A modular, immunopeptidogenomic (iPepGen) analysis pipeline for discovery, verification, and prioritization of cancer peptide neoantigen candidates

Subina Mehta, Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota; smehta@umn.edu

Reid Wagner, Minnesota Supercomputing Institute, University of Minnesota; wagnerr@umn.edu

Katherine T. Do, Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota; doxxx217@umn.edu

James E. Johnson, Minnesota Supercomputing Institute, University of Minnesota; jj@umn.edu

Fengchao Yu, Department of Pathology, University of Michigan; yufe@umich.edu

Tyler Jubenville, Masonic Cancer Center, University of Minnesota; juben002@umn.edu

Kyle Richards, Department of Pediatrics, University of Minnesota; richa868@umn.edu

Suzanne Coleman, Masonic Cancer Center, University of Minnesota; colem491@umn.edu

Flavia E. Popescu, Masonic Cancer Center, University of Minnesota; pope0014@umn.edu

Alexey I. Nesvizhskii, Department of Pathology, University of Michigan, Department of Computational Medicine and Bioinformatics, University of Michigan; nesvi@med.umich.edu

David A. Largaespada, Masonic Cancer Center, University of Minnesota, Department of Pediatrics, University of Minnesota; larga002@umn.edu

Pratik D. Jagtap, Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Masonic Cancer Center, University of Minnesota; pjagtap@umn.edu

Timothy J. Griffin, Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Masonic Cancer Center, University of Minnesota; ; Corresponding Author

Abstract

Characterizing tumor-specific neoantigen peptides, derived from genomic or transcriptomic aberrations and presented to the immune system, is critical for immuno-oncology studies. To this end, the modular iPepGen immunopeptidogenomics pipeline provides these functions: (1) Neoantigen prediction and protein database generation from genomic or transcriptomic sequencing data; (2) Peptide identification (3) Verification from immunopeptidomic mass spectral data; (4) Neoantigen classification and visualization; (5) Candidate prioritization for further study. Deployed in the Galaxy ecosystem, iPepGen can be accessed via cloud-based, scalable resources and online, interactive training materials. iPepGen offers a unique solution to the growing community of researchers characterizing neoantigens for immuno-oncology studies.

Keywords

Neoantigen, immunopeptidomics, immuno-oncology, mass spectrometry, next generation sequencing, multi-omics, Galaxy, bioinformatics, workflow

Background

Neoantigen peptides, derived from non-reference proteoforms uniquely expressed in tumor cells, are an increasingly important class of targets in cancer immunotherapy. These proteoforms may arise from a variety of genomic, transcriptomic, and post-translational mechanisms, including non-synonymous mutations, gene fusions [1], aberrant splicing, de novo transcription from non-coding regions [2], derepression of transposable elements, post-translational modification (PTM) [3] , and translation of viral open reading frames in virus-associated cancers [4]. These alterations yield tumor-specific antigens absent from normal tissues and thus bypass central tolerance [5–8]. In addition to neoantigens, tumor-associated antigens (TAAs) and tumor-specific antigens (TSAs) derived from reference proteoforms whose expression is either enriched (TAAs) or only found in cancerous cells (TSAs) are of interest for immunotherapy development [9].

Once synthesized, aberrant proteoforms are degraded by cellular proteolytic machinery. In the cytosol, peptides of 8–11 amino acids are generated and presented by Major Histocompatibility complex (MHC) class I (called Human Leukocyte Antigen complex, HLA-A, -B, -C in humans)[10]. Meanwhile, longer peptides from extracellular proteins (10–25 amino acids) are presented via MHC class II (HLA-D family in humans)[11]. These complexes are displayed on the cell surface and recognized by T cells. It has been estimated that up to 10% of these peptides may elicit immunogenic responses when derived from non-reference sequences [12]. Despite this, tumors

can escape immune surveillance via multiple mechanisms [13]. However, when MHC-bound neoantigen peptides are successfully identified, they can be leveraged for immunotherapy, including adoptive T cell transfer, immune checkpoint blockade, and prophylactic vaccination[14–18].

Identifying and predicting neoantigen candidates from genomic and transcriptomic data. The ability to identify neoantigen candidates from genomic and transcriptomic sequences has advanced rapidly due to improvements in next-generation sequencing (NGS) and customized computational pipelines [8,19]. Numerous tools have been developed for identifying potential neoantigens encoded in tumor-specific alterations [20–23]. These tools typically utilize RNA-Seq data or whole-exome sequencing to genotype alleles coding for MHC protein subunits and predict binding affinities of potential neoantigen peptides to the MHC based on biochemical and structural modeling [24–27]. Higher predicted binding affinity to specific MHC (or HLA) alleles is an indicator of immune recognition and the possibility of a peptide to elicit an immunogenic response.

Beyond prediction: confirming MHC-bound neoantigens via mass spectrometry (MS)-based immunopeptidomics. Although important, NGS-based approaches ultimately provide only a predictive view of antigen presentation, which presents limitations[8,28]. These include the inability to confirm proteoform translation, peptide processing, or MHC presentation. Predictions also fail to capture PTMs, which are known to modulate immunogenicity [29], are prone to prediction bias [30,31], and are often missed as antigenic peptides [32]. Empirical follow-up studies have shown that only a small fraction of predicted neoantigens from NGS data are processed, bound to the MHC, and truly immunogenic [17]. These constraints have motivated the development of mass spectrometry (MS)-based immunopeptidomics approaches that allow direct identification of peptides presented by MHC complexes [33–35].

In immunopeptidomics, the intact MHC and bound peptides are affinity-enriched and analyzed by liquid chromatography (LC) tandem mass spectrometry (MS/MS). Tandem mass spectra are then matched to custom protein databases that include both canonical reference sequences and tumor-specific neoantigen candidates inferred from NGS data. This strategy allows the detection of neoantigens and TAAs/TSAs in an unbiased, high-throughput manner, including peptides with PTMs that might influence immunogenicity [29]. Originally developed by Hunt and colleagues [36], MS-based immunopeptidomics has matured and now enables the routine detection of thousands of MHC-bound peptides from cellular and tissue sources [35,37], significantly expanding the scope of immuno-oncology research [8,34].

Informatics requirements for neoantigen prediction and confirmation via immunopeptidogenomics. The integration of NGS and MS data for comprehensive neoantigen

discovery, appropriately called "immunopeptidogenomics" by Purcell and colleagues [38], poses unique informatics requirements and challenges. These include: managing the heterogeneity of NGS alterations for predicting neoantigens and constructing sample-specific databases; efficient and sensitive peptide spectrum match (PSM) generation from MS/MS searched against customized protein sequence databases using non-specific enzyme constraints [29,39,40], and, if desired, analyzing PTMs; quantifying identified peptides via label-free methods [41,42].

Some pipelines have been described that focus on the analysis of MS-based immunopeptidomics data [43–45]. To varying degrees, these bioinformatic platforms offer the ability to identify potential genomic/transcriptomic aberrations from NGS data and utilize tools for predicting MHC binding to help in prioritization. Some of the platforms [43,44] use publicly available, established software such as MaxQuant [46] or commercial algorithms [38] for generating PSMs from MS/MS data; however, none of these algorithms are optimized for analysis of MHC-bound peptides using "no enzyme" settings, rather than peptides generated by trypsin digestion, as is common in proteomics applications. Recently, a sophisticated multi-omics platform called NeoDisc has been described that is openly available for non-commercial use [47]. NeoDisc offers a complete pipeline with modules for neoantigen prediction, peptide identification from immunopeptidomic MS/MS data, and advanced methods for prioritization of candidates for further study. However, it requires substantial local computational infrastructure for implementation and lacks a GUI, limiting its accessibility. Moreover, the addition of new or experimental tools into such pipelines requires software developer expertise, beyond the reach of most non-expert users.

The iPepGen pipeline and the Galaxy ecosystem as a unique solution for immunopeptidogenomic informatics. To address current limitations, we have developed the modular immunopeptidogenomic (iPepGen) pipeline for the discovery and characterization of neoantigens. Deployment in the Galaxy ecosystem [48,49] provides a web-based environment and user interface, immediately accessible via cloud-hosted servers [50], that supports reproducible bioinformatics analyses. Tools are openly available via the Galaxy Tool Shed and can be deployed on public cloud servers or institutional high-performance computing (HPC) resources. Galaxy supports comprehensive NGS and MS-based proteomics workflows [51], with built-in provenance tracking, visualizations, and shareable software workflows and analysis histories. The powerful Galaxy Training Network (GTN) [52,53] further promotes community adoption by offering interactive training resources for new users. Leveraging these strengths, iPepGen integrates validated tools for neoantigen prediction, peptide identification and verification, genomic mapping, and immunogenicity assessment within a unified and accessible framework complete with training resources, lowering the bar for adoption by new users. These collective aspects make the iPepGen pipeline a unique informatics resource built to

advance studies by a growing community of immuno-oncology researchers using immunopeptidogenomics for neoantigen discovery and characterization.

Results

We implemented the iPepGen pipeline within the Galaxy ecosystem to support end-to-end neoantigen discovery by integrative analysis of NGS and LC-MS/MS data. To demonstrate the effectiveness and reproducibility of the workflow, we analyzed paired RNA-Seq and MHC class I immunopeptidomics data from a malignant peripheral nerve sheath tumor (MPNST) cell line, STS-26T.

A modular, cloud-based pipeline is easily accessed and used via a web-based interface. The iPepGen pipeline is composed of five modular workflows (**Figure 1**), deployed in the Galaxy bioinformatics environment. These workflows integrate software tools selected and optimized to accomplish specific data analysis tasks and generate results compatible with subsequent analysis by downstream modules in the pipeline. The workflows composing each module are freely available on the European Galaxy server, through the links shown below in **Table 1.** A Galaxy workflow contains all software tools for operation on appropriate input data, along with preset, optimized settings. We have made example input files available for each workflow via Zenodo links (Table 1), which can be uploaded and recognized as appropriate input formats by the corresponding workflow. Once a workflow is run, Galaxy saves the analysis as a history, which contains all input and result files generated by the workflow. The history can be shared, re-run using adjusted settings, and/or archived. Links to completed histories for each iPepGen module are also shown in Table 1, using the example MPNST data. Workflows and histories can be imported into any Galaxy account once a user has registered on the server.

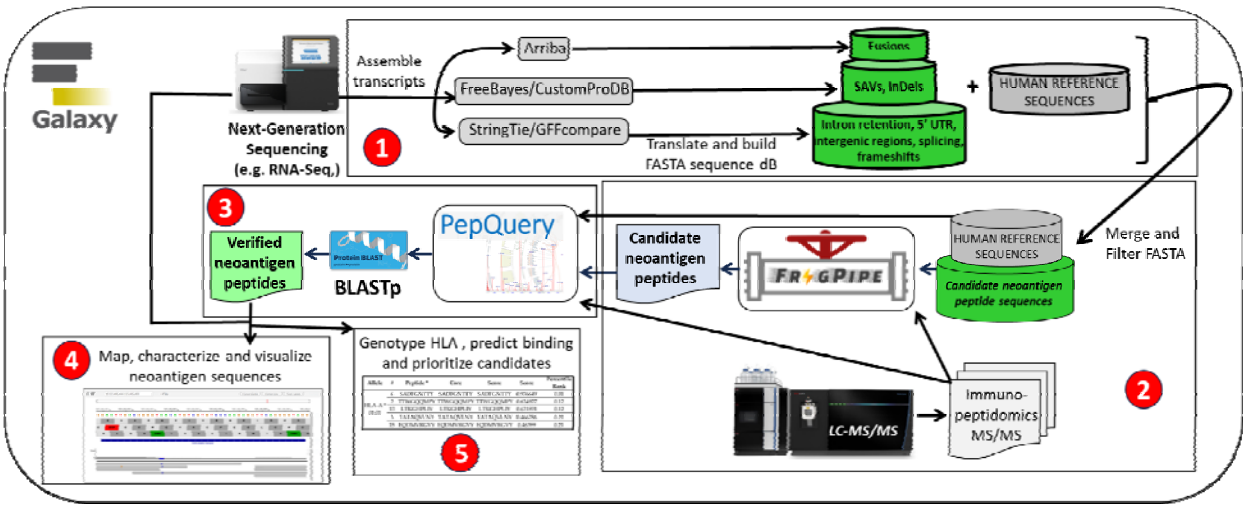


Figure 1. The modules comprising the Galaxy-based iPepGen pipeline. The modules carryout the following functions: 1) Predict neoantigen candidates from NGS data and generate customized protein sequence databases, including reference and non-reference neoantigen candidate sequences; 2) Discover neoantigen peptide candidates by sequence database searching of tandem mass spectrometry (MS/MS) immunopeptidomics data; 3) Verify discovered peptide candidates through a secondary peptide-centric evaluation against the MS/MS dataset; 4) Visualize and classify the nature of verified neoantigen peptides encoded by the genome and/or transcriptome; 5) Prioritize neoantigens for further exploration and empirical validation.

Table 1. Links for workflows, input data, analysis histories and training resources for each module of the iPepGen analysis pipeline

Module	Description	Galaxy workflow	Input data for the workflow	Galaxy history	GTN Training resources
1. Predict non-reference neoantigen peptide sequences	Predicts neoantigen peptide sequences encoded by non-reference genes and/or transcripts and generates customized protein sequence databases	***Gene fusion neoantigens*** https://tinyurl.com/fusion-workflow	***Gene fusion neoantigens*** https://zenodo.org/records/14365542	***Gene fusion neoantigens*** https://tinyurl.com/fusion-history	https://tinyurl.com/fusion-gtn
		General non-reference sequence neoantigens https://tinyurl.com/non-reference-workflow	***General non-reference sequence neoantigens*** https://zenodo.org/records/14372521	***General non-reference sequence neoantigens*** https://tinyurl.com/non-reference-history	https://tinyurl.com/non-reference-gtn
2. Discover candidate neoantigen peptide sequences and tumor-associated antigens	Merges customized and reference databases and uses the FragPipe tool suite to discover PSMs from LC-MS/MS immunopeptidomics data	https://tinyurl.com/discovery-workflow	https://zenodo.org/records/14374118	https://tinyurl.com/discovery-history	https://tinyurl.com/discovery-gtn
3. Verify PSMs	Verifies neoantigen	https://tinyurl.com/ve	https://zenodo.org/r	https://tinyurl.com/	https://tinyurl.com/v

for neoantigen candidates	peptide sequences identified in Module 2 using a secondary spectral matching algorithm PepQuery2	rification-workflow	ecords/15090517	verification-history	erification-gtn
4. Visualize and classify verified neoantigen sequences	Maps selected neoantigen peptide sequences to genes and transcripts, classifies the nature of non-reference sequences, and enables visualization	https://tinyurl.com/peptide-annotation-workflow	https://zenodo.org/records/14375693	https://tinyurl.com/peptide-annotation-history	https://tinyurl.com/peptide-annotation-gtn
5. Prioritize neoantigens for further study	Predicts binding of peptides to the HLA alleles and assesses possible immunogenicity, prioritizing candidates for further experimental evaluation	***HLA allele genotyping*** https://tinyurl.com/hla-genotyping-workflow	***HLA allele genotyping*** https://zenodo.org/records/14375917	***HLA allele genotyping*** https://tinyurl.com/hla-genotyping-history	https://tinyurl.com/hla-genotyping-gtn
		Peptide-HLA binding prediction https://tinyurl.com/binding-prediction-workflow	***Peptide-HLA binding prediction*** https://zenodo.org/records/14377365	***Peptide-HLA binding prediction*** https://tinyurl.com/binding-prediction-history	https://tinyurl.com/binding-prediction-gtn

Online, on-demand, and interactive training resources lower barriers for new users. The iPepGen pipeline takes advantage of the powerful GTN [52], which offers online, on-demand, and interactive training resources for new users. The GTN is seamlessly connected with workflows and software available on public Galaxy servers, thus providing a means for new users to adopt even sophisticated software for use in their studies. We developed an accessible "Learning Pathway" [54] within the GTN, which summarizes training resources for each module in the iPepGen pipeline, and leads users to the online documentation. The last column of Table 1 shows the links to training materials that walk new users step-by-step through each workflow, complete with links to demonstration data and software. Once a user has completed these training workflows, they are equipped with the necessary knowledge to upload their data to their Galaxy account and begin their own meaningful immunopeptidogenomic analysis.

Description of analysis modules and representative results. We describe the main functionalities and operation of each module in the iPepGen pipeline, along with results from MPNST data that demonstrate its effectiveness for immunopeptidogenomic analysis. The Methods section provides a more detailed description of software tools and key settings used in each module.

Module 1: Prediction of non-reference neoantigen peptide sequences

To demonstrate the functionality and output of the iPepGen pipeline, we analyzed RNA-Seq and immunopeptidomics data from an MPNST cell line. These demonstration analyses used the workflows and input data shown in Table 1 and produced the analysis histories shown in this table as well. For Module 1, we utilize several complementary workflows seeking to cast a wide net for possible genomic or transcriptomic sequences that may code for novel proteoforms leading to neoantigen production and may be identified via matching to immunopeptidomic MS/MS data in downstream modules.

Gene Fusion Variant Detection. One workflow targets gene fusions, a potential source of neoantigen peptides in many cancers [1], using the Arriba algorithm [55], integrated with the RNA STAR aligner [56] for spliced alignment and chimeric read detection. Outputs include a tabulated list of high-confidence fusions with supporting read metrics, visual summaries of each event in PDF format, and translated fusion junction sequences in FASTA format for downstream use (Table 1, Module 1, Item 21 in the history). Post-processing steps are applied to remove potential false positives based on read depth, mapping quality, and known benign fusion sequences. For the MPNST cell line RNA-Seq data used for demonstration, we identified 86 novel fusion sequences that were translated into potential neoantigen peptide sequences and outputted in the annotated FASTA formatted table (**Figure 2**).

Comprehensive Prediction of Non-Reference Protein Sequences. Complementing the targeted prediction of gene fusion-derived neoantigens, we employed additional workflows that broadly identify non-reference transcript sequences from RNA-Seq data. In one of these workflows, paired-end reads are aligned to the human reference genome using HISAT2 [57], followed by variant calling with FreeBayes. CustomProDB [58] then annotates predicted coding DNA sequence (CDS) variants, generating a set of non-reference protein sequences containing single amino acid variants (SAVs) and short insertion-deletions (InDels).

In the second workflow, transcript assembly is performed using StringTie [59], with transcript classification via GffCompare [60] to distinguish novel sequences compared to the reference genome. Novel, non-reference sequences, including frameshifts, novel splice events, and/or sequences from non-coding genomic regions, are translated in all three reading frames using Galaxy tools “Translate BED Transcripts” and “bed to protein map.”

The peptide sequences predicted from fusions, SAVs/InDels, and other non-reference transcripts are all merged with UniProt canonical human protein sequences and a common contaminant sequence (called the common Repository of Adventitious Proteins, cRAP) using “FASTA Merge Files and Filter Unique Sequences.” Non-reference entries were annotated with a “Generic” header, distinguishing them from UniProt canonical sequences, to support

downstream analysis in Module 2. For the RNA-Seq from the MPNST STS-26T cell line, these combined workflows generated a protein database containing approximately 2.9 million unique protein sequences (Table 1, Module 1, Item 48 in history). **Figure 2** shows the FASTA-formatted databases generated by each of the three described workflows used for the prediction of neoantigen peptides in this module. Notably, the cumulative number of entries across all individual databases exceeds that of the comprehensive database, as redundant or overlapping peptide sequences present in multiple workflow-specific outputs are collapsed into unique entries during database merging process, thereby reducing redundancy in the final merged FASTA database.

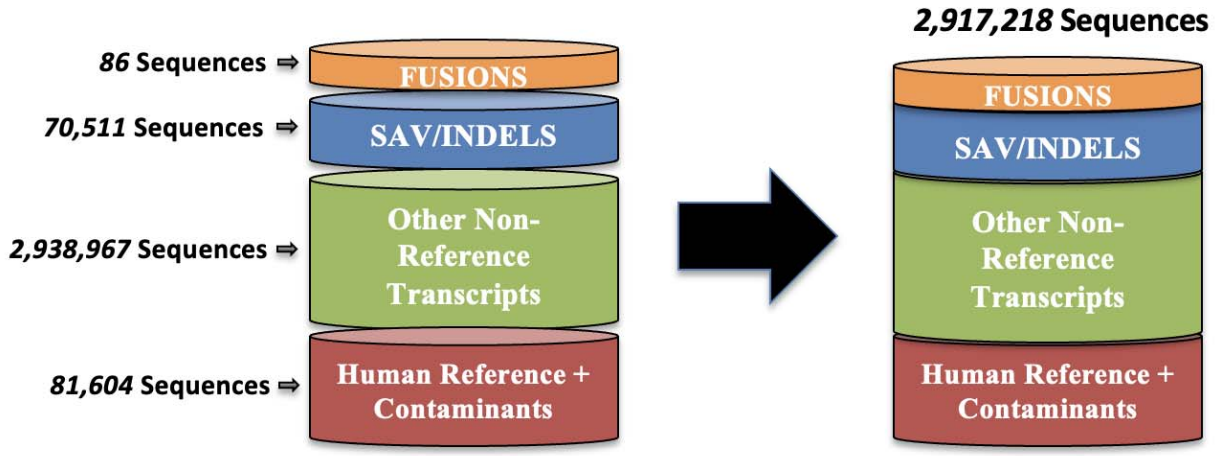


Figure 2. Relative size of databases from MPNST cell line data generated by Arriba (fusions), CustomProDB and StringTie workflows. The graphic illustrates the cumulative size of each non-reference database component and the final customized database after merging with the reference and common contaminant (cRAP) sequences.

Module 2: Discovery of candidate neoantigen peptide sequences and tumor-associated antigens

To identify candidate neoantigens presented by MHC I molecules, we analyzed MS/MS immunopeptidomics data from the same MPNST cell line used in Module 1. This module utilizes the FragPipe software suite [61], which is deployed as a tool in Galaxy. We matched MS/MS spectra against the customized FASTA database of ~2.9 million sequences generated previously (Figure 2). The workflow integrates MSFragger for PSM generation, with optional MSBooster rescoring [62], Percolator [63] for post-search validation, and Philosopher [64] for downstream filtering and false discovery rate (FDR) control at 1% for both PSM and protein levels. The "Nonspecific-HLA" search mode in FragPipe is used to identify MHC I-bound peptides ranging from 8 to 20 amino acids in length. To accommodate the large search space, the workflow automatically partitions the protein database into 200 random sections for efficient parallel processing. The Query Tabular Galaxy tool [65] extracts peptide sequences from the tabular outputs of FragPipe that did not match

any canonical reference proteins, generating a new tabular file of candidate neoantigen peptides for further analysis. Users may further stratify these peptides by variant type (e.g., SAVs, InDels, gene fusions etc.) by filtering on tag-based sequence annotations embedded during FASTA construction in Module 1. Using a single MS/MS file from the MPNST cell line STS-26T, and the comprehensive sequence database generated in Module 1, we identified 30 unique peptide sequences that mapped exclusively to non-reference proteins (Table 1, Module 2, Item 21 in history).

One key aspect of immunopeptidomics is the requirement to identify peptides using non-specific enzyme settings, as opposed to traditional bottom-up proteomics, which is based on analysis of peptides generated using proteolysis with trypsin. With non-specific enzyme settings, the possible sequence search space increases significantly, especially when using large databases including non-reference protein sequences predicted from NGS data. This challenges sequence database search algorithms to sensitively and efficiently generate PSMs for neoantigen discovery. FragPipe offers an ultrafast algorithm, which, when coupled with deployment in Galaxy on scalable compute resources, overcomes these limitations. As a demonstration, we compared FragPipe with the popular MaxQuant software, evaluating their ability to identify MHC 1-bound peptides using non-specific enzyme settings (**Figure 3A**). Both tools were run on equivalent compute environments (8 cores, 64 GB memory). FragPipe identified a greater number of peptides at a global FDR of 1% while requiring significantly less runtime. This performance advantage can be attributed to MSFragger's ultrafast search capabilities, including database splitting, and enhanced scoring strategies and post-processing methods [66], making it particularly well-suited for high throughput neoantigen discovery.

We also evaluated the effects of using the MSBooster algorithm [62] in FragPipe, which utilizes machine-learning-based prediction of fragment ion intensities and retention times, followed by a re-scoring of potential PSMs to increase confident peptide identifications. **Figure 3B** shows the use of MSBooster on an example immunopeptidomics dataset from MPNST cells, generating PSMs against a large protein sequence database. MSBooster significantly increases the number of confident PSMs generated, including identification of candidate neoantigen sequences.

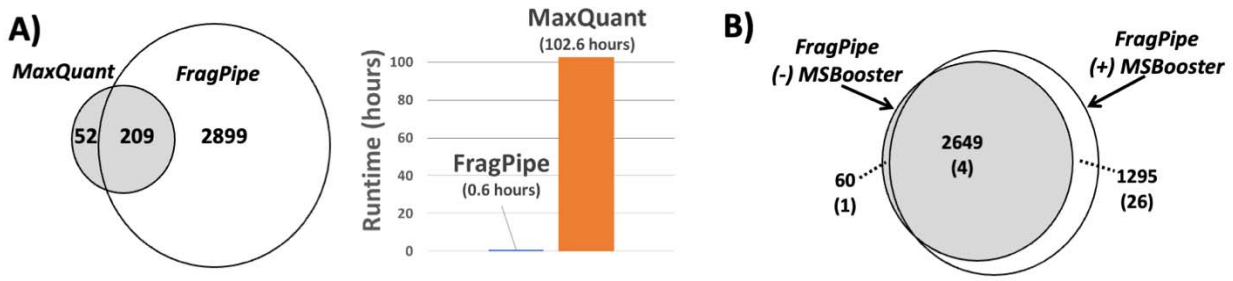


Figure 3. A. Comparison of FragPipe and MaxQuant performance for a "no-enzyme" database search of a single peptidomics MS/MS file against the reference human proteome database from the MPNST data in Module 1; both software were run using their stand-alone GUI using 8 cores and 64 GB of memory. **B.** Comparison of FragPipe run without and with MSBooster functionality on a single immunopeptidomics MS/MS datafile from MPNST cells and searched against the sequence database of 2.2 M sequences; number of unique peptides identified are shown, with number in parentheses () indicating the number of candidate neoantigen peptides identified with sequences that do not map to the reference proteome.

Module 3: Verification of neoantigen peptide-spectrum matches

To further validate candidate neoantigens identified in Module 2, we implemented a secondary peptide-centric verification step using PepQuery2 [67]. This approach addresses the potential for false-positive PSMs when searching large, non-reference databases, as previously noted in immunopeptidomics studies [68]. PepQuery2 allows for independent evaluation of individual PSMs by scoring candidate neoantigen sequences against MS/MS spectra in the presence of competing reference peptides, including common PTMs. This module also uses the BLAST-P tool for verifying that neoantigen candidates are truly novel and do not match any known reference sequences. As such, the module provides a secondary verification of candidate neoantigens identified using FragPipe against the large sequence database to ensure users are provided the most confident and high-value peptides to consider for further study.

For the demonstration MPNST immunopeptidomics dataset, we evaluated the 30 candidate neoantigen peptides from Module 2 using PepQuery2. Five neoantigen peptide sequences passed PepQuery2 hyperscore validation and BLAST-P filtering (Table 1, Module 3, Item 12 in history). These peptides represent the most confident neoantigen candidates ready for further characterization and prioritization for experimental evaluation.

Module 4: Visualization and classification of verified neoantigen sequences

This module facilitates the classification and genomic annotation of verified neoantigen peptides. The PepPointer tool [69] is used to map neoantigen peptides to their genomic coordinates using BED and GTF annotations and classify their genomic origin as coding (CDS), untranslated region (UTR), intron retaining, or intergenic sequences. PepPointer outputs a tabular file with a description of their likely genomic origin, as well as links to use with the Integrated Genome Viewer (IGV) [70] and UCSC Genome Browser [71], enabling researchers to visually inspect peptide-transcript-genome alignments. **Figure 4** shows an example output of a neoantigen candidate peptide from MPNST, mapped to the genome along with supporting transcript sequences that led to the predicted sequence that was verified in Modules 2 and 3.

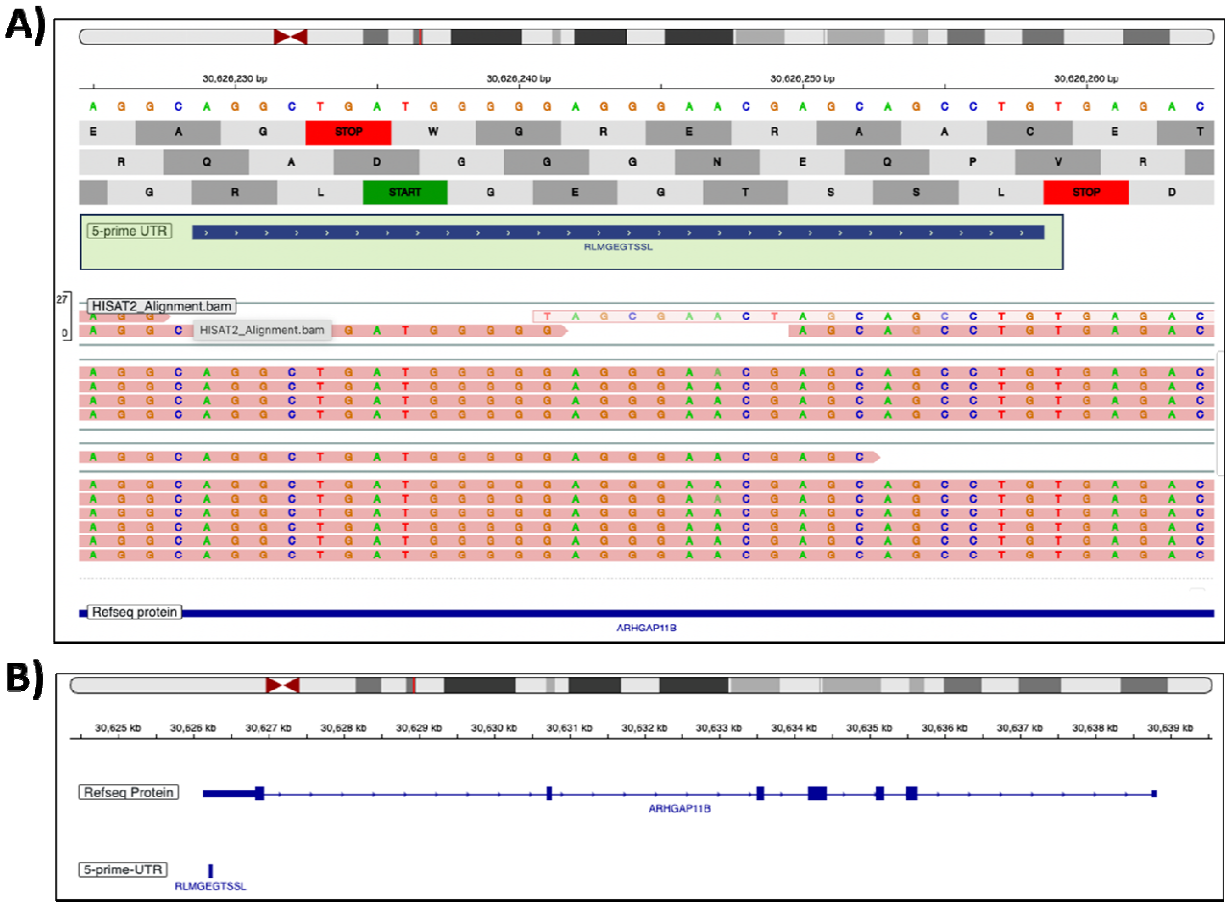


Figure 4. Example of a non-reference neoantigen peptide sequence (RLMGEGTSSL, shown in green shaded box) coded by a transcript sequence mapped to a portion of the 5' UTR of an annotated gene structure and visualized in the Integrated Genome Viewer (IGV). Assembled RNA-Seq reads coding this peptide are in pink; B) A larger view of the annotated protein coded by this genomic location (exon coding sequences shown in blocked blue) with the location of the novel neoantigen peptide mapped to the 5' UTR also shown as a reference.

PepPointer, a Galaxy tool which has been described in detail [69], is meant to offer a tabular format, initial classification of neoantigen sequences. The output enabling visualization using IGV and/or UCSC genome browsers offers users a means to more deeply understand the nature of identified neoantigens. In some cases, candidate neoantigen peptides arise from transcript coding sequences derived from complex processing of DNA or RNA sequences (e.g., unique combinations of intron and exon sequences), which are not easily annotated via PepPointer and may require manual exploration by the user for their advanced characterization. **Supplemental Figure 1** outlines one such example found in the five neoantigen candidates outputted in Module 4 for the MPNST demonstration data. **Table 2** summarizes the panel of five verified neoantigen peptides from the demonstration data. These peptides were supported by high FragPipe PSM score probabilities (≥0.92) and deemed confident by PepQuery2 verification with highly significant p-values (≤5.71E-03). Notably, peptides were derived from both canonical

(CDS, UTR) and non-canonical genomic coding regions (introns, intergenic), underscoring the potential of iPepGen to uncover promising neoantigen peptides from diverse genomic and/or transcriptomic aberrations.

Table 2. Verified and classified neoantigen peptides from MPNST demonstration data.

Peptide	Protein	Genomic location	Strand	Annotation	Fragpipe Probability	PepQuery p-value
FLARKPSAV	40S ribosomal protein S14	chr5:150447594-150447567	-	CDS	0.9996	1.27E-04
ILAQEIVKV	SUMO-activating enzyme subunit 1	chr19:47131854-47131881	+	intron	0.9365	3.80E-04
RLMGEGTSSL	Rho GTPase-activating protein 11A	chr15:30626228-30626258	+	five_prime_utr	0.976	1.06E-04
RVWQDAELGLY	Myc-associated zinc finger protein	chr16:29808646-29808679	+	CDS	0.939	1.00E-04
SPRSPSPSL	Uncharacterized	chr2:3570011-3569984	-	intergene	0.9254	5.71E-03

Module 5: Prioritization of neoantigens for further study

With verified neoantigen peptides in hand from the modules above, iPepGen offers a module to assist users in prioritizing those sequences that may be of highest interest for further evaluation in the lab. This prioritization is guided using software for predicting binding and potential recognition as immunogenic epitopes that may be of highest value for immunotherapies. Two workflows make up this module to enable this prioritization.

HLA Genotyping. As a first step to predicting peptide binding to the MHC (HLA binding in this case for human samples), HLA allele genotyping is performed on RNA-seq data. Paired-end reads in FASTQ format are analyzed using two complementary tools: OptiType [72], optimized for class I alleles, and Seq2HLA [73], which also covers class II loci, if desired. Outputs from the two tools are merged, and redundant entries are removed using Galaxy text manipulation tools. In the case of MPNST RNA-Seq data used for demonstration purposes, the final genotype set included eight HLA class 1 alleles (Table 1, Module 5, Item 13 in first history), which served as input for further neoantigen peptide-binding affinity prediction.

Peptide–HLA Binding Prediction. The genotyping results are used as one input to the IEDB MHC Binding Prediction tool [74], using the NetMHCpan_EL algorithm, trained on empirical immunopeptidomics data. The neoantigen peptides from module 3 are used as the other input. If the input peptides are 9 or more amino acids long, the software generates overlapping

shorter sequences of at least 8 amino acids and evaluates each for MHC binding, thereby expanding the list of potential peptide binders beyond the original input set. Peptide sequences are assigned a percentile rank for their binding to each HLA allele and classified as strong (percentile rank ≤ 0.5%) or weak binders (0.5–2%). Optionally, PepQuery2 can be used to verify whether any of the subset peptides created from the initial input sequences have matching MS/MS evidence from the original analysis. From the MPNST dataset, five peptides were predicted as strong binders (Table 1, Module 5, Item 21 in the second history), and three were classified as weak binders (Table 1, Module 5, Item 19 in the second history) to specific HLA alleles. The specific combination of predicted binding affinity to specific HLA alleles can be used as an indicator of immune recognition [75] and provides a means to prioritize peptides for further experimental evaluation.

Discussion

iPepGen's design addresses key challenges unique to immunopeptidomics, including the use of personalized protein databases derived from NGS data, non-tryptic peptide identification and verification of neoantigens, and computational prioritization of candidate neoantigens. Beyond performance across these key aspects of immunopeptidogenomics, implementation in the Galaxy ecosystem distinguishes it from any other available informatics resources. Ready-to-use workflows comprising each module are immediately and freely available on a scalable, cloud-based computing infrastructure. Integration with online training resources provides a unique means to empower new users to adopt iPepGen in their studies and accelerate a wide range of immuno-oncology studies. Although we have developed high-performing workflows utilizing contemporary tools, each module is flexible to the integration of new tools and/or customization of existing software. This adaptability to new software integration is a hallmark of the Galaxy bioinformatics environment, which also offers a thriving community of developers and users, ensuring sustained maintenance and operation [49]. Galaxy can also be implemented on local HPC resources, giving users the option to maintain their own customized version of iPepGen if desired.

In **Module 1**, we have developed several workflows to assemble and annotate a diverse profile of non-reference genomic and transcriptomic sequences that may give rise to novel proteoforms, which, in turn, are potentially processed into neoantigen peptides. We have developed these workflows to build protein databases personalized to the user's sample(s) with a broad profile of potential neoantigen sequences, which are confirmed via integration with MS/MS data in Module 2. This module is flexible to customization with other approaches to identifying specific types of non-reference genomic and/or transcriptomic sequences from NGS data, including single cell, long read sequencing, or other "third generation sequencing" methods [76]. Galaxy has a well-established history as an environment amenable to the

integration of emerging software for genomic and transcriptomic sequencing analysis [77]. iPepGen offers compatibility with a wide array of analysis tools in this area, providing a means to build protein sequence databases containing potential neoantigen sequences for interrogation with immunopeptidomic MS/MS data.

The use of FragPipe in **Module 2**, and specifically the MSFragger algorithm, provides the backbone for matching immunopeptidomic MS/MS data to the customized sequence databases from Module 1. MSFragger's ability to efficiently handle "no enzyme" search settings is critical in this context, as MHC-bound peptides are often produced by a variety of proteases, unlike typical proteomic samples that are digested with trypsin. The expanded search space inherent to immunopeptidomic data substantially increases memory and processing requirements, especially when searching against a custom database containing over 2 million protein sequences. To accommodate this, we enabled MSFragger's "split database" functionality, dividing the database into 200 sections. Deployment of iPepGen in a scalable, cloud-based Galaxy instance is also key. The European Galaxy instance allocates up to 250 GB of active memory, which proved sufficient for executing MSFragger under these conditions. Notably, FragPipe also supports the identification of PTMs, such as N-terminal acetylation and pyro-glutamate formation, which are known to influence MHC binding [29]. Although we did not evaluate PTM analysis in depth for this study, we do suggest that users limit the number of variable PTMs in the settings to three, in order to keep limit active memory demands for non-specific enzyme PSM generation.

We also evaluated the MSBooster tool [62] in FragPipe, which utilizes machine-learning-based scoring to predict fragment ion intensities and retention times. After this prediction, potential PSMs are re-evaluated, which results in the rescue of many peptide identifications and enhances sensitivity. In our analysis, incorporating MSBooster increased the identification of MHC I-bound peptides by 48% and neoantigen candidates by more than fivefold (**Figure 3**) without extending runtime or memory consumption. We suggest turning on MSBooster in the Galaxy FragPipe tool to maximize PSMs for further consideration. We also note that FragPipe's IonQuant module [42] is available for label-free quantification by aligning peptides across LC-MS/MS runs and calculating abundances based on area-under-the-curve metrics. This feature allows researchers to incorporate quantitative dimensions into immunopeptidomics studies with minimal additional setup.

The current workflow focuses on the analysis of data-dependent acquisition (DDA) of peptide MS/MS data. With the advent of new, faster scanning instruments, data-independent acquisition (DIA) data is increasingly used for immunopeptidomics [78,79]. DIA offers potential for deeper detection of peptides in complex mixtures and more accurate and precise

quantification. The FragPipe suite offers tools for DIA analysis [80,81], which are readily available for analysis of these data types in our pipeline.

In **Module 3**, PepQuery2 offers a peptide-centric means to verify candidate neoantigen sequences through a secondary analysis of peptides discovered in Module 2. Unlike conventional sequence database search tools, PepQuery2 focuses exclusively on user-defined candidate neoantigen peptide sequences and evaluates their match quality to potential MS/MS spectra in the dataset against human reference peptides, while also accounting for common PTMs to these peptides. We did observe that using non-specific enzyme digestion increases memory demands for PepQuery2, such that limiting the amino acid length of reference peptides for consideration to a narrow window around the input neoantigen candidates was necessary for successful operation (see Methods for more details).

Module 4 provides an initial means to classify non-reference peptides based on the aberrant genomic or transcriptomic sequences coding for the identified sequence. The simple tabular output also provides links to be imported in the popular IGV or UCSC genome browser to visualize the genomic coordinates coding for the peptide, along with any supporting transcript sequences responsible for the translation product. As we note in Supplemental Figure 1, the complex nature of some aberrant genomic and/or transcriptomic sequences coding neoantigen peptides may require deeper manual investigation in order to understand the source of some peptide sequences.

The HLA genotyping tools in **Module 5**, OptiType and Seq2HLA, provide complementary strengths for sequence analysis of HLA alleles from NGS data. OptiType is optimized for Class I alleles, while Seq2HLA genotypes both Class I and Class II alleles. Use of these two tools offers more comprehensive genotyping of HLA alleles. While Module 5 provides basic prioritization functionality using prediction of binding affinity to the genotyped alleles as a factor in immunogenicity, more advanced machine-learning-based tools are increasingly available to predict immunogenicity from multi-omics datasets [22,82–84]. Integration of such tools—either natively within Galaxy or through API connections to external platforms—could enhance neoantigen prioritization and support translational applications in cancer immunotherapy.

Conclusion

The iPepGen pipeline offers several distinguishing qualities from other available informatics solutions for immunopeptidogenomic informatics:

- Immediate access to analysis modules on scalable cloud-based infrastructure, with the option to run specific modules or together as a complete pipeline for discovery,

verification, and prioritization of neoantigen peptides for further experimental evaluation

- Step-by-step, interactive training material for each module is encapsulated in a customized learning pathway [54], seamlessly linked to accessible workflows, empowering new users with little bioinformatic expertise to learn and use the pipeline for their studies.
- Workflows are composed of tools selected to meet key challenges and common bottlenecks of neoantigen discovery and validation, such as efficient identification of peptides processed by non-specific proteases and rigorous verification of candidate neoantigen peptides using two algorithms (FragPipe and PepQuery2).
- Extensibility of core software tools composing workflows, flexible to adaptation to new developments in multi-omic technologies and data types, and extendibility of workflow modules to new software as it emerges to meet evolving user needs and/or enable user customization.
- Promotion of reproducibility and transparency via sharing of customized workflows and completed analysis histories for use by others and/or as an accessible record to accompany method descriptions in peer-reviewed publications

Taken together, these features of iPepGen make it a timely resource for the growing community of researchers conducting immunopeptidogenomics analyses. It offers an important contribution to immuno-oncology studies seeking improved therapies for treating and preventing a wide range of cancers.

Methods

Aim, Design, and Setting of the Study

This study aimed to develop and validate iPepGen—a modular, end-to-end, Galaxy-based immunopeptidogenomic pipeline designed to enable accessible and reproducible identification, verification, and prioritization of neoantigen peptides from cancer multi-omics data. The design integrates genomic, transcriptomic, and mass spectrometry (MS)-based immunopeptidomics data. Development of each module and associated workflows, along with demonstration data analysis, was conducted on the European Galaxy server [85].

Description of Materials and Data

For data demonstrating the effectiveness of iPepGen analysis modules, we used the STS-26T malignant peripheral nerve sheath tumor (MPNST) cell line. Total RNA was extracted and sequenced on an Illumina platform, yielding 100 bp paired-end reads (~80 million

reads/sample). Cells were treated with gentamicin (50 µg/mL) to promote ribosomal read-through and enhance expression of non-reference proteoforms [86]. MHC class I complexes were immunoprecipitated using the W6/32 monoclonal antibody, and bound peptides were analyzed on an Orbitrap Eclipse mass spectrometer using data-dependent acquisition (DDA) mode [87,88]. Using RNA-Seq and immunopeptidomics data from an MPNST cell line as a representative of user data inputs and analysis requirements, we describe the functionality and show representative results demonstrating the effectiveness of the iPepGen pipeline. **Table 1** provides links to the raw input data generated from the MPNST cells, as well as completed analysis histories.

iPepGen Pipeline Modules

The links in Table 1 provide access to the Galaxy software tools composing workflows of each iPepGen module, along with validated and optimized settings for their operation. Here, we describe briefly the core tools used in each workflow and the settings for specific tools necessary for effective operation and performance.

Module 1: Prediction of Non-reference Neoantigen Peptides. Paired-end RNA-Seq reads are aligned to the GRCh38 reference genome using HISAT2 [57]. In one workflow, gene fusion events are identified using Arriba [55], which utilizes spliced alignments from STAR [56] to detect chimeric reads. For broader non-canonical transcript detection, including single-nucleotide variants (SNVs), indels, and aberrant transcripts, a workflow using FreeBayes [89] was employed for variant calling coupled with CustomProDB [58], generating a protein database of SAV and short InDel sequences. A parallel workflow uses StringTie [59] and GffCompare [60] for transcript assembly and annotation of non-reference sequences, followed by the use of the Galaxy tool "Translate BED transcripts" and "bed to protein map" to generate a customized FASTA database. For both the CustomProDB and StringTie-based workflows, permissive thresholds for Reads Per Kilobase per Million mapped reads (RPKM) were used (Value of 1 for CustomProDB and 2 for StringTie) to retain low-abundance transcripts that may encode tumor-specific proteoforms. Customized protein databases with predicted non-reference sequences are merged along with the human reference sequences from UniProt [90] and common contaminant proteins using the Galaxy tool “fasta_merge_files_and_filter_unique_sequences”, generating a comprehensive, annotated FASTA-formatted database of over 2.9M sequences for downstream analysis.

This module requires the installation and configuration of HISAT2, STAR, FreeBayes, StringTie, GffCompare, Arriba, and CustomProDB within Galaxy. Of these, FreeBayes is the most computationally intensive, requiring approximately 10 CPUs and at least 17 GB of RAM per sample. A single RNA-Seq dataset may take up to 30 hours to process. Other tools in the

module typically operate efficiently with 4–8 CPUs and 8–16 GB RAM, completing in less than 4–6 hours.

Module 2: Discovery of Candidate Neoantigen Peptides. In module 2, raw LC-MS/MS files (for demonstration data, .raw formatted files from the ThermoFisher Orbitrap instrument) can be converted to mzML format using msconvert [91] or directly analyzed in native format by the FragPipe software suite. For this module, FragPipe uses these software components: MSFragger [66] for sequence database searching under no-enzyme settings (allowing identification of peptides 8–20 amino acids long), MSBooster [62] for deep-learning-based prediction of fragmentation and retention time patterns, Percolator [63] for machine-learning-based PSM validation, and Philosopher [64] for false discovery rate (FDR) estimation using a target-decoy approach. For demonstration data from MPNST cells, MS/MS spectra generated on the Orbitrap LC-MS/MS system were matched to the merged protein FASTA database generated in Module 1, generating peptide spectrum matches (PSMs). For the demonstration data, only oxidation on methionine was included as a variable PTM. Non-reference peptides were first automatically selected based on lack of direct sequence match to a known reference protein provided in the tabular FragPipe peptide report, and then extracted with annotation using Galaxy's Query Tabular tool [65], based on header annotations in the FASTA file indicating their origin from the workflows in Module 1 (e.g., fusion, SAV/InDel, or novel transcript).

This module requires the Galaxy deployed FragPipe suite, msconvert, and Query Tabular. PSM generation using non-specific enzyme settings against a large customized sequence database of several million entries is memory-intensive. To address this challenge, it is important to configure FragPipe to use a database split of at least 200 to lessen the memory burden. Additionally, the Galaxy FragPipe tool was configured with an allocation of 16 CPUs and 250 GB RAM. Using these settings, a typical DDA file (~1 GB) was processed within 6 to 8 hours on the European Galaxy server.

Module 3: Verification of Neoantigen Peptide-Spectrum Matches. This module implements a second level of PSM validation using PepQuery2 [67]. Neoantigen candidate peptides identified in Module 2 input as a tabular list of sequences are re-evaluated against the raw MS/MS spectral data, converted to MGF format via msconvert, and the UniProt reference proteome. PepQuery2 conducts peptide-centric analysis by considering reference proteome sequences of close length and mass to the candidate neoantigen sequence, along with common PTMs (e.g. phosphorylation, deamidation, acetylation, etc), as potentially competing matches to MS/MS spectra that led to the identification of the non-reference sequence. PepQuery also generates thousands of randomized sequences of similar amino acid length and mass and evaluates potential matches to the MS/MS spectra under consideration. Through these evaluations, a hyperscore is assigned for sequences matching the MS/MS spectrum and statistical modeling to

assign p-values to the best sequence match to the spectrum is conducted. Verified neoantigen peptides are those that remain the top match to the MS/MS spectrum (as opposed to a reference sequence or a randomized sequence) and are assigned a p-value indicating strong statistical significance. PepQuery2 assigns a Yes or No Confidence to each input peptide based on these assigned values. In our workflow, neoantigen peptides deemed confident are further evaluated using BLAST-P against the NCBI reference database to confirm that these are truly novel sequences, with no exact matching sequences found in the reference proteome.

To ensure the performance of PepQuery2, when considering peptides from non-specific enzyme processing, using narrow amino acid length windows for reference sequences to consider is critical. For example, for neoantigen candidate peptide sequences nine amino acids in length, a window of 8–10 amino acids is suggested to minimize the number of reference peptides that need to be generated initially by the software, controlling active memory required. These settings provide the algorithm with thousands of reference peptides to consider as competing for matching to MS/MS. Using these optimized conditions, PepQuery2 completed the evaluation of 10 candidate sequences in under 20 minutes, using ≤8 GB RAM and 2 CPUs. BLAST-P queries using the Galaxy tool were also completed in under 30 minutes using similar compute allocations, making this module both computationally tractable and scalable.

Module 4: Visualization and Classification of Neoantigen Origins. This module utilizes the PepPointer tool [69] to map non-reference neoantigen peptides to their genomic coordinates using BED and GTF annotations assigned to the sequences. It also classifies their origin as being from coding DNA sequence (CDS), untranslated regions (UTR), intron-retained, or other intergenic sequences. The tabular output provides links with necessary information to be input into the IGV [70] or the UCSC Genome Browser [71], enabling researchers to visually inspect peptide-transcript-genome alignments. Galaxy's Query Tabular tool was used to format, annotate, and extract relevant features from intermediate outputs and generate tabular results with formats compatible with PepPointer. The tools in this module are not compute-intensive, typically completing in <15 minutes using 1–2 CPUs and ≤4 GB RAM, even when analyzing multiple peptide entries.

Module 5: Prioritization of Neoantigens via HLA Binding Prediction. HLA genotyping is performed using two complementary tools: OptiType [72], optimized for class I alleles, and Seq2HLA [73], which also covers class II loci. Paired RNA-Seq fastq files serve as input, and genotype outputs are merged using AWK in Galaxy and formatted via Query Tabular into tabular formats compatible with the IEDB MHC Binding Prediction tool [74]. The NetMHCpan_EL algorithm, trained on empirical data for eluted peptides from immunopeptidomics studies, predicts binding affinity of verified neoantigen peptides (which

are provided as input) to specific HLA alleles. Peptides are classified as strong (binding affinity percentile rank ≤ 0.5%) or weak binders (percentile rank 0.5–2%) to specific HLA sequences. Peptides of 9 or more amino acids are processed into shorter sub-sequences by the IEDB tools, and their binding affinity to HLA alleles is also assessed. Optionally, PepQuery2 can be used to verify whether these shorter sub-sequences were present in the original MS/MS dataset.

This module requires Galaxy-deployed OptiType, Seq2HLA, the IEDB binding prediction software, PepQuery2 (optional), and native Galaxy text processing tools (AWK, Query Tabular). OptiType required 8 CPUs and 32 GB RAM, with runtimes of ~32 hours, while Seq2HLA required 4 CPUs and 16 GB RAM, completing in ~24 hours. The IEDB predictor operated efficiently with 2–4 CPUs and ≤8 GB RAM, processing ~100 peptides in ~30 minutes.

Abbreviations

Browser Extensible Data (BED) formatted file
common Repository of Adventitious Proteins (cRAP)
coding DNA sequence (CDS)
data-dependent acquisition (DDA)
data-independent acquisition (DIA)
eluted ligand (EL)
false-discovery rate (FDR)
Galaxy Training Network (GTN)
Gene Transfer File (GTF)
high-performance computing (HPC)
human leukocyte antigen (HLA)
immunopeptidogenomics (iPepGen)
Insertions and Deletion (InDel)
Integrated Genome Viewer (IGV)
label-free quantification (LFQ)
liquid chromatography (LC)
major histocompatibility complex (MHC)
malignant peripheral nerve sheath tumor (MPNST)
Mascot generic format (.mgf)
mass spectrometry (MS)
National Library of Medicine-National Center for Biotechnology Information (NLM-NCBI)
Next-generation sequencing (NGS)
post-translational modification (PTM)
reads per kilobase per million mapped reads (RPKM)
single amino acid variants (SAVs)

tandem mass spectrometry (MS/MS)
tumor-associated antigen (TAA)
Tumor-specific antigen (TSA)
untranslated region (UTR)

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All of the software tools composing the iPepGen pipeline and its workflow modules are available through the Galaxy Tool Shed [92]. The Tool Shed page for each software tool also provides links to GitHub repositories with source code and documentation, along with license information. Most software used in the iPepGen modules is developed using permissive licensing (e.g., GNU GPL, MIT). Some of the tools contained within the FragPipe software platform (MSFragger and the associated IonQuant software are developed under an academic license and are available freely for academic research, non-commercial, or educational purposes. Other software contained in FragPipe is developed under permissive licensing [93]. The Galaxy bioinformatics platform is freely available and open-source, released under the MIT license [94]. All data, including representative input files, intermediate processed results, and final outputs, are available on the publicly accessible European Galaxy server, as described in Table 1.

Competing interests

A.I.N. and F.Y. receive royalties from the University of Michigan for the sale of MSFragger, IonQuant, and diaTracer software licenses to commercial entities. All license transactions are managed by the University of Michigan Innovation Partnerships office, and all proceeds are subject to the university technology transfer policy. D.A.L. is the co-founder and co-owner of NeoClone Biotechnologies, Inc., Discovery Genomics, Inc. (recently acquired by Immusoft, Inc.), B-MoGen Biotechnologies, Inc. (recently acquired by Bio-Techne Corporation), and Luminary

Therapeutics, Inc. The business of all these companies is unrelated to the contents of this abstract.

Funding

The work reported here was funded in part by National Institutes of Health (NIH) grants UH3CA244687 to D.A.L. and U01CA288888 to T.J.G., P.D.J., and A.I.N., and American Cancer Society Research Professor Award (RP-17-216-06-COUN to D.A.L.). Data used for testing and demonstrating the software described here was generated using an Orbitrap Eclipse instrumentation platform purchased through High-end Instrumentation Grant S10OD028717 from the NIH.

Authors' contributions

S.M. conceptualized the study, curated data, performed formal analysis, developed software, created visualizations, and drafted and revised the manuscript. R.W. contributed to data curation, software development, and manuscript review. K.T.D. participated in data curation, formal analysis, software development, and manuscript editing. J.E.J. contributed to conceptualization, data curation, and software implementation. F.Y. provided data resources, contributed to software development, and participated in manuscript review. T.J. supported data acquisition and manuscript review. K.R. contributed resources, software support, and manuscript editing. S.C. provided resources and reviewed the manuscript. F.P. assisted in data curation, resource provision, and manuscript editing. A.I.N. contributed software, supervision, funding acquisition, and manuscript review. D.A.L. led conceptualization, resource acquisition, supervision, funding, and manuscript review. P.D.J. contributed to conceptualization, software development, supervision, funding acquisition, and manuscript review. T.J.G. led conceptualization, software development, supervision, and funding acquisition and contributed to drafting and revising the manuscript.

All authors read and approved the final manuscript.

Acknowledgements

We thank the Center for Metabolomics and Proteomics at the University of Minnesota for providing services to generate MS-based immunopeptidomics data used for developing, testing, and producing training resources for the iPepGen pipeline as described in the manuscript. We also thank Dr. Bing Zhang (Baylor College of Medicine, Houston, TX) for advice on the use of PepQuery2 within our pipeline.

References:

1. Wang Y, Shi T, Song X, Liu B, Wei J. Gene fusion neoantigens: Emerging targets for cancer immunotherapy. Cancer Lett. 2021;506:45–54.

2. Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure J-P, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018;10:eaau5516.

3. Yi X, Liao Y, Wen B, Li K, Dou Y, Savage SR, et al. caAtlas: An immunopeptidome atlas of human cancer. iScience. 2021;24:103107.

4. Wang E, Aifantis I. RNA Splicing and Cancer. Trends Cancer. 2020;6:631–44.

5. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. Nat Methods. 2013;10:186–7.

6. Capietto A-H, Hoshyar R, Delamarre L. Sources of Cancer Neoantigens beyond Single-Nucleotide Variants. Int J Mol Sci. 2022;23:10131.

7. Xie N, Shen G, Gao W, Huang Z, Huang C, Fu L. Neoantigens: promising targets for cancer therapy. Signal Transduct Target Ther. 2023;8:9.

8. Ahn R, Cui Y, White FM. Antigen discovery for the development of cancer immunotherapy. Semin Immunol. 2023;66:101733.

9. Minati R, Perreault C, Thibault P. A Roadmap Toward the Definition of Actionable Tumor-Specific Antigens. Front Immunol. 2020;11:583287.

10. Archbold JK, Macdonald WA, Gras S, Ely LK, Miles JJ, Bell MJ, et al. Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition. J Exp Med. 2009;206:209–19.

11. Kaufman JF, Auffray C, Korman AJ, Shackelford DA, Strominger J. The class II molecules of the human and murine major histocompatibility complex. Cell. 1984;36:1–13.

12. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature. 2017;541:321–30.

13. Kim SK, Cho SW. The Evasion Mechanisms of Cancer Immunity and Drug Intervention in the Tumor Microenvironment. Front Pharmacol. 2022;13:868695.

14. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nat Rev Clin Oncol. 2021;18:215–29.

15. Chen F, Zou Z, Du J, Su S, Shao J, Meng F, et al. Neoantigen identification strategies enable personalized immunotherapy in refractory solid tumors. J Clin Invest. 129:2056–70.

16. Braun DA, Moranzoni G, Chea V, McGregor BA, Blass E, Tu CR, et al. A neoantigen vaccine generates antitumour immunity in renal cell carcinoma. Nature. 2025;639:474–82.

17. Zhang X, Goedegebuure SP, Chen MY, Mishra R, Zhang F, Yu YY, et al. Neoantigen DNA vaccines are safe, feasible, and induce neoantigen-specific immune responses in triple-negative breast cancer patients. Genome Med. 2024;16:131.

18. Redwood AJ, Dick IM, Creaney J, Robinson BWS. What's next in cancer immunotherapy? - The promise and challenges of neoantigen vaccination. Oncoimmunology. 11:2038403.

19. Andersen MH. Tumor microenvironment antigens. Semin Immunopathol. 2023;45:253–64.

20. Bais P, Namburi S, Gatti DM, Zhang X, Chuang JH. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. Bioinformatics. 2017;33:3110–2.

21. Chai S, Smith CC, Kochar TK, Hunsucker SA, Beck W, Olsen KS, et al. NeoSplice: a bioinformatics method for prediction of splice variant neoantigens. Bioinform Adv. 2022;2:vbac032.

22. Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, et al. pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. Cancer Immunol Res. 2020;8:409–20.

23. Vensko SP, Olsen K, Bortone D, Smith CC, Chai S, Beckabir W, et al. LENS: Landscape of Effective Neoantigens Software. Bioinformatics. 2023;39:btad322.

24. Dhanda SK, Mahajan S, Paul S, Yan Z, Kim H, Jespersen MC, et al. IEDB-AR: immune epitope database—analysis resource in 2019. Nucleic Acids Res. 2019;47:W502–6.

25. Kalemati M, Darvishi S, Koohi S. CapsNet-MHC predicts peptide-MHC class I binding based on capsule neural networks. Commun Biol. 2023;6:492.

26. Koşaloğlu-Yalçın Z, Blazeska N, Vita R, Carter H, Nielsen M, Schoenberger S, et al. The Cancer Epitope Database and Analysis Resource (CEDAR). Nucleic Acids Res. 2022;51:D845–52.

27. Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. PLoS Comput Biol. 2013;9:e1003266.

28. Becker JP, Riemer AB. The Importance of Being Presented: Target Validation by Immunopeptidomics for Epitope-Specific Immunotherapies. Front Immunol. 2022;13:883989.

29. Kacen A, Javitt A, Kramer MP, Morgenstern D, Tsaban T, Shmueli MD, et al. Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. Nat Biotechnol. 2023;41:239–51.

30. Zeng H, Gifford DK. Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design. Cell Syst. 2019;9:159-166.e3.

31. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. PLoS Comput Biol. 2018;14:e1006457.

32. Feola S, Chiaro J, Martins B, Russo S, Fusciello M, Ylösmäki E, et al. A novel immunopeptidomic-based pipeline for the generation of personalized oncolytic cancer vaccines. eLife. 11:e71156.

33. Gfeller D, Liu Y, Racle J. Contemplating immunopeptidomes to better predict them. Semin Immunol. 2023;66:101708.

34. Shapiro IE, Bassani-Sternberg M. The impact of immunopeptidomics: From basic research to clinical implementation. Semin Immunol. 2023;66:101727.

35. Terai YL, Huang C, Wang B, Kang X, Han J, Douglass J, et al. Valid-NEO: A Multi-Omics Platform for Neoantigen Detection and Quantification from Limited Clinical Samples. Cancers (Basel). 2022;14:1243.

36. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, et al. Pillars article: Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. Science 1992. 255: 1261-1263. J Immunol. 2007;179:2669–71.

37. Rivero-Hinojosa S, Grant M, Panigrahi A, Zhang H, Caisova V, Bollard CM, et al. Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. Nat Commun. 2021;12:6689.

38. Scull KE, Pandey K, Ramarathinam SH, Purcell AW. Immunopeptidogenomics: Harnessing RNA-Seq to Illuminate the Dark Immunopeptidome. Mol Cell Proteomics. 2021;20:100143.

39. Sandalova T, Sala BM, Achour A. Structural aspects of chemical modifications in the MHC-restricted immunopeptidome; Implications for immune recognition. Front Chem. 2022;10:861609.

40. Capturing the diversity of protein modifications on presented tumor antigens. Nat Biotechnol. 2023;41:195–6.

41. Mehta S, Easterly CW, Sajulga R, Millikin RJ, Argentini A, Eguinoa I, et al. Precursor Intensity-Based Label-Free Quantification Software Tools for Proteomic and Multi-Omic Analysis within the Galaxy Platform. Proteomes. 2020;8:15.

42. Yu F, Haynes SE, Nesvizhskii AI. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. Mol Cell Proteomics. 2021;20:100077.

43. Li Y, Wang G, Tan X, Ouyang J, Zhang M, Song X, et al. ProGeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection. BMC Med Genomics. 2020;13:52.

44. Liu C, Zhang Y, Jian X, Tan X, Lu M, Ouyang J, et al. ProGeo-Neo v2.0: A One-Stop Software for Neoantigen Prediction and Filtering Based on the Proteogenomics Strategy. Genes (Basel). 2022;13:783.

45. Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat Commun. 2020;11:1759.

46. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–72.

47. Huber F, Arnaud M, Stevenson BJ, Michaux J, Benedetti F, Thevenet J, et al. A comprehensive proteogenomic pipeline for neoantigen discovery to advance personalized cancer immunotherapy. Nat Biotechnol. 2024;

48. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res. 2022;50:W345–51.

49. Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Res. 2024;52:W83–94.

50. Galaxy Platform Directory: Servers, Clouds, and Deployable Resources - Galaxy Community Hub [Internet]. [cited 2025 Mar 24]. Available from: https://galaxyproject.org/use/

51. Mehta S, Bernt M, Chambers M, Fahrner M, Föll MC, Gruening B, et al. A Galaxy of informatics resources for MS-based proteomics. Expert Rev Proteomics. 2023;20:251–66.

52. Hiltemann S, Rasche H, Gladman S, Hotz H-R, Larivière D, Blankenberg D, et al. Galaxy Training: A powerful framework for teaching! PLoS Comput Biol. 2023;19:e1010752.

53. Batut B, Hiltemann S, Bagnacani A, Baker D, Bhardwaj V, Blank C, et al. Community-Driven Data Analysis Training for Biology. Cell Syst. 2018;6:752-758.e1.

54. Learning Pathway: Prediction of potential neoantigens [Internet]. [cited 2025 Jun 17]. Available from: https://training.galaxyproject.org/training-material/learning-pathways/neoantigen.html

55. Uhrig S, Ellermann J, Walther T, Burkhardt P, Fröhlich M, Hutter B, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. 2021;31:448–60.

56. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

57. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.

58. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics. 2013;29:3235–7.

59. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.

60. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9:ISCB Comm J-304.

61. FragPipe [Internet]. FragPipe. [cited 2025 Jun 17]. Available from: https://fragpipe.nesvilab.org/

62. Yang KL, Yu F, Teo GC, Li K, Demichev V, Ralser M, et al. MSBooster: improving peptide identification rates using deep learning-based features. Nat Commun. 2023;14:4539.

63. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007;4:923–5.

64. da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang H-Y, Shanmugam AK, Mellacheruvu D, et al. Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods. 2020;17:869–70.

65. Johnson JE, Kumar P, Easterly C, Esler M, Mehta S, Eschenlauer AC, et al. Improve your Galaxy text life: The Query Tabular Tool. F1000Res. 2018;7:1604.

66. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods. 2017;14:513–20.

67. Wen B, Zhang B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. Nat Commun. 2023;14:2213.

68. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods. 2014;11:1114–25.

69. Kumar P, Johnson JE, McGowan T, Chambers MC, Heydarian M, Mehta S, et al. Discovering Novel Proteoforms Using Proteogenomic Workflows Within the Galaxy Bioinformatics Platform. Methods Mol Biol. 2025;2859:109–28.

70. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.

71. Raney BJ, Barber GP, Benet-Pagès A, Casper J, Clawson H, Cline MS, et al. The UCSC Genome Browser database: 2024 update. Nucleic Acids Res. 2024;52:D1082–8.

72. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014;30:3310–6.

73. Boegel S, Bukur T, Castle JC, Sahin U. In Silico Typing of Classical and Non-classical HLA Alleles from Standard RNA-Seq Reads. Methods Mol Biol. 2018;1802:177–91.

74. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017;199:3360–8.

75. Pagliuca S, Gurnari C, Rubio MT, Visconte V, Lenz TL. Individual HLA heterogeneity and its implications for cellular immune evasion in cancer and beyond. Front Immunol. 2022;13:944872.

76. Chen Z, He X. Application of third-generation sequencing in cancer research. Med Rev (2021). 2021;1:150–71.

77. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, et al. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010;Chapter 19:Unit 19.10.1-21.

78. diaPASEF Analysis for HLA-I Peptides Enables Quantification of Common Cancer Neoantigens - PubMed [Internet]. [cited 2025 Jun 17]. Available from: https://pubmed.ncbi.nlm.nih.gov/40044040/

79. Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries, Data-Independent Acquisition, and MS/MS Prediction - PubMed [Internet]. [cited 2025 Jun 17]. Available from: https://pubmed.ncbi.nlm.nih.gov/33845167/

80. Li K, Teo GC, Yang KL, Yu F, Nesvizhskii AI. diaTracer enables spectrum-centric analysis of diaPASEF proteomics data. Nat Commun. 2025;16:95.

81. Yu F, Teo GC, Kong AT, Fröhlich K, Li GX, Demichev V, et al. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. Nat Commun. 2023;14:4154.

82. Xia H, Hoang MH, Schmidt E, Kiwala S, McMichael J, Skidmore ZL, et al. pVACview: an interactive visualization tool for efficient neoantigen prioritization and selection. Genome Med. 2024;16:132.

83. Pyke RM, Mellacheruvu D, Dea S, Abbott C, Zhang SV, Phillips NA, et al. Precision Neoantigen Discovery Using Large-Scale Immunopeptidomes and Composite Modeling of MHC Peptide Presentation. Mol Cell Proteomics. 2023;22:100506.

84. Xu L, Yang Q, Dong W, Li X, Wang K, Dong S, et al. Meta learning for mutant HLA class I epitope immunogenicity prediction to accelerate cancer clinical immunotherapy. Brief Bioinform. 2024;26:bbae625.

85. Galaxy [Internet]. [cited 2025 Jun 17]. Available from: https://usegalaxy.eu/

86. Diop D, Chauvin C, Jean-Jean O. Aminoglycosides and other factors promoting stop codon readthrough in human cells. C R Biol. 2007;330:71–9.

87. Lim Kam Sian TCC, Goncalves G, Steele JR, Shamekhi T, Bramberger L, Jin D, et al. SAPrIm, a semi-automated protocol for mid-throughput immunopeptidomics. Front Immunol. 2023;14:1107576.

88. Pandey K, Ramarathinam SH, Purcell AW. Isolation of HLA Bound Peptides by Immunoaffinity Capture and Identification by Mass Spectrometry. Curr Protoc. 2021;1:e92.

89. freebayes/freebayes: Bayesian haplotype-based genetic polymorphism discovery and genotyping. [Internet]. [cited 2025 Jun 17]. Available from: https://github.com/freebayes/freebayes

90. Bowler-Barnett EH, Fan J, Luo J, Magrane M, Martin MJ, Orchard S, et al. UniProt and Mass Spectrometry-Based Proteomics-A 2-Way Working Relationship. Mol Cell Proteomics. 2023;22:100591.

91. Adusumilli R, Mallick P. Data Conversion with ProteoWizard msConvert. Methods Mol Biol. 2017;1550:339–68.

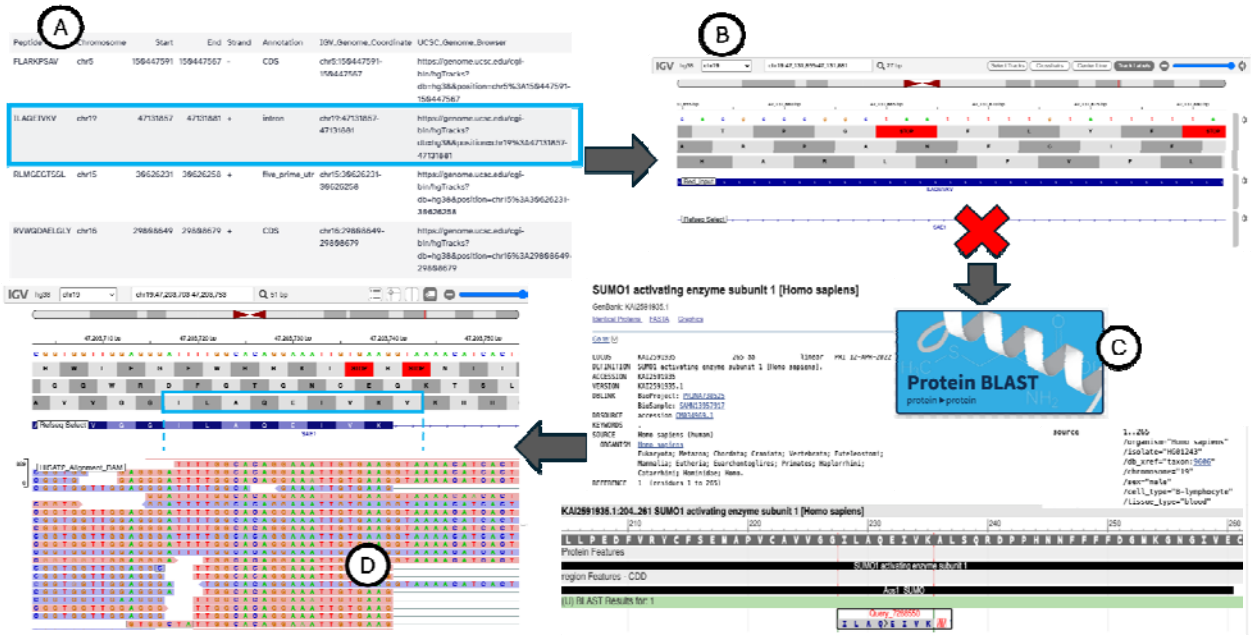
92. Galaxy | Tool Shed [Internet]. [cited 2025 Mar 24]. Available from: https://toolshed.g2.bx.psu.edu/

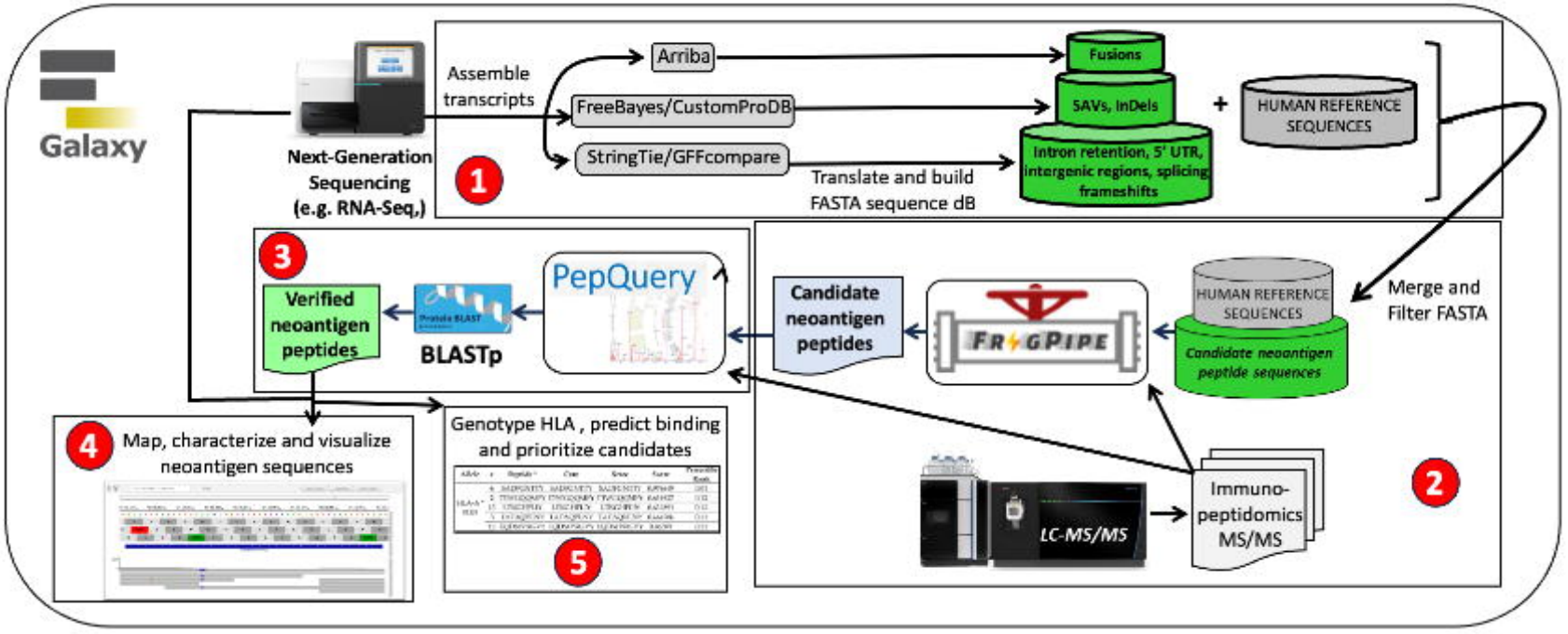
93. FragPipe/LICENSE at develop · Nesvilab/FragPipe [Internet]. GitHub. [cited 2025 Mar 24]. Available from: https://github.com/Nesvilab/FragPipe/blob/develop/LICENSE

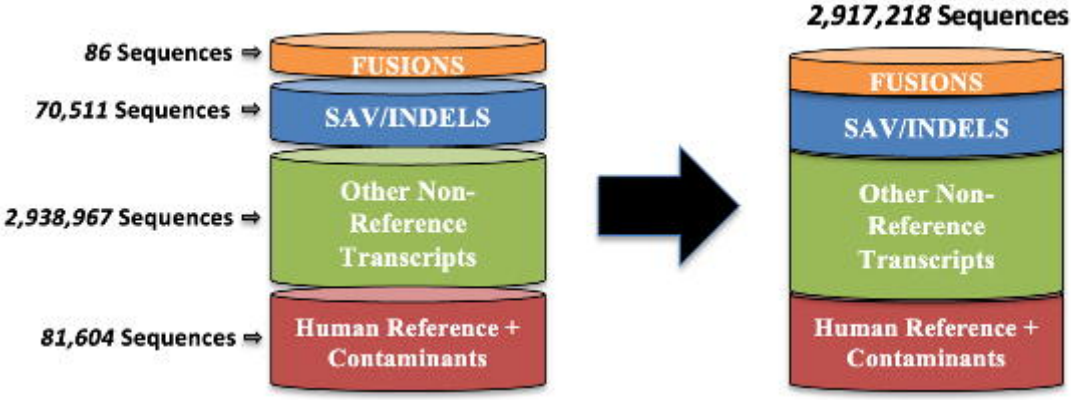
94. galaxy/LICENSE.txt at dev · galaxyproject/galaxy [Internet]. GitHub. [cited 2025 Mar 24]. Available from: https://github.com/galaxyproject/galaxy/blob/dev/LICENSE.txt

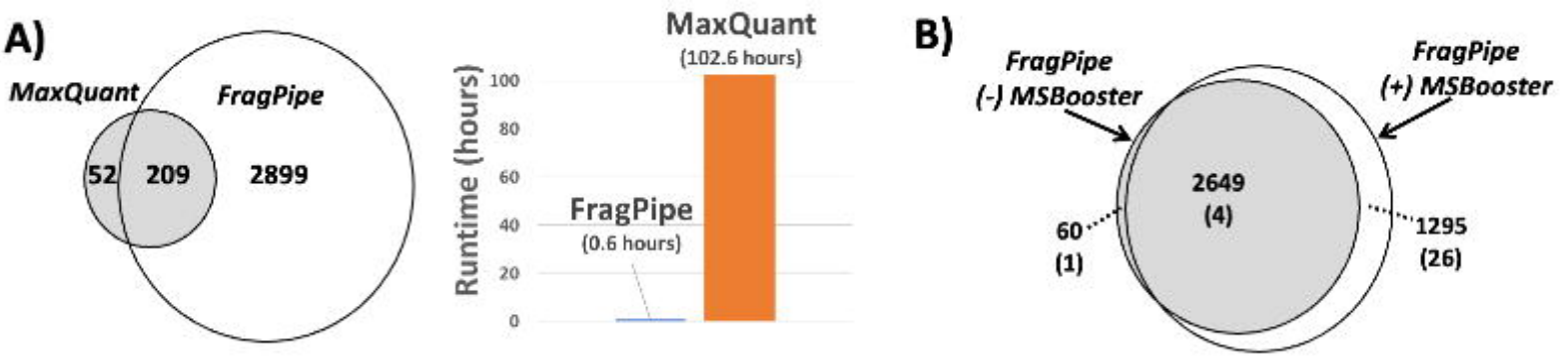
Supplemental Figure

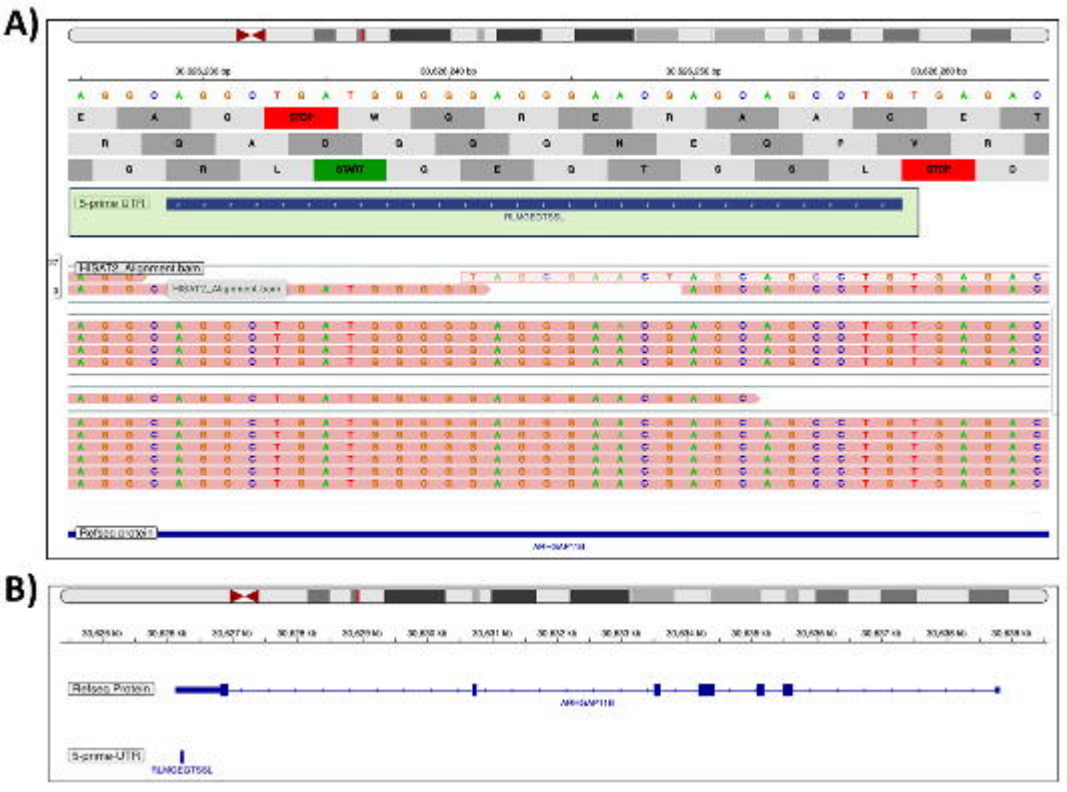
Supplemental Figure 1. Manual characterization of the genomic mapping of a complex transcript coding an identified neoantigen candidate (A) PepPointer analysis annotates the peptide (sequence ILAQEIVKY, **blue highlight**) as being located within an intronic region. **(B)** Genomic alignment indicates that the peptide is absent from the predicted genomic location. **(C)** BLAST analysis confirms the peptide's alignment to a known protein, supporting its chromosomal location, sequence alignment with an amino acid substitution in the SAE1 protein. **(D)** A re-analysis using IGV with a three-frame translation (**blue highlight**) demonstrates that eight of the nine amino acids composing the peptide are coded from a transcript mapping to the SAE1 gene structure; however the transcript also contains additional novel sequence mapping to an intron and adding amino acid, making this a novel non-reference neoantigen candidate coded by a combination of exon and intron sequences. Such cases are novel and not easily annotated by the PepPointer tool and require manual exploration.




Galaxy
Next-Generation Sequencing (e.g. RNA-Seq,)
Assemble transcripts
Arriba
FreeBayes/CustomProDB
StringTie/GFFcompare
Translate and build FASTA sequence dB
Fusions
SAVs, InDels
Intron retention, 5' UTR, intergenic regions, splicing frameshifts
+
HUMAN REFERENCE SEQUENCES
1
Merge and Filter FASTA
HUMAN REFERENCE SEQUENCES
Candidate neoantigen peptide sequences
FRIGPIPE
Candidate neoantigen peptides
LC-MS/MS
Immuno-peptidomics MS/MS
2
3
PepQuery
BLASTp
Verified neoantigen peptides
4
Map, characterize and visualize neoantigen sequences
Genotype HLA , predict binding and prioritize candidates
5


86 Sequences ⇒
FUSIONS
70,511 Sequences ⇒
SAV/INDELS
2,938,967 Sequences ⇒
Other Non-Reference Transcripts
81,604 Sequences ⇒
Human Reference + Contaminants
2,917,218 Sequences
FUSIONS
SAV/INDELS
Other Non-Reference Transcripts
Human Reference + Contaminants


A)
MaxQuant
FragPipe
52
209
2899
Runtime (hours)
100
80
60
40
20
0
MaxQuant
(102.6 hours)
FragPipe
(0.6 hours)
B)
FragPipe
(-) MSBooster
FragPipe
(+) MSBooster
2649
(4)
60
(1)
1295
(26)


A)
B)
5-prime UTR
Relsec Protein
5-prime-UTR