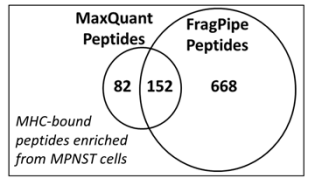


FragPipe performance in Galaxy



Example verified TAA peptides (MPNST cells)

Gene	Peptide	Description
MAGEA4	KVLEHVVRV	Melanoma-associated antigen 4
MAGEC2	GVYAGREHFV	Melanoma-associated antigen C2
MAGEA10	GLYDGMEHL	Melanoma-associated antigen 10
MAGEA2	QVMPKTGL	Melanoma-associated antigen 2
MAGEA3	EVDPIGHLY	Melanoma-associated antigen 3
PRAME	SIQSRYISM	Melanoma antigen expressed in tumors
SOX9	KLADQYPHL	Transcription factor SOX-10



FUTURE PLANS

- Optimize and harden workflows for production-ready deployment
- Publish optimized workflows and tools on accessible Galaxy gateways and in the Galaxy Tool Shed
- Develop Galaxy Training Network tutorials for online, on-demand access

Relevant References

Galaxy: Galaxy, C. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res* **50**, W345-W351 (2022).
Galaxy Training Network: Hiltemann, S. et al. Galaxy Training: A powerful framework for teaching! *PLoS Comput Biol* **19**, e1010752 (2023).
Galaxy-P tools: Chambers, M.C. et al. An Accessible Proteogenomics Informatics Resource for Cancer Researchers. *Cancer Res* **77**, e43-e46 (2017).
Arriba: Uhrig, S. et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res* **31**, 448-460 (2021).
CustomProDB: Wang, X. & Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29**, 3235-3237 (2013).
TE-transcripts: Jin, Y., Tam, O.H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593-3599 (2015).
Seq2HLA: Boegel, S., Bukur, T., Castle, J.C. & Sahin, U. In Silico Typing of Classical and Non-classical HLA Alleles from Standard RNA-Seq Reads. *Methods Mol Biol* **1802**, 177-191 (2018).
OptiType: Szolek, A. HLA Typing from Short-Read Sequencing Data with OptiType. *Methods Mol Biol* **1802**, 215-223 (2018).
NetMHCpan: Reynisson, B., et al. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **48**, W449-W454 (2020).
MS-Fragger/FragPipe: Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D. & Nesvizhskii, A.I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-520 (2017).
PepQuery: Wen, B. & Zhang, B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. *Nat Commun* **14**, 2213 (2023).

Acknowledgements

We thank the Minnesota Supercomputing Institute for maintenance of computational infrastructure used for development and testing, and the Center for Metabolomics and Proteomics at the University of Minnesota for support generating immunopeptidomic MS/MS data. This work is supported by NIH grant UG3CA244687 as part of the Immuno-oncology Translational Network (IOTN) program and by the Minnesota Masonic Cancer NIH grant P30CA077598

The authors declare no competing financial interest.