A modular, Galaxy-based immunopeptidogenomic (iPepGen) analysis pipeline for discovery, verification, and prioritization of candidate cancer neoantigen peptides

Subina Mehta[1], Reid Wagner[2], Katherine T. Do[1], James E. Johnson[2], Fengchao Yu[3], Tyler Jubenville[4], Kyle Richards[5], Suzanne Coleman[4], Flavia E. Popescu[4], Alexey I. Nesvizhskii[3], David A. Largaespada[4,5], Pratik D. Jagtap[1,4], Timothy J. Griffin[1,4*]

Institutional affiliations

[1]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota

[2]Minnesota Supercomputing Institute, University of Minnesota

[3] Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan

[4]Masonic Cancer Center, University of Minnesota

[5]Department of Pediatrics, University of Minnesota

*Corresponding author
University of Minnesota
6-155 Jackson Hall
321 Church St SE
Minneapolis, MN 55455
tgriffin@umn.edu

Abstract

Background

Characterizing peptide antigens, processed from tumor-specific proteoforms, and bound to the major histocompatibility complex, is critical for immuno-oncology research. Next-generation sequencing predicts candidate neoantigen peptides derived from DNA mutations and/or RNA transcripts coding proteoform sequences that differ from the reference proteome. Mass spectrometry (MS)-based immunopeptidomics identifies predicted, MHC-bound neoantigen peptides and other tumor antigens. This "immunopeptidogenomic" approach requires multi-omic software integration, challenging researchers with limited bioinformatics expertise and resources. As a solution, we developed the immunopeptidogenomic (iPepGen) pipeline in the Galaxy ecosystem. iPepGen is composed of five core workflow modules, available via publicly accessible, scalable Galaxy instances, accompanied by training resources to empower community adoption.

Findings

Using representative multi-omic data from malignant peripheral nerve sheath tumors, we demonstrate the operation of iPepGen modules with these functions: 1) Predict neoantigen candidates from sequencing 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 method 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.

Conclusions

We demonstrate the effectiveness of the iPepGen pipeline for candidate neoantigen discovery and characterization. With tools, workflows, and training resources available in the open Galaxy ecosystem, iPepGen should provide cancer researchers with a flexible and accessible informatics resource tailored to accelerating immuno-oncology studies.

Keywords (3-10 keywords)

Neoantigen, immunopeptidomics, immuno-oncology, cancer, mass spectrometry, multi-omics, Galaxy, bioinformatics, workflow

FINDINGS

BACKGROUND

Peptide neoantigens and their promise in cancer immunotherapy. Neoantigen peptides derive from proteolytic processing of unique non-reference proteoforms [1] expressed in pre-cancerous or cancerous, tumor-forming cells. These tumor-specific proteoforms arise from various mechanisms[2–4], including non-synonymous genomic sequence mutations and structural variants (e.g. gene fusions[5]), dysregulated RNA splicing and processing, novel transcription/translation from putative non-coding regions[6], derepression of transposable element expression, aberrant post-translational modification (PTM) of proteins[7], and from translation of viral encoded open reading frames for virus-mediated cancers[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].

Some proteoforms with non-reference sequences are recognized as foreign to the cell, and undergo proteolysis in the proteasome, resulting in neoantigen peptides that are 8-11 amino acids in length[7], bound to the cell-surface localized class I major histocompatibility complex (MHC I); in humans, the MHC is commonly referred to as the human leukocyte antigen (HLA) complex, where expression of HLA-A, HLA-B and HLA-C genes encode for the human form of MHC I[10]. Alternatively, extracellular proteoforms undergo proteolysis in the endosome, resulting in non-reference neoantigen peptides 10-25 amino acids in length[7] bound to the cell surface localized class II major histocompatibility complex (MHC II), encoded by HLA-D genes in humans[11]. The peptides bound to the MHC I or II complex are recognized by specialized T-cells. It's been estimated that 10% of processed, non-reference neoantigen peptide sequences could be MHC I-bound and elicit an immunogenic response[12]. Unfortunately, despite MHC I presentation of neoantigens, tumor cells can evade immune activation and promote continued proliferation and tumor growth[13]. However, when MHC-bound neoantigen peptide sequences are known, numerous studies have demonstrated effective exploitation of these peptides for personalized immunotherapy treatment, (e.g., see [14–17]), and even as powerful prophylactic tumor vaccines for cancer immunoprevention[14,18].

Identifying and predicting neoantigen candidates from genomic and transcriptomic data. Given their promise, research focused on identifying tumor-specific neoantigens has intensified in recent years[4,19]. These studies utilize genomic and/or transcriptomic next-generation sequencing (NGS) data to identify cancer-specific, non-reference coding sequences that may generate neoantigens. A collection of software tools and pipelines exist [20–23] for predicting candidate neoantigens from NGS data, covering a diverse array of genome/transcriptome alterations. Additionally, tools now exist to predict the binding of in-silico translated predicted

neoantigen peptides to MHC I or II (e.g., see [24–26]). These tools commonly analyze RNA-Seq data to determine MHC allele genotypes (HLA allele genotypes in human samples). The binding affinity of peptides to specific MHC alleles, which offers an initial indicator of their potential immunogenicity, can be predicted based on biophysical and structural properties of the genotyped MHC using customized software, complemented by growing empirical data on MHC peptide binders [27].

Beyond prediction: confirming MHC-bound neoantigens via mass spectrometry (MS)-based immunopeptidomics. Although valuable, NGS data, coupled with MHC genotyping and binding algorithms for modeling neoantigen presentation and potential immune response [28–30] are only predictive. As such, these approaches have a number of limitations [4,31]: 1) No direct confirmation that novel DNA or transcript sequences are actually translated to stable proteoforms; 2) Even if translated, these tools provide no direct confirmation of their processing and binding to the MHC; 3) These methods cannot predict PTMs that are known to underlie the binding and immunogenicity of many neoantigens or TAAs [4,32]; 4) Predictions have inherent biases [33,34] which may overlook true peptide binders, necessitating a more unbiased, direct approach to discover bound peptides associated with specific MHC genotypes within individual samples , which also is a key towards personalized therapy [35]; 5) For neoantigens predicted from NGS data, a low number of these prove to be immunogenic in follow-up experiments using immune monitoring assays[36].

These limitations have led to the development of an immunopeptidomics approach (**Figure 1**), using liquid chromatography (LC) tandem mass spectrometry (MS/MS), to directly enrich and detect MHC-bound peptides [37–39]. These approaches leverage predicted neoantigen peptide sequences from genomic or transcriptomic data analysis to build sample-specific sequence databases, which include candidate neoantigen sequences and also canonical, reference protein sequences that may be the source of TAA or TSA peptides. These are used with sequence database searching algorithms to generate peptide spectrum matches (PSMs) to MS/MS spectra, offering the ability to also identify antigen peptides carrying PTMs mediating their immunogenicity [32]. The wet-bench and instrumental methods for MS-based immunopeptidomics first demonstrated in pioneering studies by Hunt and colleagues [40], have continued to progress and proliferate, enabling the routine detection of thousands of MHC-bound peptides from both cellular and even solid tissue samples, greatly accelerating immuno-oncology studies[4,38,39,41].

Informatic requirements for neoantigen prediction and confirmation via immunopeptidogenomics. The coupling of genomic and transcriptomic data and MS-based immunopeptidomics to predict and identify neoantigens, an approach coined "immunopeptidogenomics" by Purcell and colleagues [42], presents a number of informatics requirements, including: 1) Integration of tools for neoantigen prediction from NGS data, with

MS-based peptidomic analysis tools [43], as well as tools to interpret results and prioritize discovered MHC-bound peptides for further experimental evaluation; 2) Generation of PSMs from MS/MS data using customized sequence databases containing predicted neoantigen sequences, using "no enzyme" search settings that challenge traditional proteomic software tools which are tailored towards protein digestion with trypsin to constrain the sequence search space for PSM generation [39,44,45]; 3) An option for analysis of PTMs, given their prominent role in regulating neoantigen and/or TAA/TSA peptide binding to the MHC [32,46,47]; 4) Availability of peptide-level abundance measurements via label-free quantification (LFQ) from LC-MS data [48,49], to further prioritize candidate neoantigen peptides when comparing experimental conditions.

Unfortunately, limited informatics resources exist that meet these requirements for complete immunopeptidogenomic analysis, especially those accessible by non-expert bench researchers. Most tools available for neoantigen discovery are focused on the analysis of NGS data to detect unique sequences expressed in cancerous cells or tissues arising from DNA or RNA alterations [50–53]. Algorithms following established rules integrating MHC allele genotyping and predicted peptide binding are used for further prioritization of potential neoantigens[54], and some, such as pVACtools [22] and the CEDAR resource [26,55] incorporate additional functionalities to aid in prioritizing predicted neoantigen candidates. However, none of these include integrated functionalities for the analysis of MS-based immunopeptidomic data.

Some pipelines have been described that focus on analysis of MS-based immunopeptidomics data (e.g. see [56–58]) 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 [56,57] use publicly available, established software such as MaxQuant [59] or other algorithms [58] for generating PSMs from MS/MS data; however, none of these algorithms are optimized for analysis of MHC-bound peptides using "no enzyme" settings. Others utilize commercial software for generating PSMs [42], limiting accessibility by the greater research community. Recently, a sophisticated multi-omics platform called NeoDisc has been described that is openly available for non-commercial use [60]. NeoDisc offers several tools for predicting neoantigen peptides from NGS data, as well as generating PSMs via analysis of MS-based immunopeptidomics data using sample-specific and publicly available protein sequence databases. Downstream tools for assessing the immunogenicity of identified peptides and prioritizing further experimental validation, including machine learning-based algorithms, are a main focus of this advanced platform. However, NeoDisc lacks an accessible Graphical User Interface for users with limited informatics expertise and/or scalable computational infrastructure necessary for local implementation. Moreover, the addition of new customized software within the pipeline (e.g.

new tools for neoantigen prediction from NGS data types) would require extensive software developer effort and expertise.

The Galaxy ecosystem as a solution for immunopeptidogenomic informatics. The Galaxy ecosystem brings several strengths as a solution for immunopeptidogenomics informatics. Galaxy's thriving community of thousands of active users and developers worldwide [61] offers a mature computing environment, with a web-based user interface geared towards usage by bench scientists, capable of integrating disparate command-line tools from across 'omic domains into reproducible workflows. Software tools are open and freely available for implementation within locally maintained Galaxy instances on high-performance computing (HPC) infrastructure [61], via the Galaxy Tool Shed, or through a network of freely available, scalable, cloud-based public servers [62]. The ecosystem provides a comprehensive catalog of contemporary software tools for analyzing NGS data, as well as tools for analysis of MS/MS-based peptidomic and proteomic data [63]. Furthermore, a sophisticated provenance tracking system offers complete records of analysis histories as well as the ability to build, save, and share optimized analysis workflows, promoting reproducible and transparent analyses. All of these features are complemented by the unique Galaxy Training Network (GTN) [64,65] tailored to training new, non-expert users through online and on-demand resources aimed at empowering adoption by the research community.

OPERATION

We describe a series of Galaxy-based modules that make up the openly available iPepGen pipeline (**Figure 2** and **Table 1**). We demonstrate its effectiveness using transcriptomic and MS-based immunopeptidomic data from malignant peripheral nerve sheath tumor (MPNST) cells. This is the first description of a Galaxy-based solution for immunopeptidogenomics, with a focus on: a) predicting potential neoantigen sequences from NGS data; b) confidently generating PSMs to these sequences from MS-based immunopeptidomics data; c) characterizing the DNA and/or RNA sequences coding identified candidate neoantigens; and d) initial prioritization of candidate neoantigen peptides to guide further experimental validation. The iPepGen offers several unique advantages not currently available in any other informatics pipeline:

- The modular design in Galaxy offers flexibility for users to utilize the well-tested and optimized tools using default settings, with the option for more advanced users to customize settings in these modules for their data, and extend with new software as they emerge; users also have the option to utilize only select modules of interest, or use them together in a complete pipeline
- Accessibility on existing publicly available Galaxy gateways [62,66] that offer scalable computing resources necessary for compute-intensive steps inherent to

immunopeptidogenomic analysis; optionally, tools and workflow modules are available for local installation through the Tool Shed on high-performance computing platforms

- Implementation of rigorously validated workflows into the GTN, which offers an on-ramp for new users to learn each module, access publicly available tools, and adopt them for use with their own data

Module 1: Predict non-reference neoantigen peptide sequences. This module capitalizes on the wide variety of upstream tools and workflows available and actively maintained in Galaxy for the analysis of DNA or RNA NGS data (e.g. see [67,68]) that identify non-reference sequences that are potentially translated into novel proteoforms. These workflows offer users a diverse selection of analysis approaches to characterize aberrant gene sequences and/or transcript sequences that may give rise to neoantigen peptides. For the demonstration we present here, we use two workflows for predicting non-reference gene products that may encode neoantigens, both using RNA-Seq NGS data as inputs: 1) a workflow for identifying novel gene fusions, drivers of many cancers [69] , and potential sources of neoantigen peptides [3] using the Arriba software [70]; 2) a workflow based on CustomProDB [71] and the StringTie transcript assembly and annotation tool [72] which compare assembled sequences to the reference genome to identify a wide-variety of non-reference transcripts. We have utilized an embodiment of this workflow for Galaxy-based proteogenomic analysis [73–75]. These transcripts code proteins containing single amino acid variants (SAVs), short insertion-deletions (InDels), splice variants, intron-containing translation products, and proteins from putatively non-coding untranslated regions (UTRs) or intergenic regions. All of these predicted proteins may be processed and generate non-reference peptide neoantigens. Transcripts identified by StringTie from non-coding or frame-shifted sequences are 3-frame translated, and all included in a FASTA-formatted database of novel proteoform sequences. StringTie offers an established tool [72] recognized for its accuracy and efficiency in identifying a very broad profile of potential non-reference transcripts and their translated protein sequences, tailored towards integration with immunopeptidomics MS/MS data to maximize identification of neoantigen candidates. The database of predicted novel proteoforms is also merged with a database of human reference protein sequences from the UniProt knowledgebase. Although the workflows demonstrated here offer prediction of a wide variety of potential neoantigens, more advanced users have the option of developing customized workflows drawing on software for NGS analysis found in the Galaxy Tool Shed [67,68]. Moreover, our workflows use reference genome comparisons to identify possible neoantigen sequences, however, users may also draw upon de novo assembly tools such as Trinity [76] in Galaxy as an alternative.

Module 2: Discover candidate peptide sequences and tumor-associated antigens. This module uses a Galaxy tool (called "fasta_merge_files_and_filter_unique_sequences") designed to merge the FASTA-formatted, predicted non-reference protein sequences generated in Module

1. The merging filters out redundant sequences, creating a single FASTA database of protein sequences with customized annotations classifying the source of the non-reference sequences (e.g. Arriba annotated fusions, CustomProDB annotated variants, broadly defined non-reference StringTie-assembled sequences) along with reference human sequences from UniProt.

The merged protein sequence database serves as input to the FragPipe suite of tools [49,77–82] implemented within Galaxy. FragPipe uses the MSFragger algorithm [77] for database searching and PSM generation. MSFragger, coupled with the downstream tools in FragPipe which include deep-learning-prediction-based rescoring using MSBooster [79], machine-learning-based reranking using Percolator [80], and false-discovery rate (FDR) estimation using Philosopher [82] offers an efficient means to analyze immunopeptidomics MS/MS data using "no enzyme" constraints and large customized protein sequence databases generated from NGS data (See **Performance** section below). Both reference TAA/TSA sequences as well as non-reference neoantigen candidates are discovered in this module. Furthermore, a relatively wide length of peptides can be identified (8-25 amino acids), which makes FragPipe amenable to identifying either MHC I or MHC II bound peptides depending on the enrichment protocol used.

Module 3: Verify PSMs for neoantigen candidates and process for further characterization. The potential for false positives is a main challenge in the use of sequence database searching to identify non-reference peptide sequences from MS/MS data, especially when using large databases containing millions of predicted non-reference sequences from NGS data [83]. Despite the rigorous analysis using FragPipe in Module 2, neoantigen candidates matching non-reference sequences require a cautious approach in assessing those sequences of highest quality for further study. For this, we use the PepQuery2 software [84] for a secondary analysis step of FragPipe-identified neoantigen peptides. Using a list of candidate neoantigen peptides identified by FragPipe as input, along with the raw MS/MS spectral files and reference protein sequence database, PepQuery2 rigorously evaluates putative PSMs from the Discovery module. For any given peptide of interest, the algorithm selects MS/MS spectra generated from precursors of similar mass; it then generates thousands of randomized sequences along with human reference sequences of similar length and mass to evaluate the quality of predicted fragmentation patterns to the empirically derived MS/MS spectrum. PepQuery2 also considers reference sequences containing common PTM mass shifts, evaluating those with correct precursor masses as potential matches to the MS/MS spectra in question. A hyperscore is assigned to the best sequence match to the MS/MS spectrum in question, as well as the p-value for significance. These are used to assign any given input peptide a "Yes" or "No" in terms of confidence of the PSM.

Our secondary analysis of selected peptides from Module 2 follows a tradition of utilizing multiple algorithms for PSM generation on the same data, working off the underlying

assumption that PSMs identified by more than one algorithm are of highest quality [85], especially for non-reference sequences known to have a higher potential for false positives [32].

Module 4: Visualize and classify verified neoantigen sequences. This module integrates the assembled gene and/or transcript sequences from the NGS data in Module 1 with the verified neoantigen peptide sequences. Here, the Galaxy tool PepPointer [73] create a Browser Extensible Data (BED) formatted file, enabling mapping of peptides and their coding transcripts to the genome. PepPointer uses the BED file along with the Gene Transfer File (GTF) containing gene annotations for the reference human genome to initially classify the genomic source of the non-reference neoantigen peptide sequence. This produces a tabular output with annotation for the genomic source of each peptide of interest - such as those derived from known coding DNA sequence (CDS), intron-containing transcripts, frameshifts, or an untranslated region (UTR). Importantly, PepPointer uses the BED file to create inputs to popular genome browsers including the Integrated Genome Viewer (IGV) [86] and the UCSC Genome Browser [87]. These browsers allow users to more deeply explore the nature of the sequences coding neoantigen peptides of interest.

Module 5: Prioritize neoantigens for further study. This module uses established tools implemented in Galaxy for further prioritizing neoantigens that may have the most potential as immunogenic tumor markers. The module is built upon two complementary algorithms that first genotype the HLA alleles from RNA-Seq or genomic NGS data: OptiType [88] and Seq2HLA [89]. OptiType is ideal for MHC I allele genotyping, while Seq2HLA provides broader genotyping of both MHC I and MHC II, if desired. The workflow combines and formats these genotyping outputs for input to the IEDB MHC Binding prediction suite of tools [90]. Using the list of neoantigen peptide sequences of interest as the other input, the IEDB tools use a rules-based algorithm, informed by empirical data, to predict binding affinity to specific allele genotypes. Peptide sequences larger than 8 amino acids are further processed in-silico into potential cleavage product peptides, which are also evaluated for binding affinity. The module also offers an additional PepQuery2 verification step for these predicted cleavage product peptides to determine whether there is any MS/MS evidence for these being present in the sample. Ultimately, the tabular output allows users to prioritize neoantigens that may be strong or weak binders to their corresponding HLA alleles - all potential indicators of immunogenicity.

FUNCTIONALITY

Using RNA-Seq and MS-based immunopeptidomics data from an MPNST cell line as representative user data, we describe the functionality and results output for each module of the iPepGen pipeline. **Table 1** provides links to the Galaxy workflows that include all the necessary tools for each module, and the last column contains the Galaxy history containing demonstration input data and results for our representative MPNST data.

Module 1: Predict non-reference neoantigen peptide sequences. We provide two separate workflows for predicting non-reference neoantigen peptides from RNA-Seq data, aimed at casting a wide net of possibilities for matching to MS/MS data in Module 2.

Gene fusion variant prediction. One workflow in this module uses Arriba [70] that combines with RNA STAR aligner [91] to identify gene fusion variants. RNA STAR detects spliced alignments and identifies chimeric reads inherent to fusion events. Arriba processes these candidate chimeric reads, filters and annotates them to output likely fusion sequences. The workflow also includes post-processing steps that refine the results by excluding potential false positives based on read depth and quality thresholds, and elimination of known, non-pathogenic fusions.

The input data for this workflow consists of paired-end RNA-Seq reads in fastq format. The workflow produces various outputs, which include a tabular summary of detected fusions that lists fused gene pairs and read quality metrics. Graphical representations of each fusion event are generated in pdf format for easy interpretation of the results. Our workflow also reformats these outputted results into simplified tabular and FASTA formats of predicted peptide sequences coded by these gene fusion events for use in downstream analysis modules. Item number 21 in the corresponding history in **Table 1** shows the 86 non-reference peptides predicted from these fusions in FASTA format, generated from our demonstration MPNST data.

Comprehensive non-reference neoantigen prediction. Complementing the Arriba workflow for fusions, we demonstrate a workflow for broadly identifying non-reference transcripts from RNA-Seq data. It starts with comprehensive alignment of paired-end reads to a reference genome using HISAT2 [92] operated with default parameters, followed by variant calling with FreeBayes. Predicted variants are further annotated and collated using CustomProDB [71] creating a customized protein database of protein sequences containing predicted single amino acid variants (SAVs), as well as non-reference sequences arising from small InDels.

Complementing the identification of annotated variants by CustomProDB, the workflow uses the program StringTie[72] combined with GffCompare [93] to assemble transcripts, align to the reference genome and identify transcripts that do not match the reference. These are outputted in a BED formatted file, containing nucleotide sequences and their encoded proteins, using the Galaxy tools “Translate BED Transcripts” and “bed to protein map”. Sequences that do not align with a known CDS, as well as frame shifted sequences, are translated in all three reading frames to include all possible translation products. The translated proteins are then merged with canonical, reference human protein sequences from the UniProt database [94] and common contaminant protein sequences using “FASTA Merge Files and Filter Unique Sequences”. In this merging process, annotations of the non-reference sequences are appended with a header “Generic”, which is compatible for FragPipe for PSM generation in Module 2.

Both of these workflows are implemented using parameters intended to produce a broad catalog of non-reference sequences that may encode novel proteoforms, which are further processed into novel neoantigens peptides. As such, parameters such as read depth thresholds for accepting assembled transcripts are kept rather low. The CustomProDB workflow uses a setting of one of one read per kilobase per million mapped reads (RPKM) threshold value, while StringTie uses a default RPKM value of two. This large database of possible non-reference sequences maximizes the potential to generate PSMs to neoantigen candidates, which are then stringently assessed for quality in downstream modules. As with all of the Galaxy-based tools, more experienced users have the option to customize parameters set in any of the tools comprising these workflows (e.g. increase read depth thresholds, filters for variant calling), or even implement their own workflows for predicting non-reference neoantigen peptides if desired.

The input data for this workflow consists of paired-end RNA-Seq reads in fastq format and human reference GTF file. The final output includes the merged FASTA formatted protein databases, consisting of both annotated non-reference and reference proteins. The results from the demonstration MPNST RNA-Seq dataset using the workflow combining CustomProDB and StringTie to annotate non-reference sequences resulted in a merged FASTA database containing approximately 2.9 million protein sequence entries (item 48 in the corresponding history shown in **Table 1**). This database also included entries from the human reference Uniprot database.

Module 2: Discover candidate peptide sequences and tumor-associated antigens. This workflow discovers candidate neoantigens from MS-based immunopeptidomics data by matching MS/MS data to the merged FASTA database generated in Module 1, using the FragPipe suite of tools. FragPipe uses the MSFragger algorithm for PSM generation [77], along with the option of using MSBooster [79] (see **Performance** section below), Percolator [80], and Philosopher [82] to maximize results and ensure false discovery rate (FDR) control (1% global PSM- and protein-level as a default using target-decoy methods). For identifying MHC I-bound peptides, we use a workflow within FragPipe called "Nonspecific-HLA", using no enzyme settings for matched peptide sequences and identifying peptides from 8-20 amino acids in length. When searching a large, customized database (e.g. over 2 million sequences for the demonstration data), database splitting into 200 sections is used as a default, to save on active memory requirements during the analysis. In the demonstration workflow shown in Table 1, oxidation on methionine was used as the only variable PTM, although other PTMs can be included if desired. See the **Performance** section below for more on the importance of these settings for neoantigen discovery.

The input files required for this workflow include unprocessed MS/MS data files (e.g., Thermo.raw or mzML converted), FASTA protein sequence database, and an experimental

design tabular file listing the names of the MS/MS datafiles used for input. The file names in the experimental design file must match the file names of the input MS/MS data files; if multiple files are being analyzed, these can be annotated with experimental conditions and/or any other groupings of samples based on the study design. The workflow output includes peptide identification results, inferred proteins, and tabular-formatted output for all PSMs, identified peptides, and inferred proteins. Users can select which of these outputs are produced, with the most important output being identified peptides that may be of interest as neoantigens.

The workflow utilizes the Select and Query Tabular tool [95], which uses regular expressions to extract from the tabular list of identified peptides neoantigen candidates with sequences that do not map to any known, annotated proteins in the reference proteome. It should be noted that users have the flexibility to filter peptides identified by FragPipe according to their preferences. By utilizing the Select tool, they can extract specific classes of variants that were included in their FASTA protein sequence database, such as SAVs, INDELs, and fusion proteins, by matching annotations contained in the headers for these different classes. Analyzing our demonstration data consisting of a single raw MS/MS data file from immunopeptidomics LC-MS/MS analysis from an MPNST cell line and our FASTA sequence database of over 2.9 million entries, we obtained 40 candidate neoantigen peptides matching non-reference sequences (item 24 in the corresponding history shown in **Table 1**).

Module 3: Verify PSMs for neoantigen candidates and process for further characterization. Module 2 utilizes a rigorous algorithm as part of FragPipe for generating PSMs and assessing confidence via FDR control and quality measures of annotated MS/MS identifying MHC-bound peptides. Module 3 offers a secondary analysis of neoantigen peptides of interest from Module 1, acknowledging the need for caution when evaluating PSMs to non-reference sequences [96]. For this secondary analysis, we utilize PepQuery2 [84]. PepQuery2 takes a peptide-focused approach to evaluate putative PSMs to any given MS/MS spectrum, as described above in the **Operation** section. This step offers a means to confirm PSMs for neoantigen candidates, using a second algorithm, following an established tradition in the MS-based proteomics community using multiple algorithms to increase confidence in PSM quality [85].

To utilize PepQuery2, raw MS/MS data files are first converted to standardized formats using msconvert [97]. The Mascot generic format (.mgf) conversion is preferred here as a lightweight option for input to PepQuery2. Other inputs to PepQuery2 include the tabular formatted list of candidate neoantigen sequences to be evaluated (from Module 2) as well as the reference proteome FASTA sequence database, which is selected and downloaded automatically through the Galaxy tool. Reference peptide sequences, including the addition of common PTMs, are evaluated against MS/MS derived from peptides within a narrow mass range matching the input sequence being verified, as alternatives to the putative match to non-reference

neoantigen sequences. Neoantigen sequences that are deemed the best match to the MS/MS spectrum via hyperscore assignment and statistical analysis after this rigorous re-analysis are deemed confident.

A number of settings are key to the successful use of PepQuery2 for neoantigen verification. Setting the amino acid length to a narrow range to match the input candidate sequences of interest being evaluated (e.g., 9 to 11 amino acids for MHC I-bound peptides) is essential when using no enzyme sequence settings. Using a wider range that is outside the length of the input sequences forces the program to generate all possible peptides across the reference proteome, which overwhelms the active memory. Additionally, the tabular results from PepQuery2 containing scores for each peptide evaluated are further manipulated using Query Tabular [95], extracting those neoantigen sequences deemed "confident". These sequences are converted to FASTA format using Tabular-to-FASTA to facilitate further analysis.

As a final evaluation step, the BLAST-P sequence alignment tool is used on these verified peptide sequences to further confirm that verified sequences do not match any known sequences in the human reference proteome. In this step, only peptide sequences not found in the National Library of Medicine-National Center for Biotechnology Information (NLM-NCBI) reference database are retained for further characterization, filtering out matches to canonical reference sequences, or even commonly identified SAVs or fusions. The resulting novel, non-reference peptides are those of highest priority for further characterization as potential neoantigens. A final Query Tabular step extracts verified, non-reference peptide sequences after BLAST-P analysis and creates a tabular file listing these sequences and their annotations. In the demonstration MPNST immunopeptidomics dataset, we verified five neoantigen peptide candidates from Module 3 using this stringent approach when considering peptides 9-11 amino acids in length using PepQuery2 analysis and BLAST-P confirmation (item 12 in the corresponding history shown in **Table 1**).

Module 4: Visualize and classify verified neoantigen sequences. With verified neoantigen peptides in hand, module four offers functionalities aimed at interpretation of these peptides. Tools in this module map and annotate peptide sequences with their coding genomic coordinates and strand-specific orientation. The workflow uses Query Tabular [95] and its capabilities to build and query SQL databases from tabular inputs and use regular expressions to extract relevant information on genomic coordinates and annotate and reformat into outputs compatible with downstream tools. One of these outputs is a BED file used by the PepPointer tool [73] to classify the nature of the non-reference coding sequences as being from a CDS, some combination of exonic, intronic, and/or intergenic sequences, or an unexpected gene product from a 5'-UTR or intergenic genomic locations. PepPointer also uses the genomic coordinates in the BED file to enable deeper assessment of neoantigen peptide sequences mapped against their coding genomic locations, and if desired, their corresponding transcript

reads from RNA-Seq data, providing inputs for IGV [86] and the UCSC Genome Browser [87]. Finally, a FASTA formatted file containing genomic annotations for each peptide is generated for input to Module 5.

The input for this workflow is a BED file containing chromosomal coordinates for the entire reference human genome, a peptide report with all peptides discovered in Module 2, a tabular file containing verified neoantigens from Module 3, and the human genome GTF annotation file. For the MPNST demonstration dataset, the tabular file contained five verified neoantigens from Module 3. These five peptides were annotated and classified, summarized as a tabular output (item 15 in the corresponding history shown in **Table 1**). **Figure 3** shows a screenshot of one of these peptides viewed in the IGV visualization tool, mapped to its genomic coordinates along with the associated transcripts assembled from the RNA-Seq data and contained in the BAM file from Module 1. PepPointer is meant to offer a tabular formatted, 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. Utilization of PepPointer and these functionalities have been described in detail [73]. We do note that some 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 output in Module 4 for the MPNST demonstration data.

Module 5: Prioritize neoantigens for further study. The final module offers a means for users to prioritize verified neoantigen sequences for further experimental validation, such as testing the immunogenicity of prioritized peptides in model tumor systems in the laboratory. The software within this module utilizes established algorithms to predict the binding of antigenic peptide sequences to specific MHC protein sequences (HLA in humans) translated from allele genotypes. This module provides a means to genotype HLA alleles and measure binding properties to these for verified neoantigen peptides from Module 3. This combination of HLA allele and predicted peptide binding affinity offers criteria to infer immunogenicity and prioritize neoantigens for further study.

HLA allele genotyping. The first workflow in this module involves genotyping the alleles coding for the HLA complex from human NGS data. Inputs are sequencing data (e.g. paired end RNA-Seq) in the fastq format. Alleles are genotyped using two established software (OptiType [88] and Seq2HLA [89]) to ensure comprehensive genotyping from NGS data; results from each tool are merged and organized, removing redundant results using the AWK regular expression tool for text reformatting in Galaxy. Query Tabular is used in the final step to generate tabular outputs of allele genotypes combined from both programs for analysis in the second workflow

for predicting peptide binding affinities. From processing the demonstration data from MPNST cells**,** we obtained 8 genotyped HLA alleles (item 13 in the corresponding history shown in **Table 1**).

Peptide HLA Binding Prediction. The second workflow uses established rules to predict the binding affinity of verified neoantigen peptide sequences from Module 3 to the HLA genotypes from the first workflow of this module. Using the verified neoantigen peptides (tabular format) from Module 3 as input, the IEDB MHC binding affinity prediction tool [90] is used to identify peptides with strong (percentile rank <=0.5) and weak (percentile rank between 0.5 and 2) binding affinity to specific HLA alleles. The IEDB prediction tool offers different methods for predicting affinity peptide binding to the HLA. For our workflows, we used the "netMHCpan_el" method (NetMHCpan predicts for HLA class I genotypes), which employs "eluted ligand" (EL) considerations that use a model trained on empirical immunopeptidomics data for predicting the likelihood a peptide sequence will be processed and HLA-bound and potentially recognized by immune system cells.

The IEDB MHC binding algorithm also processes the input peptides that are greater than 8 amino acids in length into smaller sequences and predicts their binding affinity to the HLA alleles. Therefore, even a small number of peptides from Module 3 used as input results in potentially dozens of smaller sequences analyzed for HLA allele binding affinity. This workflow also includes another optional PepQuery2 step to verify not only the initial peptides used as input but also in-silico processed sequences derived from these peptides, which may have gone undetected in the original discovery module. Peptides deemed confident by PepQuery2 are separated into tabular outputs for strong and weak binders and annotated with their corresponding HLA alleles, using Query Tabular for manipulating and reformatting results. It is possible that a user could skip the PepQuery2 step here if only desiring to focus on results for the input peptide sequences already verified via Module 3.

The input for this workflow includes peptide sequences from the Module 3 workflow in FASTA format, tabular results of HLA allele genotypes, and the raw MS/MS files used for verification in the PepQuery2 step. PepQuery2 downloads the human reference proteome for its verification analysis. The output consists of a refined list of verified neoantigens, annotated with their respective HLA alleles and binding affinities. For the MPNST demonstration data, the five neoantigen peptides had strong binding to specific HLA alleles (item 21 in the corresponding history in **Table 1**), with the three of these peptides also having weak predicted binding to other specific HLA alleles (item 19 in the history).

METHODS

Generation of NGS data. Data for predicting neoantigen candidates is generated via standard NGS methods. For demonstration data utilized in this work, RNA-Seq data was generated from

a human MPNST cell line STS-26T, which was treated with gentamicin as a means to increase ribosome readthrough of premature stop codons and increase translation of non-reference protein sequences in STS-26T cells [98]. RNA-Seq data was generated using paired-end sequencing, at a depth of 80 million reads and an average length of 100 base pairs. Other NGS data could also be compatible with Galaxy-based tools for NGS assembly and variant annotation, including whole genome sequencing and/or exome sequencing [67,68]. Regardless of the NGS method used for generating data, the analysis produces coding sequences that do not match known reference sequences, which are then translated in-silico to produce predicted neoantigen peptide sequences for inclusion in the FASTA formatted protein sequence database used in the Discovery module.

Generation of MS-based immunopeptidomics data. Numerous methods have been described for the reliable generation of MS-based immunopeptidomics data [99–102]. These all rely on the use of antibodies that recognize the MHC I or MHC II, enabling immunoprecipitation of these protein complexes along with bound peptides. In this work, the MHC I was enriched from STS-26T cells treated with gentamicin. Peptides bound to MHC I were eluted and analyzed by liquid chromatography (LC) MS/MS. The raw MS/MS files are input to Module 2 (for use by FragPipe) and also Module 3 (PepQuery2). Our current pipeline utilizes data-dependent acquisition (DDA)-based MS/MS data using high-resolution instrumentation produced by the popular Orbitrap MS instruments from ThermoFisher Scientific. The msconvert tool [97] within Galaxy transforms any instrument raw file into community standard mzML format, which our pipeline can utilize. FragPipe also offers tools for data-independent acquisition (DIA) analysis [103–105], which could provide input to identify candidate neoantigens as an alternative approach to more traditional DDA.

Software tools implemented in iPepGen. The links to Galaxy workflows for each iPepGen module (**Table 1**) provide a listing of all software used in the analysis. All software is freely available for non-commercial research. All tools have been reviewed and tested by the members of the Galaxy Intergalactic Utilities Commission [106] prior to making these available for installation within any Galaxy instance within the Galaxy Tool Shed. The Tool Shed listing also contains links to GitHub repositories for each software.

Table 1. Modules comprising the Galaxy-based iPepGen analysis pipeline

Module	Description	Accessible Galaxy workflow	Accessible Galaxy history
1.Predict non-reference neoantigen peptide sequences	Predicts neoantigens peptide sequences encoded by non-reference genes and/or	*Gene fusion neoantigens* https://usegalaxy.eu/u/galaxyp/w/gigasciencefusionsdemonstrationsts26t	*Gene fusion neoantigens* https://usegalaxy.eu/u/galaxyp/h/neoantigenfu

	transcripts and generates customized protein sequence databases	-gentworkflow	siontranscriptsdbgeneration
		General non-reference sequence neoantigens https://usegalaxy.eu/u/galaxyp/w/gigascienceindelssavnon-normaldemonstrationsts26t-gentworkflow-1	*General non-reference sequence neoantigens* https://usegalaxy.eu/u/galaxyp/h/neoantigennon-normaltranscriptsdbgeneration
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://usegalaxy.eu/u/galaxyp/w/gigasciencedatabasemergefragpipests26tdemonstration	https://usegalaxy.eu/u/galaxyp/h/neoantigen-db-merge-fragpipe-discovery
3. Verify PSMs for neoantigen candidates	Verifies neoantigen peptide sequences identified in Module 2 using a secondary spectral matching algorithm PepQuery2	https://usegalaxy.eu/u/galaxyp/w/gigasciencepepquery2demonstrationsts26tneoantigencandidatesworkflow	https://usegalaxy.eu/u/galaxyp/h/neoantigen-pepquery2-verification
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://usegalaxy.eu/u/galaxyp/w/gigasciencepeptideannotationdemonstrationsts26tneoantigencandidatesworkflow	https://usegalaxy.eu/u/galaxyp/h/neoantigenvariant-peptide-annotation
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://usegalaxy.eu/u/galaxyp/w/gigascience-rnaseq-optitype-seq2hla-to-iedb-alleles	*HLA allele genotyping* https://usegalaxy.eu/u/galaxyp/h/neoantigenpredictinghlabinding
		Peptide-HLA binding prediction https://usegalaxy.eu/u/galaxyp/w/gigascience-iedb-pepquery-neoantigen	*Peptide-HLA binding prediction* https://usegalaxy.eu/u/galaxyp/h/neoantigenhlabindingnovelpeptides

IMPLEMENTATION

As shown in Table 1, all workflows have been tested using example multi-omic data from MPNST cell lines and analyzed using publicly available workflows within the Galaxy Europe public instance. Each tool within each workflow module may also be implemented via download and local Galaxy install from the Galaxy Tool Shed. Complete workflows can be saved as .ga files and implemented in any Galaxy instance as well, provided individual tools have been installed and tested for operation.

ACCESSIBILITY

As shown in Table 1, all workflows as well as completed analysis histories, are accessible on the European Galaxy public instance. The histories contain all demonstration input data files, as well as the intermediate and final results of each workflow module. All software tools used within each module are accessible within the Galaxy Tool Shed [107] for local install or use on one of the global network of public Galaxy instances [62].

DOCUMENTATION AND TRAINING

For each of the modules comprising the iPepGen pipeline, we have developed training tutorials available on the GTN. These include documentation describing the inputs, software tools, and outputs for each module, and step-by-step instructions guiding new users through the workflows on representative input data. Once trained via these GTN tutorials, users are able to import the workflows into their user account on Galaxy Europe and use them on their own data. **Table 2** provides an overview of these GTN tutorials along with the links where they can be accessed.

Table 2. GTN training tutorial links for each iPepGen module

Module	GTN link
1. Predict non-reference neoantigen peptide sequences and generate customized protein sequence databases	https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/neoantigen-1-fusion-database-generation/tutorial.html https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/neoantigen-2-non-reference-database-generation/tutorial.html
2. Discover neoantigen candidate peptide sequences and tumor-associated antigens	https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/neoantigen-3-fragpipe-discovery/tutorial.html

3. Verify PSMs for neoantigen candidates	https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/neoantigen-4-peptide-verification/tutorial.html
4. Visualize and classify verified neoantigen sequences	https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/neoantigen-5-variant-annotation/tutorial.html
5. Prioritize neoantigens for further study	https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/neoantigen-6-predicting-hla-binding/tutorial.html https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/neoantigen-7-hla-binding-novel-peptides/tutorial.html

INPUTS

For each of the iPepGen modules, we have made accessible links to Zenodo repositories containing the appropriate demonstration input data generated from our MPNST cells, as shown in **Table 2**. These data can be uploaded into any public or local Galaxy instance, creating a new history. These inputs can then be operated on using the corresponding workflows shown in Column 3 of **Table 1**. Column 4 of **Table 1** shows the links to these histories.

PERFORMANCE

Runtime and compute resources considerations. Most of the software tools comprising the modules of iPepGen do not require special consideration for allocating computing resources to ensure acceptable performance. There are, however, several tools that do deserve some discussion related to their performance and computing requirements to ensure their reliable operation.

FreeBayes. This is a community-standard tool for calling and annotating sequence variants from NGS data and is used in Module 1. Because FreeBayes must compare the assembled DNA or RNA sequences against the entire human genome while scanning for multiple different variant classes, it can take over 24 hours to complete. For example, using the European Galaxy instance with 10 core CPUs for processing and 17 GB of active memory, FreeBayes can take over 30 hours to process a single RNA-Seq file. Users should plan their analyses accordingly to accommodate this runtime.

FragPipe. A key component of Module 2 (Discovery of MHC-bound peptides) is the use of FragPipe, specifically the MSFragger algorithm [77] to generate PSMs to the protein sequence database. Traditional MS-based proteomics analysis uses trypsin to digest intact proteins, following well-established rules for amino acids recognized by the protease, and constraining

the search space for possible peptides matching to MS/MS spectra. For immunopeptidomics, peptides bound to the MHC have been processed by a variety of proteases. Consequently, PSMs must be generated using "nonspecific" or "no enzyme" settings, which greatly increases the potential sequences that could match any MS/MS spectrum, increasing memory demands on sequence database searching programs; this is exacerbated by the potentially large, customized sequence database that is used in immunopeptidomics containing millions of protein sequences that must be considered for PSM generation.

MSFragger is highly efficient at generating PSMs using no enzyme settings. Two settings are key to its performance. First, the "split database" setting allows for the matching of MS/MS from very large sequence databases, by splitting the database into smaller sections, and matching spectra to each of these separately, saving on memory requirements. For our work using a sequence database of over two million sequences, we split the database into 200 sections. We also took advantage of the scalability of Galaxy when installed on cloud-based high-performance computing resources. The installation of FragPipe on the European Galaxy instance provides 250 GB of active memory to the tool, which we found accommodates the algorithm's requirement for successful operation on immunopeptidomics data. These settings allow for MSFragger to generate PSMs across a relatively wide amino acid length (8-20 amino acids), making it suitable for immunopeptidomics methods enriching either MHC I or II-bound peptides.

FragPipe also offers additional functionalities of value for MS-based immunopeptidomics data analysis. One is the ability to identify PTMs on peptides. These are known to be potential regulators of peptides bound to the MHC [32]. Common variable modifications on MHC-bound peptides include acetylation on protein N-terminal, pyro-glu from Q, pyro-glu from E, and cysteinylation on C. The addition of variable PTM mass shifts to peptides for PSM generation further expands the space for potential sequence matches to MS/MS, increasing memory requirements for analysis. For no enzyme settings and using the reference human proteome sequence database (e.g. Uniprot), it is recommended that only 2-3 PTM possibilities are included in any given search to maintain feasible active memory requirements. In addition, the MSBooster tool [79] offers the ability to predict the fragment peak intensities and retention time for any given peptide sequence, which can be used to rescore the PSMs with and boost sensitivity. We show the effectiveness of MSBooster below in our evaluation of the depth of results enabled by the use of FragPipe below. Lastly, the IonQuant tool [49] within FragPipe is designed to quantify peptide-level data via label-free quantification. It employs established methods for aligning peptides across separate LC-MS/MS analyses, determining abundance by area-under-the-curve measurements for extracted LC profiles of detected peptides, and normalization for accurate quantification. This tool is available if users desire quantitative analysis of immunopeptidomics data.

PepQuery2. PepQuery2 operates differently than normal sequence database searching algorithms such as FragPipe, as it focuses on evaluating quality of matching to MS/MS spectra for known sequences of interest input by the user (e.g. neoantigen candidates identified in Module 2). The reference human proteome sequences from UniProt are used for this evaluation, while also considering common PTMs as other factors that may better explain matches to MS/MS compared to the putative candidate neoantigen sequences. In our example dataset, we enriched MHC class I peptides from MPNST cells, identifying neoantigen candidates of interest that are 8-10 amino acids in length. These served as input sequences for analysis by PepQuery2, which was set to consider all possible sequences from the reference proteome within this relatively narrow amino acid length, using non-specific enzyme digestion. We found that if a larger range of amino acid lengths were used (e.g. 8-20 amino acids), the generation of all of these potential sequences using no enzyme constraints overwhelmed the memory use of the program. Such a large mass range is unnecessary for PepQuery2 analysis, as it only considers peptides of similar length and mass to the sequence input for consideration. Using a narrow mass range, a list of 10 neoantigen peptide sequences could be evaluated in under 20 minutes using PepQuery2 using no enzyme settings. Importantly, even with this reduced range of peptide lengths, using no enzyme settings generates thousands of sequences from the reference proteome as possible matches to the MS/MS spectrum in question, allowing for statistical modeling and assignment of a hyperscore for the peptide sequences used as input.

HLA allele genotyping. Our pipeline incorporates HLA allele genotyping using two complementary tools, OptiType and Seq2HLA, to ensure robust and comprehensive analysis of raw NGS data. Using RNA-Seq data as input, OptiType, optimized for HLA Class I alleles (HLA-A, HLA-B, HLA-C), typically takes about 32 hours to complete using computing resources available on the European Galaxy server. Seq2HLA provides broader coverage, genotyping both Class I and Class II alleles (HLA-DRB1, HLA-DQA1, HLA-DQB1), and completes analyses in approximately 24 hours. The combined runtime for both tools could be reduced by allocating additional processors and memory. OptiType is ideal for high-accuracy Class I genotyping, while Seq2HLA's broader scope makes it suitable for studies requiring both Class I and II allele coverage. Together, these tools provide a reliable framework for HLA genotyping, balancing accuracy and computational efficiency.

Depth of results: performance of FragPipe for PSM generation from immunopeptidomics MS/MS data. The use of MSFragger for the analysis of immunopeptidomics MS/MS data in Module 2 is key to the iPepGen pipeline. It uses an ultrafast algorithm for PSM generation [77] offering efficient performance, even using no enzyme settings and matching MS/MS to a very large sequence database, which has the potential to decrease sensitivity and increase false positives [108].

As an evaluation of performance, we analyzed our representative data using FragPipe in Galaxy and compared it to results from MaxQuant, a prominently used program PSM generation [59]. In both cases, a single high-resolution raw file of immunopeptidomics MS/MS data was matched to a large sequence database of over 2.9 million protein sequences from MPNST cells generated in Module 1. **Figure 4A** shows a comparison of results for peptides identified using FragPipe and MaxQuant. Although both FragPipe and MaxQuant are available as Galaxy tools, for comparison purposes these analyses were run using the standalone GUI using the same computing resources (8 cores and 64 GB of active memory). We assessed the number of peptides identified by both software at an estimated global FDR of 1%, as well as the runtime required to generate results. FragPipe outperforms MaxQuant in both regards.

We also evaluated the use of MSBooster in FragPipe [79] to increase PSMs from immunopeptidomics data. MSBooster uses a machine-learning-based algorithm from the DIA-NN tool [109] to predict fragment peak intensities and retention times from a list of peptides generated by the initial MSFragger database search. It then calculates additional scores based on reanalysis using predicted MS/MS spectra compared to the observed spectra to boost the sensitivity of peptide identification. **Figure 4B** shows an example of the increase in identified peptide sequences using standard PSM processing via the Percolator tool [80], versus MSBooster combined with Percolator processing. For the example dataset and large sequence database of over two million protein sequence, the use of MSBooster increased the overall identification of MHC I bound peptides by 48%; importantly, it also increased the identification of non-reference neoantigen candidate peptide sequences by over 5-fold. To utilize this functionality, a user only needs to select "Run MSBooster and Percolator" under the "Validation" settings in the FragPipe tool as part of Module 2 workflow of the iPepGen pipeline. Adding this analysis did not increase the runtime or active memory requirements.

FUTURE CONSIDERATIONS

The modular iPepGen pipeline, implemented in the Galaxy ecosystem, offers a unique and accessible informatics resource for researchers generating immunopeptidogenomic data. In particular, the accessibility to training material and the optimized workflows should prove valuable for those lacking the infrastructure and/or expertise necessary to implement such tools independently. Although the pipeline offers the tools necessary for complete immunopeptidogenomic analysis, future extensions could enhance its value. For example, Galaxy offers flexibility to integrate emerging approaches for predicting candidate neoantigen sequences from newer "third generation sequencing" technologies [110] which may provide increased depth in identifying neoantigen candidates (Module 1).

Additionally, although Module 5 offers a means to prioritize candidates for further study, sophisticated tools for predicting the most promising immunogenic peptide sequences from multi-omics data continue to emerge. Many of these combine increasing empirical knowledge linking HLA allele genotypes, peptide binding, and other criteria, to model potential therapeutic efficacy and prioritize peptides for further evaluation [22,30]. Promising tools employing machine-learning algorithms trained on increasing volume of multi-omic data and empirical validations also continue to emerge, aimed at more accurate prediction of the most promising peptide candidates for immunotherapy [111,112]. Such tools for improved prioritization of neoantigen candidates could be directly integrated into iPepGen via direct deployment as Galaxy tools. Alternatively, web-based knowledgebases could be accessed programmatically through Galaxy's API. Our Galaxy-based iPepGen pipeline has been developed with an eye towards interoperability with new tools that will continue to emerge in the expanding area of immuno-oncology informatics.

AVAILABILITY OF SUPPORTING SOURCE CODE AND REQUIREMENTS

All of the software tools composing the iPepGen pipeline and its workflow modules are available through the Galaxy Tool Shed [107]. 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 available freely for academic research, non-commercial, or educational purposes. Other software contained in FragPipe is developed under permissive licensing [113]. The Galaxy bioinformatics platform is freely available and open-source and released under the MIT license [114].

DATA AVAILABILITY

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. Table 3 also provides access to all the input files for the workflow modules that make up the iPepGen pipeline.

DECLARATIONS

List of abbreviations

Browser Extensible Data (BED) formatted file

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 Deletions (INDELs)

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)

untranslated region (UTR)

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

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 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 manuscript.

Funding

The work reporter 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 demonstration of the software described here was generated using an Orbitrap Eclipse instrumentation platform purchased through High-end Instrumentation Grant S10OD028717 from the NIH. The Proteogenomics Shared Resource in the University of Minnesota Masonic Cancer Center (NCI grant P30CA077598) provided additional support for S.M., P.D.J. and T.J.G.

Authors' contributions

S.M.: Conceptualization, Data curation, Formal Analysis, Software, Visualization, Writing - original draft, Writing - review and editing

R.W.: Data curation, Resources, Software, Writing - review and editing

K.T.D.: Data curation, Formal Analysis, Software, Writing - review and editing

J.E.J.: Conceptualization, Data curation, Software

F.Y.: Data curation, Resources, Software, Writing - review and editing

T.J.: Data curation, Resources, Writing - review and editing

K.R.: Resources, Software, Writing - review and editing

S.C.: Resources, Writing - review and editing

F.P.: Data curation, Resources, Writing - review and editing

A.I.N.: Software, Supervision, Funding acquisition, Writing - review and editing

D.A.L.: Conceptualization, Resources, Supervision, Funding acquisition, Writing - review and editing

P.D.J.: Conceptualization, Software, Supervision, Funding acquisition, Writing - review and editing

T.J.G.: Conceptualization, Software, Supervision, Funding acquisition, Writing - original draft, Writing - review and editing

Acknowledgments

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 for advice on the use of PepQuery2 within our pipeline.

REFERENCES

1. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. *Nat Methods*. 2013; doi: 10.1038/nmeth.2369.

2. Capietto A-H, Hoshyar R, Delamarre L. Sources of Cancer Neoantigens beyond Single-Nucleotide Variants. *Int J Mol Sci*. 2022; doi: 10.3390/ijms231710131.

3. Xie N, Shen G, Gao W, Huang Z, Huang C, Fu L. Neoantigens: promising targets for cancer therapy. *Signal Transduct Target Ther*. 2023; doi: 10.1038/s41392-022-01270-x.

4. Ahn R, Cui Y, White FM. Antigen discovery for the development of cancer immunotherapy. *Semin Immunol*. 2023; doi: 10.1016/j.smim.2023.101733.

5. Wang Y, Shi T, Song X, Liu B, Wei J. Gene fusion neoantigens: Emerging targets for cancer immunotherapy. *Cancer Lett*. 2021; doi: 10.1016/j.canlet.2021.02.023.

6. 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; doi: 10.1126/scitranslmed.aau5516.

7. Yi X, Liao Y, Wen B, Li K, Dou Y, Savage SR, et al. caAtlas: An immunopeptidome atlas of human cancer. *iScience*. 2021; doi: 10.1016/j.isci.2021.103107.

8. Wang E, Aifantis I. RNA Splicing and Cancer. *Trends Cancer*. 2020; doi: 10.1016/j.trecan.2020.04.011.

9. Minati R, Perreault C, Thibault P. A Roadmap Toward the Definition of Actionable Tumor-Specific Antigens. *Front Immunol*. 2020; doi: 10.3389/fimmu.2020.583287.

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; doi: 10.1084/jem.20082136.

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; doi: 10.1016/0092-8674(84)90068-0.

12. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. *Nature*. 2017; doi: 10.1038/nature21349.

13. Kim SK, Cho SW. The Evasion Mechanisms of Cancer Immunity and Drug Intervention in the Tumor Microenvironment. *Front Pharmacol*. 2022; doi: 10.3389/fphar.2022.868695.

14. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nat Rev Clin Oncol*. 2021; doi: 10.1038/s41571-020-00460-2.

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*. doi: 10.1172/JCI99538.

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; doi: 10.1038/s41586-024-08507-5.

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; doi: 10.1186/s13073-024-01388-3.

18. Redwood AJ, Dick IM, Creaney J, Robinson BWS. What's next in cancer immunotherapy? - The promise and challenges of neoantigen vaccination. *Oncoimmunology*. doi: 10.1080/2162402X.2022.2038403.

19. Andersen MH. Tumor microenvironment antigens. *Semin Immunopathol*. 2023; doi: 10.1007/s00281-022-00966-0.

20. Bais P, Namburi S, Gatti DM, Zhang X, Chuang JH. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics*. 2017; doi: 10.1093/bioinformatics/btx375.

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; doi: 10.1093/bioadv/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; doi: 10.1158/2326-6066.CIR-19-0401.

23. Vensko SP, Olsen K, Bortone D, Smith CC, Chai S, Beckabir W, et al. LENS: Landscape of Effective Neoantigens Software. *Bioinformatics*. 2023; doi: 10.1093/bioinformatics/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; doi: 10.1093/nar/gkz452.

25. Kalemati M, Darvishi S, Koohi S. CapsNet-MHC predicts peptide-MHC class I binding based on capsule neural networks. *Commun Biol*. 2023; doi: 10.1038/s42003-023-04867-2.

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; doi: 10.1093/nar/gkac902.

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; doi: 10.1371/journal.pcbi.1003266.

28. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med*. 2016; doi: 10.1186/s13073-016-0264-5.

29. James CA, Ronning P, Cullinan D, Cotto KC, Barnell EK, Campbell KM, et al. In silico epitope prediction analyses highlight the potential for distracting antigen immunodominance with allogeneic cancer vaccines. *Cancer Res Commun*. 2021; doi: 10.1158/2767-9764.crc-21-0029.

30. 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; doi: 10.1186/s13073-024-01384-7.

31. Becker JP, Riemer AB. The Importance of Being Presented: Target Validation by Immunopeptidomics for Epitope-Specific Immunotherapies. *Front Immunol*. 2022; doi: 10.3389/fimmu.2022.883989.

32. 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; doi: 10.1038/s41587-022-01464-2.

33. Zeng H, Gifford DK. Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design. *Cell Syst*. 2019; doi: 10.1016/j.cels.2019.05.004.

34. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLoS Comput Biol*. 2018; doi: 10.1371/journal.pcbi.1006457.

35. 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*. doi: 10.7554/eLife.71156.

36. Levitsky DO, Benevolensky DS. Effects of changing Ca2+-to-H+ ratio on Ca2+ uptake by cardiac sarcoplasmic reticulum. *Am J Physiol*. 1986; doi: 10.1152/ajpheart.1986.250.3.H360.

37. Gfeller D, Liu Y, Racle J. Contemplating immunopeptidomes to better predict them. *Semin Immunol*. 2023; doi: 10.1016/j.smim.2022.101708.

38. Shapiro IE, Bassani-Sternberg M. The impact of immunopeptidomics: From basic research to clinical implementation. *Semin Immunol*. 2023; doi: 10.1016/j.smim.2023.101727.

39. 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; doi: 10.3390/cancers14051243.

40. 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*. 179:2669–712007;

41. 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; doi: 10.1038/s41467-021-26936-y.

42. Scull KE, Pandey K, Ramarathinam SH, Purcell AW. Immunopeptidogenomics: Harnessing RNA-Seq to Illuminate the Dark Immunopeptidome. *Mol Cell Proteomics*. 2021; doi: 10.1016/j.mcpro.2021.100143.

43. Menschaert G, David F. Special Issue of Mass Spectrometry Reviews on “Computers in Mass Spectrometry”: “Proteogenomics from a bioinformatics angle: a growing field.” *Mass Spectrom Rev*. 2017; doi: 10.1002/mas.21483.

44. Alvarez B, Barra C, Nielsen M, Andreatta M. Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes. *Proteomics*. 2018; doi: 10.1002/pmic.201700252.

45. Kote S, Pirog A, Bedran G, Alfaro J, Dapic I. Mass Spectrometry-Based Identification of MHC-Associated Peptides. *Cancers (Basel)*. 2020; doi: 10.3390/cancers12030535.

46. Sandalova T, Sala BM, Achour A. Structural aspects of chemical modifications in the MHC-restricted immunopeptidome; Implications for immune recognition. *Front Chem*. 2022; doi: 10.3389/fchem.2022.861609.

47. . Capturing the diversity of protein modifications on presented tumor antigens. *Nat Biotechnol*. 2023; doi: 10.1038/s41587-022-01465-1.

48. 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; doi: 10.3390/proteomes8030015.

49. Yu F, Haynes SE, Nesvizhskii AI. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol Cell Proteomics*. 2021; doi: 10.1016/j.mcpro.2021.100077.

50. Diao K, Chen J, Wu T, Wang X, Wang G, Sun X, et al. Seq2Neo: A Comprehensive Pipeline for Cancer Neoantigen Immunogenicity Prediction. *Int J Mol Sci*. 2022; doi: 10.3390/ijms231911624.

51. Rieder D, Fotakis G, Ausserhofer M, René G, Paster W, Trajanoski Z, et al. nextNEOpi: a comprehensive pipeline for computational neoantigen prediction. *Bioinformatics*. 2021; doi: 10.1093/bioinformatics/btab759.

52. Wang T-Y, Wang L, Alam SK, Hoeppner LH, Yang R. ScanNeo: identifying indel-derived neoantigens using RNA-Seq data. *Bioinformatics*. 2019; doi: 10.1093/bioinformatics/btz193.

53. Zhou C, Wei Z, Zhang Z, Zhang B, Zhu C, Chen K, et al. pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. *Genome Med*. 2019; doi: 10.1186/s13073-019-0679-x.

54. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, et al. Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics*. 2015; doi: 10.1093/bioinformatics/btv123.

55. Blazeska N, Kosaloglu-Yalcin Z, Vita R, Peters B, Sette A. IEDB and CEDAR: Two Sibling Databases to Serve the Global Scientific Community. *Methods Mol Biol*. 2023; doi: 10.1007/978-1-0716-3239-0_9.

56. 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; doi: 10.1186/s12920-020-0683-4.

57. 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; doi: 10.3390/genes13050783.

58. Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat Commun*. 2020; doi: 10.1038/s41467-020-15456-w.

59. 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; doi: 10.1038/nbt.1511.

60. 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; doi: 10.1038/s41587-024-02420-y.

61. . The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res*. 2022; doi: 10.1093/nar/gkac247.

62. : Galaxy Community Hub. https://galaxyproject.org/use/ Accessed 2025 Mar 25.

63. 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; doi: 10.1080/14789450.2023.2265062.

64. 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; doi: 10.1371/journal.pcbi.1010752.

65. Batut B, Hiltemann S, Bagnacani A, Baker D, Bhardwaj V, Blank C, et al. Community-Driven Data Analysis Training for Biology. *Cell Syst*. 2018; doi: 10.1016/j.cels.2018.05.012.

66. Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res*. 2024; doi: 10.1093/nar/gkae410.

67. Backofen R, Engelhardt J, Erxleben A, Fallmann J, Grüning B, Ohler U, et al. RNA-bioinformatics: Tools, services and databases for the analysis of RNA-based regulation. *J Biotechnol*. 2017; doi: 10.1016/j.jbiotec.2017.05.019.

68. Larivière D, Abueg L, Brajuka N, Gallardo-Alba C, Grüning B, Ko BJ, et al. Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nat Biotechnol*. 2024; doi: 10.1038/s41587-023-02100-3.

69. Gao Q, Liang W-W, Foltz SM, Mutharasu G, Jayasinghe RG, Cao S, et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep*. 2018; doi: 10.1016/j.celrep.2018.03.050.

70. 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; doi: 10.1101/gr.257246.119.

71. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics*. 2013; doi: 10.1093/bioinformatics/btt543.

72. 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; doi: 10.1038/nbt.3122.

73. 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; doi: 10.1007/978-1-0716-4152-1_7.

74. Chambers MC, Jagtap PD, Johnson JE, McGowan T, Kumar P, Onsongo G, et al. An Accessible Proteogenomics Informatics Resource for Cancer Researchers. *Cancer Res*. 2017; doi: 10.1158/0008-5472.CAN-17-0331.

75. Jagtap PD, Johnson JE, Onsongo G, Sadler FW, Murray K, Wang Y, et al. Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. *J Proteome Res*. 2014; doi: 10.1021/pr500812t.

76. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011; doi: 10.1038/nbt.1883.

77. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods*. 2017; doi: 10.1038/nmeth.4256.

78. Teo GC, Polasky DA, Yu F, Nesvizhskii AI. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *J Proteome Res*. 2021; doi: 10.1021/acs.jproteome.0c00544.

79. 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; doi: 10.1038/s41467-023-40129-9.

80. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*. 2007; doi: 10.1038/nmeth1113.

81. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003; doi: 10.1021/ac0341261.

82. 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; doi: 10.1038/s41592-020-0912-y.

83. Kumar P, Johnson JE, Easterly C, Mehta S, Sajulga R, Nunn B, et al. A Sectioning and Database Enrichment Approach for Improved Peptide Spectrum Matching in Large, Genome-Guided Protein Sequence Databases. *J Proteome Res*. 2020; doi: 10.1021/acs.jproteome.0c00260.

84. Wen B, Zhang B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. *Nat Commun*. 2023; doi: 10.1038/s41467-023-37462-4.

85. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. Combining results of multiple search engines in proteomics. *Mol Cell Proteomics*. 2013; doi: 10.1074/mcp.R113.027797.

86. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011; doi: 10.1038/nbt.1754.

87. 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; doi: 10.1093/nar/gkad987.

88. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. 2014; doi: 10.1093/bioinformatics/btu548.

89. 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; doi: 10.1007/978-1-4939-8546-3_12.

90. 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; doi: 10.4049/jimmunol.1700893.

91. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; doi: 10.1093/bioinformatics/bts635.

92. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019; doi: 10.1038/s41587-019-0201-4.

93. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Res*. 2020; doi: 10.12688/f1000research.23297.2.

94. 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; doi: 10.1016/j.mcpro.2023.100591.

95. 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; doi: 10.12688/f1000research.16450.2.

96. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*. 2014; doi: 10.1038/nmeth.3144.

97. Adusumilli R, Mallick P. Data Conversion with ProteoWizard msConvert. *Methods Mol Biol*. 2017; doi: 10.1007/978-1-4939-6747-6_23.

98. Diop D, Chauvin C, Jean-Jean O. Aminoglycosides and other factors promoting stop codon readthrough in human cells. *C R Biol*. 2007; doi: 10.1016/j.crvi.2006.09.001.

99. 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; doi: 10.3389/fimmu.2023.1107576.

100. Pandey K, Ramarathinam SH, Purcell AW. Isolation of HLA Bound Peptides by Immunoaffinity Capture and Identification by Mass Spectrometry. *Curr Protoc*. 2021; doi: 10.1002/cpz1.92.

101. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat Protoc*. 2019; doi: 10.1038/s41596-019-0133-y.

102. Bassani-Sternberg M. Mass Spectrometry Based Immunopeptidomics for the Discovery of Cancer Neoantigens. *Methods Mol Biol*. 2018; doi: 10.1007/978-1-4939-7537-2_14.

103. Li K, Teo GC, Yang KL, Yu F, Nesvizhskii AI. diaTracer enables spectrum-centric analysis of diaPASEF proteomics data. *Nat Commun*. 2025; doi: 10.1038/s41467-024-55448-8.

104. 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; doi: 10.1038/s41467-023-39869-5.

105. Demichev V, Szyrwiel L, Yu F, Teo GC, Rosenberger G, Niewienda A, et al. dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat Commun*. 2022; doi: 10.1038/s41467-022-31492-0.

106. : Intergalactic Utilities Commission. https://galaxyproject.org/iuc/ Accessed 2025 Mar 24.

107. : Galaxy | Tool Shed. https://toolshed.g2.bx.psu.edu/ Accessed 2025 Mar 24.

108. Cargile BJ, Bundy JL, Stephenson JL. Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res*. 2004; doi: 10.1021/pr049946o.

109. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods*. 2020; doi: 10.1038/s41592-019-0638-x.

110. Chen Z, He X. Application of third-generation sequencing in cancer research. *Med Rev (2021)*. 2021; doi: 10.1515/mr-2021-0013.

111. 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; doi: 10.1016/j.mcpro.2023.100506.

112. 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; doi: 10.1093/bib/bbae625.

113. : FragPipe/LICENSE at develop · Nesvilab/FragPipe. GitHub. https://github.com/Nesvilab/FragPipe/blob/develop/LICENSE Accessed 2025 Mar 24.

114. : galaxy/LICENSE.txt at dev · galaxyproject/galaxy. GitHub. https://github.com/galaxyproject/galaxy/blob/dev/LICENSE.txt Accessed 2025 Mar 24.

FIGURE LEGENDS

Figure 1. The standard workflow for MS-based immunopeptidomics. Cells or tissues are lysed, and anti-MHC antibody is used to affinity purify the peptide-MHC complex. MHC-bound peptides are eluted from the enriched complex and analyzed by LC-MS/MS. The peptide MS/MS spectra are matched to sequences contained in a protein database, which contains both the reference proteome sequence and potential non-reference neoantigen sequences predicted from NGS data.

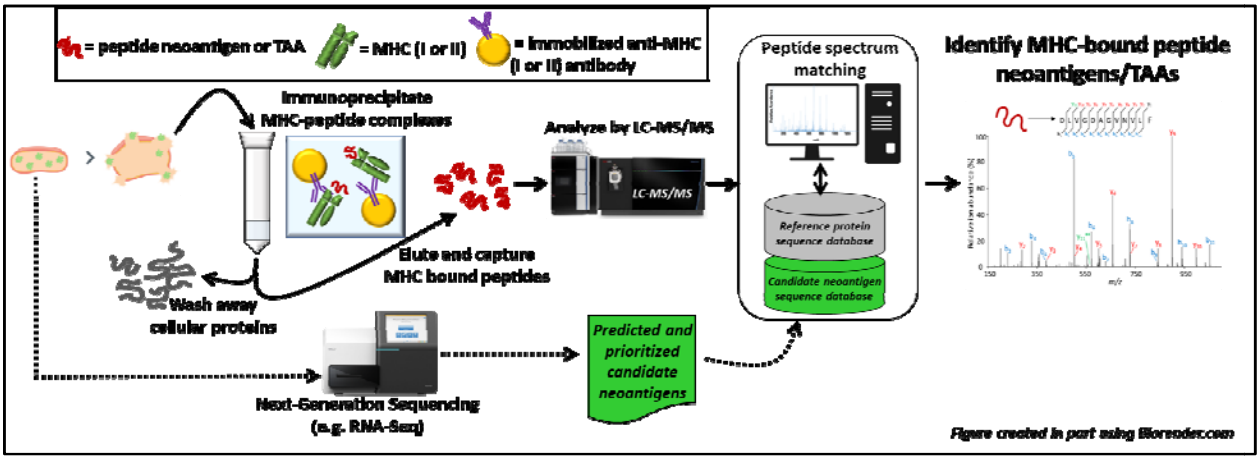


Figure 2. The modules comprising the Galaxy-based iPepGen pipeline. The modules carryout the following functions: 1) Predict neoantigen candidates from NGS data and generate customized protein sequence databases, including reference and non-reference neoantigen candidate sequences; 2) Discover neoantigen 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.

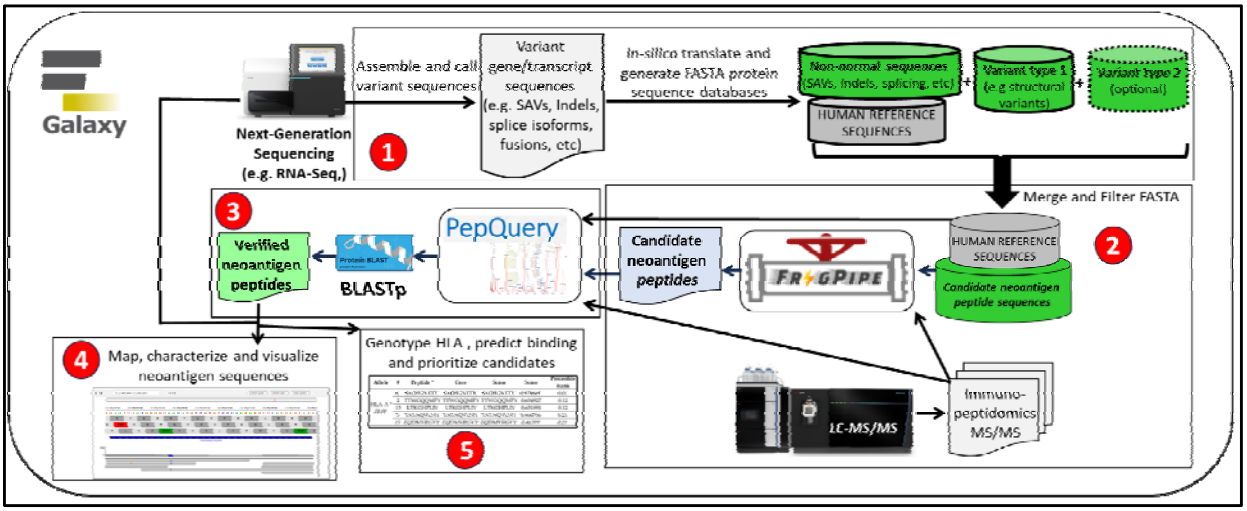


Figure 3. Visualization of a non-reference neoantigen peptide coded by a 5' UTR transcript region: Example of a non-reference neoantigen peptide sequence (RLMGEGTSSL, shown in green shaded box) coded by a transcript sequence mapped to a portion of the 5' UTR of an annotated gene structure and visualized in the Integrated Genome Viewer (IGV). Assembled RNA-Seq reads coding this peptide are in pink; B) A zoom-out 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.

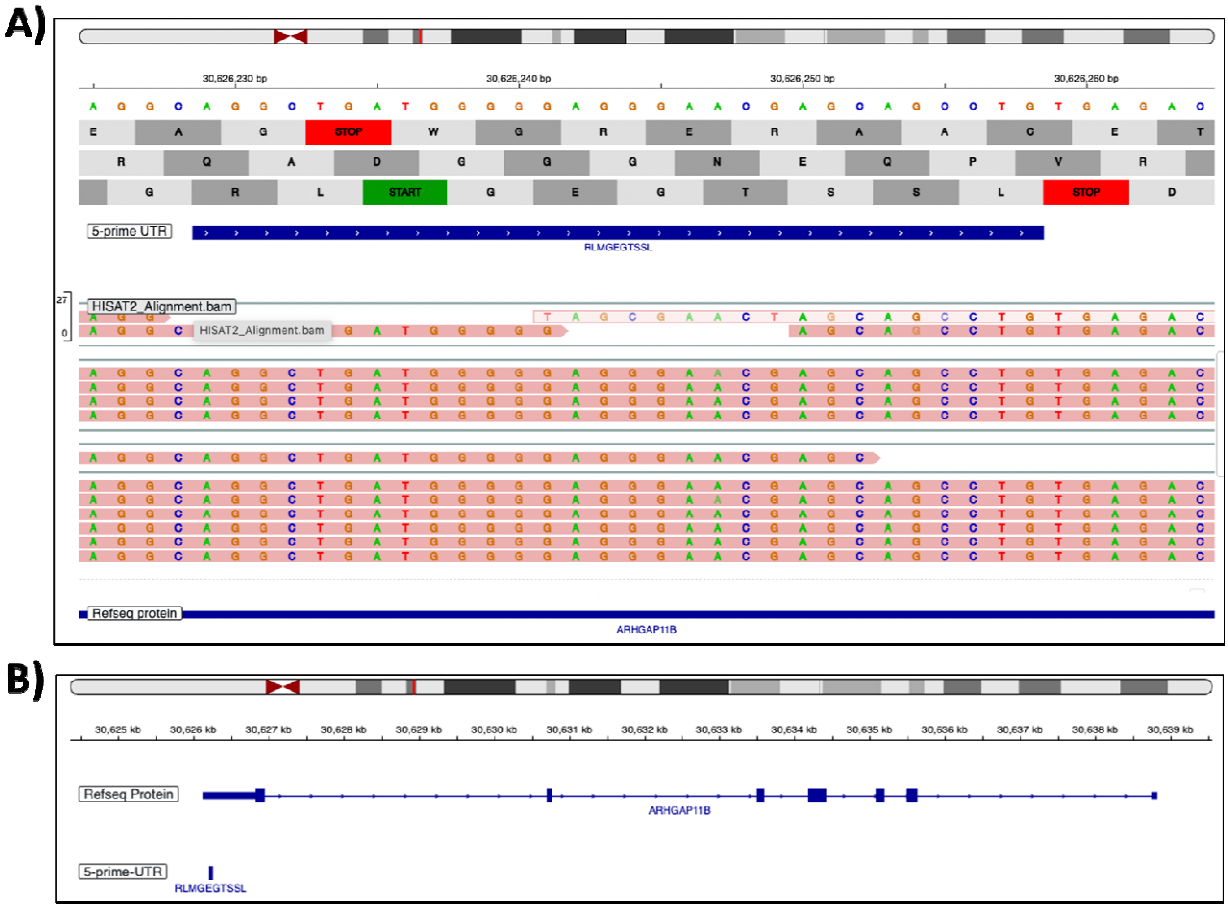
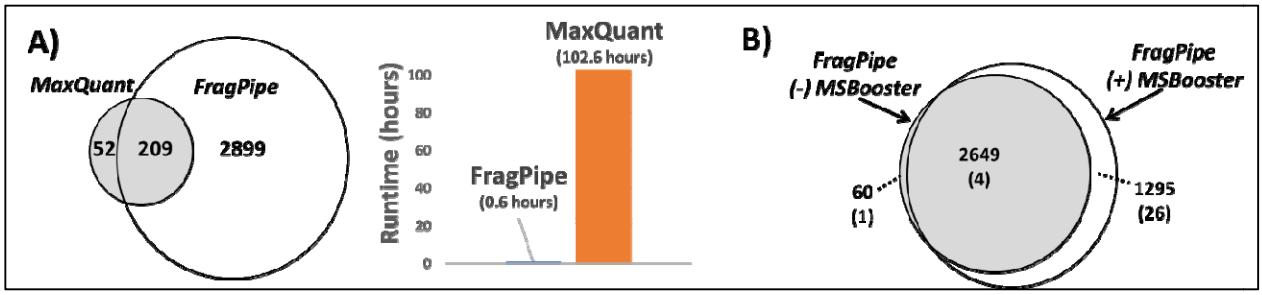
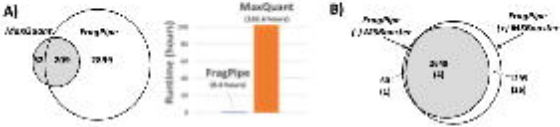


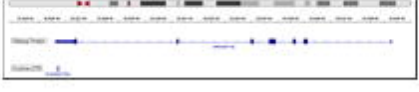
Figure 4. Performance comparison of peptide identification tools for immunopeptidomics data analysis: **A.** Comparison of FragPipe and MaxQuant performance for a "no enzyme" database search of a single peptidomics MS/MS file against the reference human proteome database from MPNST data in Module 1; both software were run using their stand-alone GUI using 8 cores and 64GB memory. **B.** Comparison of FragPipe run without and with MSBooster functionality on a single immunopeptidomics MS/MS datafile from MPNST cells and searched against a sequence database of over two million protein sequences, including the reference and annotated non-reference candidate neoantigen sequences; number of unique peptides identified are shown, with numbers in parentheses () indicating the number of candidate neoantigen peptides identified with sequences that do not map to the reference proteome.

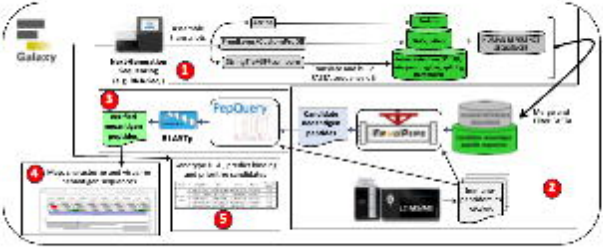
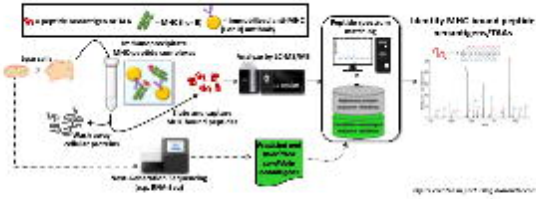



A)
MaxQuant
FragPipe
Runtime (hours)
MaxQuant
FragPipe
B)
FragPipe
FragPipe

A)

B)




Galaxy
1
2
3
4
5
PepQuery
BLAST