

# MhcVizPipe: A Quality Control Software for Rapid Assessment of Small- to Large-Scale Immunopeptidome Datasets

## Authors

Kevin A. Kovalchik, Qing Ma, Laura Wessling, Frederic Saab, Jérôme D. Duquette, Peter Kubiniok, David J. Hamelin, Pouya Faridi, Chen Li, Anthony W. Purcell, Anne Jang, Eustache Paramithiotis, Marco Tognetti, Lukas Reiter, Roland Bruderer, Joël Lanoix, Éric Bonnell, Mathieu Courcelles, Pierre Thibault, Etienne Caron, and Isabelle Sirois

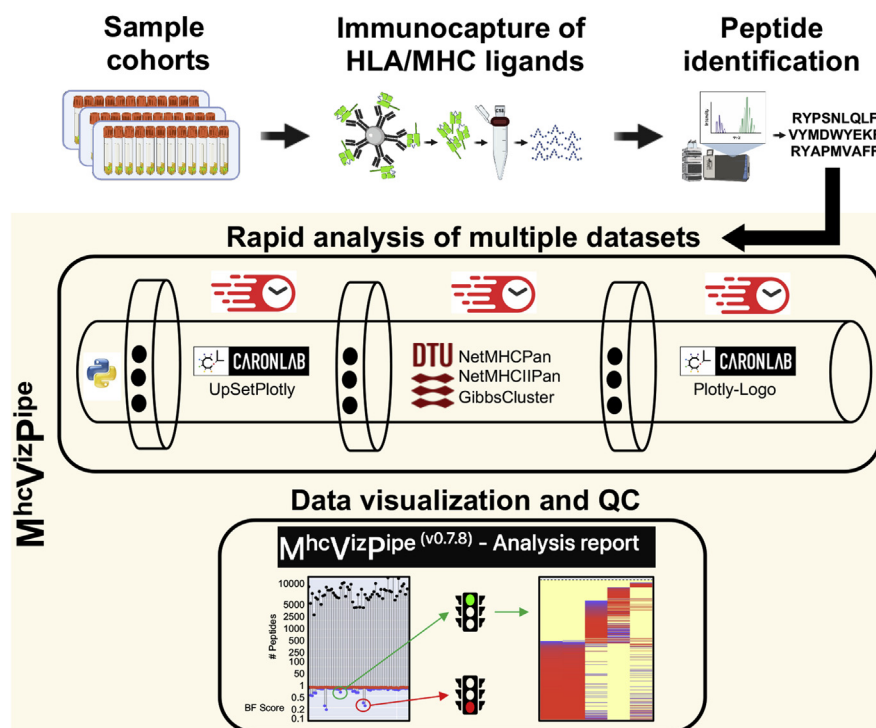
## Correspondence

[etienne.caron@umontreal.ca](mailto:etienne.caron@umontreal.ca);  
[isabelle.sirois.phd@gmail.com](mailto:isabelle.sirois.phd@gmail.com)

## Graphical Abstract

### In Brief

Automated QC software tools capable of detecting sample and/or measurement issues are important for downstream data interpretation. We created MhcVizPipe, the first semiautomated GUI-based QC software tool that enables rapid and simultaneous assessment of multiple MHC class I and II immunopeptidomic datasets generated by MS. MVP is intuitive to use, generate QC reports in HTML formats, and will find utility in any immunopeptidomic laboratory.



## Highlights

- MhcVizPipe is a software tool for QC in immunopeptidomics.
- It provides numerical scores for QC analysis of tumor biopsies.
- It performs rapid QC analysis of large clinical immunopeptidomic sample cohorts.
- It generates organized and easy-to-understand reports in HTML format.



# MhcVizPipe: A Quality Control Software for Rapid Assessment of Small- to Large-Scale Immunopectidome Datasets

Kevin A. Kovalchik<sup>1</sup>, Qing Ma<sup>2</sup>, Laura Wessling<sup>1</sup>, Frederic Saab<sup>1</sup>, Jérôme D. Duquette<sup>1</sup>, Peter Kubiniok<sup>1</sup>, David J. Hamelin<sup>1</sup>, Pouya Faridi<sup>3</sup>, Chen Li<sup>3</sup>, Anthony W. Purcell<sup>3</sup> , Anne Jang<sup>4</sup>, Eustache Paramithiotis<sup>4</sup>, Marco Tognetti<sup>5</sup> , Lukas Reiter<sup>5</sup>, Roland Bruderer<sup>5</sup>, Joël Lanoix<sup>6</sup>, Éric Bonneil<sup>6</sup>, Mathieu Courcelles<sup>6</sup> , Pierre Thibault<sup>6,7</sup>, Etienne Caron<sup>1,8,\*</sup>, and Isabelle Sirois<sup>1,\*</sup>

MS-based immunopectidomics is maturing into an automatized and high-throughput technology, producing small- to large-scale datasets of clinically relevant major histocompatibility complex (MHC) class I-associated and class II-associated peptides. Consequently, the development of quality control (QC) and quality assurance systems capable of detecting sample and/or measurement issues is important for instrument operators and scientists in charge of downstream data interpretation. Here, we created MhcVizPipe (MVP), a semiautomated QC software tool that enables rapid and simultaneous assessment of multiple MHC class I and II immunopectidomic datasets generated by MS, including datasets generated from large sample cohorts. In essence, MVP provides a rapid and consolidated view of sample quality, composition, and MHC specificity to greatly accelerate the “pass-fail” QC decision-making process toward data interpretation. MVP parallelizes the use of well-established immunopectidomic algorithms (NetMHCpan, NetMHCIIpan, and GibbsCluster) and rapidly generates organized and easy-to-understand reports in HTML format. The reports are fully portable and can be viewed on any computer with a modern web browser. MVP is intuitive to use and will find utility in any specialized immunopectidomic laboratory and proteomics core facility that provides immunopectidomic services to the community.

The importance of MS-based immunopectidomics for the discovery of T-cell targets in autoimmunity (1, 2), cancer (3–5), and infectious diseases (6–10)—including pandemic pathogens (11, 12)—has attracted the interest of investigators from a wide range of clinical disciplines, leading to the creation of

the human immunopectidome project (13–15). In fact, the growing interest for clinical immunopectidomics was recently accelerated by the unquestionable contribution to the 2018 Nobel prize winners James P. Allison and Tasuku Honjo for their work on cancer immunotherapy (16). In addition, the development of next-generation MS technologies and methods has fostered the discovery of new classes of actionable tumor-specific antigens in various cancer types (17–22). In its simplest form, clinical immunopectidomics involves the isolation of human leukocyte antigen (HLA) class I-associated and class II-associated peptides from patient biospecimens by immunoaffinity capture, followed by peptide release, and subsequent peptide sequence identification by MS combined with advanced bioinformatics (23). Once the HLA ligands have been confidently identified, their immunogenicity can be evaluated to further guide the development of vaccines and T-cell-based therapies in translational laboratories (24).

Data quality is a cornerstone of solid research, demanding repeatability and reproducibility (25). In that respect, assessing and controlling the quality of immunopectidomic data generated by MS is of utmost importance. In genomics and MS-based proteomics, the importance of quality control (QC) and quality assurance has been long acknowledged, and various grades of QC samples (i.e., QC1, QC2, and QC3) (26) as well as QC software solutions have been extensively developed and applied over the years (27–36). In contrast, in MS-based immunopectidomics, QC samples and QC software tools specialized for major histocompatibility index (MHC) class I-associated and class II-associated peptides

From the <sup>1</sup>CHU Sainte-Justine Research Center, Montreal, Quebec, Canada; <sup>2</sup>School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, Ontario, Canada; <sup>3</sup>Infection and Immunity Program and Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Victoria, Australia; <sup>4</sup>CellCarta, Montreal, Quebec, Canada; <sup>5</sup>Biognosys, Schlieren, Switzerland; <sup>6</sup>Institute of Research in Immunology and Cancer, Montreal, Quebec, Canada; <sup>7</sup>Department of Chemistry, Université de Montréal, Montreal, Quebec, Canada; <sup>8</sup>Department of Pathology and Cellular Biology, Faculty of Medicine, Université de Montréal, Quebec, Canada

\* For correspondence: Isabelle Sirois, [isabelle.sirois.phd@gmail.com](mailto:isabelle.sirois.phd@gmail.com); Etienne Caron, [etienne.caron@umontreal.ca](mailto:etienne.caron@umontreal.ca).

remain poorly documented in spite of their importance for successful therapeutic development (37).

To date, only two studies have focused on QC measures to validate the quality of immunopeptidomic data generated by MS (37, 38). In the first study, Ghosh *et al.* (37) described the importance of (i) the mass accuracy of the obtained peptides, (ii) the fitness of the observed and calculated retention times, (iii) the repeatability of the retention times and the signal intensities of the detected peptides, (iv) the use of synthetic peptides (light or heavy) to determine the specificity and limit of detection of the HLA ligands of interest, and (v) the technical and biological reproducibility. The identification scores of the peptides and their identification with different search engines can also be applied to assess data quality (39). In the second QC study, Fritsche *et al.* (38) presented (i) statistics to enable discrimination of true HLA ligands from coisolated HLA-independent proteolytic fragments, (ii) the necessary steps to ensure system suitability of the chromatographic system, (iii) an algorithm for detection of source fragmentation events that are introduced by electrospray ionization during MS, and (iv) an experimental pipeline that enables high-throughput sequence verification through similarity of fragmentation patterns and coelution of synthetic isotope-labeled internal standards. Akin to MS-based proteomics, such QC approaches are useful to show the overall quality of immunopeptidomic datasets in addition to point out limitations and pitfalls that are critical for individual peptides.

Other basic and essential steps to assess the overall quality and MHC specificity of immunopeptidomic datasets are to quantify the total number of peptides per sample, length of the detected peptides, number of strong binders (SBs), weak binders (WBs), and non-MHC binders (NBs) per sample, number of SBs, WBs, and NBs per allele per sample, number of peptides making up each sequence motif, and fraction of each sequence motif attributed to each MHC allele. Current software tools used to assess such QC measures to determine the MHC specificity of immunopeptidomic datasets include MHC peptide-binding prediction algorithms and clustering tools, such as NetMHCpan (40, 41), MHCFlurry (42, 43), GibbCluster (44), and MoDec (45). Although widely used, these algorithms were not purposely built for QC in MS-based immunopeptidomics, and as a result, can process only one sample at a time and generally require further human-based data manipulations (e.g., in Excel)—a relatively time-intensive and error-prone procedure that is not sustainable for QC in large-scale MS-based immunopeptidomics studies, as recently reported (46, 47). Hence, the development of automated or semiautomated QC software tools for rapid and simultaneous quality assessment of the MHC specificity of multiple immunopeptidomic datasets generated by MS has yet to be developed.

In this technical report, we document QC in MS-based immunopeptidomics by presenting MhcVizPipe (MVP). MVP

is an open-source and freely available QC software tool with an intuitive graphical user interface (GUI), intended to be used by any immunopeptidomic laboratory. MVP builds upon the algorithms mentioned previously (NetMHC suite tools and GibbCluster) and, once installed, provides a semi-automated and fast postdata acquisition system to assess and control sample quality and MHC specificity through effective visualization of one or multiple immunopeptidomic datasets in an HTML report (Fig. 1A). Reported QC metrics are peptide length, numbers of MHC binders, distribution of binders per MHC allele, and prominent MHC peptide-binding motifs (see examples of HTML reports in supplemental Data S1–S4). In addition, MVP computes length fraction (LF) scores and binding fraction (BF) scores for each sample to rapidly highlight problematic samples while analyzing large immunopeptidomic sample cohorts. In this regard, a command line interface (CLI) is also available for large-scale/“batch” analyses (<https://github.com/CaronLab/MhcVizPipe/wiki/Command-line-interface>). Later, we describe (i) the installation procedure of MVP, (ii) the MVP GUI for data upload, (iii) the multiple computational steps that are automatically performed within the MVP-HTML reporting pipeline, (iv) the content of an HTML report, (v) QC analysis of low-quality to high-quality biopsies, (vi) the speed performance of MVP, and (vii) QC analysis of a large cohort composed of 152 samples. We also discuss the current limitations of MVP and how the software could be further developed to support QC in large-scale clinical immunopeptidomic studies.

## EXPERIMENTAL PROCEDURES

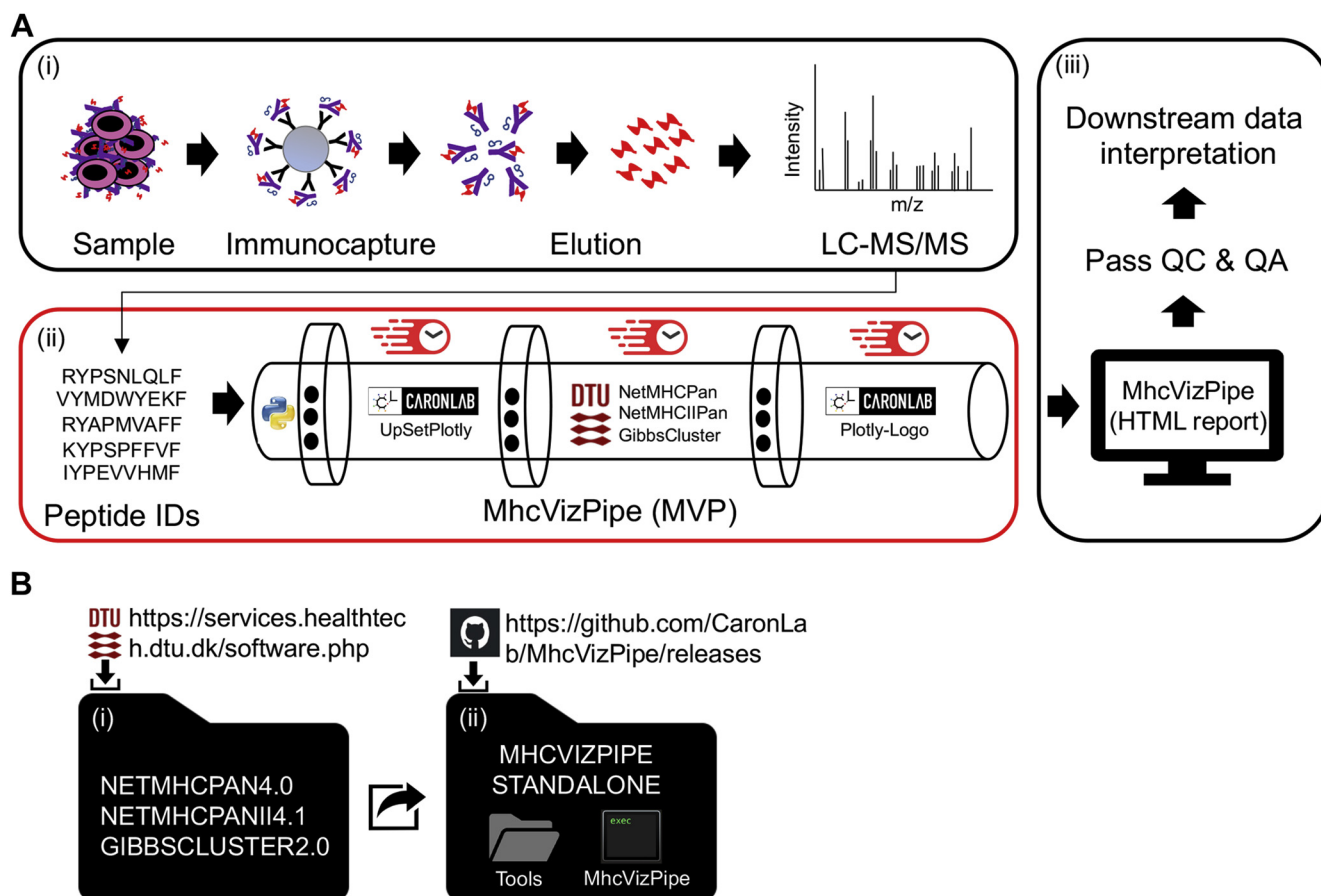
### System Requirements

**Operating System**—MVP runs natively on Linux (e.g., Ubuntu) or MacOS systems. It can also be installed on Windows 10 systems by using the Windows Subsystem for Linux (WSL) to run NetMHCpan, NetMHCIIpan, and GibbCluster. MVP was tested on the following systems: Ubuntu 16.04, 18.04, and 20.04; Linux Mint 18; MacOS 10.13 HighSierra; 10.14 Mojave; 10.15 Catalina; 11.2.3 Big Sur; and Windows 10 with both WSL 1 and WSL 2.

**Memory and Processor**—There are no hard memory or central processing unit requirements for MVP. However, because MVP utilizes multithreading, performance increases on systems with higher numbers of central processing units. The memory usage is minimal and should not present an issue to any recent desktop or laptop computer.

### Components of MVP

MVP connects the bioinformatics tools NetMHCpan, NetMHCIIpan, and GibbCluster with Python visualization libraries to create portable HTML reports for interpretation of immunopeptidomics MS data. In addition to NetMHCpan, NetMHCIIpan, and GibbCluster, it makes use of the following third-party Python libraries: Plotly, PlotlyDash, Dash Bootstrap Components, Pandas, Numpy, Dominate, Upset-Plotly, Waitress, and PlotlyLogo (PlotlyLogo and UpsetPlotly developed in-house and described later). The GUI of MVP is built using the PlotlyDash library and runs as a local web application (i.e., it runs in a web browser).



**FIG. 1. Overview and installation of the MhcVizPipe (MVP) software.** A, illustration showing the (i) experimental workflow for the isolation and identification of MHC-associated peptides by LC-MS/MS, (ii) the software tools integrated into the MVP pipeline for rapid data processing, and (iii) the HTML report generated to support (or not) downstream immunopeptidomic data interpretation through quality control (QC) and quality assurance (QA). The MVP pipeline was created with the Python language and parallelizes in-house (UpSetPlotly and Plotly-Logo) and established (NetMHCpan, NetMHCIIpan, and GibbsCluster) algorithms. B, installation steps. The user can download and install the MVP software from <https://github.com/CaronLab/MhcVizPipe> on Linux, Mac, and Windows 10 (using the Windows Subsystem for Linux). The installation requires the download of the third-party software tools and the installation of MVP.

### Peptide List Preprocessing

Prior to analysis, peptide lists are stripped of chemical modifications (e.g., oxidation of methionine and carbamidomethylation), any peptides containing nonstandard amino acids are removed, and flanking amino acids are removed if found (e.g., P and V in P.KAPDNRETL.V). It is assumed that flanking amino acids are separated from the main sequence with dots, as indicated in the example. Formats differing from this will require preprocessing by the user.

### Length Distributions and Intersecting Peptide Sets

Length distributions are visualized for all peptides  $\leq 30$ -mer in the input. If multiple samples are analyzed, the intersections of the peptide sequence sets arising from the samples are visualized in an UpSetPlot-type figure (48) using the Python library UpSetPlotly.

### NetMHCpan and NetMHCIIpan

Prior to analysis with NetMHCpan or NetMHCIIpan, subsets of the peptide lists are created as follows: for class I, the lengths are restricted to between 8 and 12 mer inclusive; for class II, the lengths are restricted to between 9 and 22 mer inclusive. These lists are

analyzed using either NetMHCpan or NetMHCIIpan to yield binding predictions (eluted ligand [EL] percent rank) used to annotate the peptides as follows: for class I, the rank of SBs is  $\leq 0.5$ , the rank of WBs is  $\leq 2.0$ , and the rank of NBs is  $> 2.0$ ; for class II, the rank of SBs is  $\leq 2.0$ , the rank of WBs is  $\leq 10$ , and the rank of NBs is  $> 10$ . The results are presented in tabular format as well as with a bar plot and heatmap. The bar plot shows the number of peptides *versus* aggregate binding strength (e.g., a peptide that has a weak affinity for one allele and a strong affinity for another is counted as an SB). The heatmap is sorted by percent rank EL score from left to right (i.e., the peptides are ordered first by the left-most column, then the second-left-most, and so on). To prevent the convolution of the ordering by poorly scoring peptides, all values greater than 2.5 for class I or 12 for class II are set to 2.5 or 12, respectively. The colormap of the heatmap is set such that red approximately represents SBs, blue approximately represents WBs, and yellow represents NBs with a linear gradient between the colors.

### LF and BF Scores

The LF score is the fraction of all peptides that are 8 to 14 mers in length for class I peptides or 8 to 25 mers in length for class II



peptides. The BF score is the fraction of all peptides within the above length range, which are predicted to be either WBs or SBs by NetMHCpan or NetMHCIIpan.

### GibbsCluster

MVP performs two GibbsCluster routines, which we have termed “Unsupervised GibbsCluster” and “Allele-Specific GibbsCluster.” There is a tab for each of these in the “Sequence Motifs” section of the report. The unsupervised GibbsCluster is a standard GibbsCluster run using all peptides in the subset described previously. The following parameters are used, which are the recommended defaults for class I and II peptides on the GibbsCluster-2.0 server (<https://services.healthtech.dtu.dk/service.php?GibbsCluster-2.0>): -g 1-6 -T -j 2 -C -D 4 -I 1 (class I); -g 1-6 -k 1 -T -j 2 (class II). Note that the grouping setting (-g) is actually run one group at a time, as explained in the following section. The grouping with the highest Kullback-Leibler Distance (KLD) score is presented in the report. The “Allele-specific GibbsCluster” is dependent upon the results of NetMHCpan or NetMHCIIpan. A subset of peptides is created for each allele, such that all the peptides in a subset are SBs or WBs for the respective allele. An additional subset is created in which peptides predicted to be NBs for all the alleles are combined (*i.e.*, the peptides in this set are not predicted binders for any allele). Any subset containing less than 20 peptides is discarded. Each of the remaining subsets is run in GibbsCluster using the aforementioned parameters with -g set to 1, forcing GibbsCluster to look for only one peptide group. An exception is the subset of NBs, in which -g is set to 1 to 5, allowing GibbsCluster to look for multiple groups in these unannotated peptides. As aforementioned, the grouping with the highest KLD score is presented in the report. The results are reported in the “Allele-Specific GibbsCluster” tab, where we see one motif for each allele present in the sample, and up to five motifs for the nonbinding peptides. All the peptide groups shown in the “Sequence Motifs” section are visualized using the PlotlyLogo library as described later.

### Multiprocessing in NetMHCpan, NetMHCIIpan, and GibbsCluster

To shorten the overall analysis time, MVP parallelizes the use of NetMHCpan, NetMHCIIpan, and GibbsCluster. Because NetMHCpan and NetMHCIIpan do not take advantage of multiprocessing, in MVP, the peptide lists are broken into smaller lists, which are concurrently analyzed by separate instances of the respective software. The results are then combined back into a single list. GibbsCluster does offer multiprocessing, analyzing the different grouping possibilities (*e.g.*, 1, 2, 3, 4, or 5 groups) concurrently. Unfortunately, with larger peptide lists, the single-group clustering is often long running, which makes it difficult to schedule consecutive GibbsCluster analyses in an efficient manner. To address this, MVP splits the analysis into individual instances of GibbsCluster, each analyzing only a single grouping possibility. For example, instead of using the command line parameter “-g 1-5” as indicated in the previous sections, in actuality, MVP creates five jobs with the parameters -g 1, -g 2, -g 3, and so on. With a single sample, this does not result in a speed improvement, but with multiple samples, MVP is able to efficiently schedule the jobs from multiple samples to utilize all available processors.

### UpSetPlotly

UpSetPlotly is an open-source Python package, based upon UpSet (48), developed in-house for visualizing intersecting sets using the Plotly Python library. The source code is available at <https://github.com/kevinkovalchik/UpSetPlotly>, and it can be installed as “upsetplotly” from PyPI using the Pip package manager.

### PlotlyLogo

PlotlyLogo is an open-source Python package developed in-house for generating sequence logos from sequence alignment data and is available at <https://github.com/kevinkovalchik/Plotly-Logo>. PlotlyLogo was designed for the specific purpose of generating sequence logos from sequence alignments as native Python objects using the Plotly plotting framework. As such, the version used in this report (plotly-logo, version 0.0.2) does not include much of the functionality of more complete solutions such as Seq2Logo, but it has the advantage of utilizing a modern Python framework (compatible with Python3) and of generating figures as Python objects, which can be directly used in other Python code. The choice to develop PlotlyLogo rather than using an existing Python solution such as LogoMaker (49) was influenced by the desire to generate live and interactive figures in a portable HTML format. It is developed in the Python programming language and requires the Python plotting library Plotly. It is installable as “plotly-logo” from PyPI using the Pip package manager. The algorithms in PlotlyLogo are based upon the methods described by Thomsen *et al.* for Seq2Logo (50) and in Immunological Bioinformatics (51). In brief, sequence alignments are read from text-formatted files, and probability matrices are generated using sequence weighting from Hobohm 1 clustering and pseudocount correction, as described (50, 51). Two types of sequence logos can be generated: Shannon and Kullback–Leibler.

### Datasets

To test the performance of EL *versus* binding affinity (BA) in annotating peptide for specific MHC alleles, the following datasets were used: H2-K<sup>b</sup> and H2-D<sup>b</sup> class I peptides extracted from mouse liver tissue (PXD008733) (52), HLA-A and HLA-B class I peptides extracted from peripheral blood mononuclear cell (PXD001872) (39), H2-IA<sup>d</sup> and H2-IE<sup>d</sup> class II peptides extracted from the mouse A20 cell line (53), and HLA-DQA10101, -DQB10501 and HLA0DQA10103, -DQB10603 class II peptides extracted from the human MAVER-1 cell line (54). To test the speed performance of MVP, the following datasets were used: HLA-ABC-associated peptides extracted from the JY cell line and multiple peripheral blood mononuclear cell samples (PXD001872) (39) and H2-K<sup>b</sup>/H2-D<sup>b</sup> class I peptides extracted from various mouse tissues (PXD008733) (52). Supplemental data in the study by Rijensky *et al.* (55) and Shraibman *et al.* (46) were used to calculate the QC scores.

### Time Speed Estimation Procedure

To benchmark the speed performance of MVP, the selected datasets were processed manually (human based) or using MVP (computer based). Time estimates were measured using a stopwatch. For the computer-based approach, the data were processed with MVP using an iMac18,2 (MacOS Big Sur, version 11.2.3) with four cores and 16 GB memory. For the human-based approach, the exact same datasets were processed manually by an experienced researcher with expertise in immunopeptidomics, with the goal of generating figures and tables that are produced in an HTML report. In other words, the human-based approach mimicked as much as possible the computer-based approach to produce the equivalent of an HTML report. Hence, for the human-based approach, time estimates were measured using a stopwatch for the following actions: (1) make graphs and tables for peptide length distribution and specificity in Microsoft Excel, (2) run NetMHCpan 4.1 and GibbsCluster online for each dataset, (3) extract output files and manipulate the data manually for making histograms and heatmaps in Microsoft Excel from the predicted MHC peptide BA scores, and (4) combine the results in a document.

### Batch Analysis

To facilitate automation or batch analysis—which might benefit from running MVP in a bash script, a python script, or as a scheduled process—MVP contains a CLI. The CLI runs MVP and saves the HTML report, NetMHCpan predictions, and PDF files of all figures in the report in a specified location. The use of the CLI is described at the following link: <https://github.com/CaronLab/MhcVizPipe/wiki/Command-line-interface>.

## RESULTS

### Installation of MVP

MVP is freely available at <https://github.com/CaronLab/MhcVizPipe> and can be downloaded and installed on Linux (e.g., Ubuntu), Mac, and Windows 10 (using the Windows Subsystem for Linux). A detailed explanation of the installation process can be found on the MVP GitHub wiki: <https://github.com/CaronLab/MhcVizPipe/wiki>. Note that for installing and running MVP on Windows 10, a few additional steps are required (see <https://github.com/CaronLab/MhcVizPipe/wiki/Windows-installation> for details). In brief, two installation options are available: (1) a ZIP file containing a standalone Python distribution, and all the Python packages required by MVP can be downloaded from the MVP GitHub repository (<https://github.com/CaronLab/MhcVizPipe/releases>) and (2) MVP can be installed from the Python Package Index into an existing Python environment. Both options also require the separate acquisition of copies of NetMHCpan-4.1, NetMHCIIpan-4.0, and GibbsCluster-2.0 from DTU Health Tech (<https://services.healthtech.dtu.dk/software.php>). Note that when new versions of NetMHCpan and GibbsCluster will be made publicly available, users will need to download the new versions themselves as there is no possibility at this time to automatically update the tools. When new versions are available, we will ensure that MVP functions with them.

Option 1 consists of a ZIP file that contains MVP, a standalone Python distribution, and all the required Python packages. The user must then copy or move the extracted DTU Health Tech tools (NetMHCpan, NetMHCIIpan, and GibbsCluster) into the “tools” folder found inside the MVP directory (Fig. 1B). MVP can then be run by double clicking the MVP executable file or running the accompanying MhcVizPipe.sh script file. This installation option does not require any use of the terminal, and the final program is fully portable (e.g., you could put it on a USB drive and use it on any compatible computer, though use in Windows requires you to copy MVP from the USB to the hard drive before use).

While we have tried to ensure the portable installation is compatible with most systems, we cannot guarantee it will work on every computer. This stems from the fact that the standalone distribution of Python has system dependencies that are usually, but not always, present. To address this issue, it is also possible to install MVP into an existing Python environment as described in “option 2.”

Option 2 is intended for users who are familiar with the command line and Python and wish to install the MVP Python package themselves or who are unable to use the portable distribution for the reasons mentioned previously. MVP can be installed in an existing Python environment ( $\geq 3.7$ ) from the Python Package Index by invoking “pip install MhcVizPipe” from the terminal. As aforementioned, the user must also acquire copies of NetMHCpan-(4.0 or 4.1), NetMHCIIpan-4.0, and GibbsCluster-2.0. For option 2, the user needs to ensure that the tools from DTU Health Tech are properly configured (according to their accompanying documentation) and reference their locations in the MVP settings window after starting the program for the first time. With this type of installation, the program is started from a terminal using this command: “python -m MhcVizPipe.gui.”



Because MVP is run as a web server, it can be accessed by any computer connected to the local network, facilitating use by multiple users. If unexpected issues are encountered during the installation procedure, please contact the authors or open an issue at the following link: <https://github.com/CaronLab/MhcVizPipe/issues>. The proper installation of MVP can be tested using example peptide lists and accompanying HTML reports available at the following link: [https://github.com/CaronLab/MhcVizPipe/tree/master/test\\_data](https://github.com/CaronLab/MhcVizPipe/tree/master/test_data).

### The MVP GUI

MVP provides a simple and intuitive GUI in any web browser (tested in Firefox, Safari, Chrome, Chromium, and Edge) (Fig. 2). Peptide lists can be uploaded in different formats (.csv, .tsv, or .txt) or copy-pasted into the GUI. Any file of the mentioned formats may be loaded provided they have columns headers or are a simple list. If a multicolumn file is opened (e.g., database search results in .csv format), the user is asked by MVP to select the header of the column that contains the peptide sequences. There is no limit to the number of files that can be analyzed at one time, though processing time increases with the number of files and the interpretability of figures and tables will decrease with excessive numbers of samples. Once the peptide lists have been uploaded, the MHC class (I or II) and alleles corresponding to the sample(s) are specified. Up to six alleles may be specified for each sample. Many technical details related to the samples can also be specified and will appear in the final report, which is generated in a portable HTML format and can be viewed in the browser or saved to the computer. MVP also generates zip files containing editable PDF versions of all plots in the report and all the data used to generate the report (i.e., NetMHCpan predictions, GibbsCluster output files).

### Automated Steps Performed Within the MVP Reporting Pipeline

Peptide lists are stripped of peptides containing chemical modifications. The complete lists are then used for generating the length histogram and UpSet plot in the “Sample overview”

A quick and user-friendly visualization tool for mass spectrometry data of MHC class I and II peptides.

[Click here for help and resources](#)

[CHECK FOR UPDATES](#) [SETTINGS](#)

To load data, use the light blue area below or copy-and-paste into the text box below (you may load any number of samples):

Drag and drop one or more files in this area, or [CLICK TO SELECT FILE\(S\)](#) (.txt, .csv, .tsv)

Note: you cannot sequentially load files with identical contents (i.e. duplicates).

Paste a peptide list or select a file using the above interface

**Sample information:**

Sample name

Sample description (optional)

[LOAD DATA](#)

**Loaded data:**

	Sample name	Description (optional)	Alleles (required)
×	biopsy_P2		HLA-A2402, HLA-B5801, HLA-C0701
×	biopsy_P4		HLA-A2301, HLA-A0201, HLA-B4403, HLA-B3801, HLA-C0401, HLA-C1203

**Allele search (click arrow to add to selected cells):**

[Select...](#)

**MHC class:**

Class I

**General information:**

Experiment description (optional)

Submitter name (optional)

**Experimental information (optional):**

Species:

# of cells:

Lysis buffer:

Type of beads:

Antibody:

Incubation time:

MHC-ligand complex elution buffer:

Peptide elution buffer:

Type of MS/MS:

Peptide identification software:

Peptide FDR:

(Enter any further information using the same format)

[GO!](#)

FIG. 2. **Overview of the MhcVizPipe (MVP) graphical user interface (GUI).** To run an analysis, the user can copy-paste a list of peptides or upload more than one file at a time. Samples can be labeled with detailed information. The user click the “GO!” button to start an analysis, and a loading screen appears while the analysis is running, followed by a pop-up window with a link to the HTML report. More details about the GUI are available at <https://github.com/CaronLab/MhcVizPipe/wiki/Usage>. A command line interface (CLI) is also available for “batch” analyses at <https://github.com/CaronLab/MhcVizPipe/wiki/Command-line-interface>.

section of the HTML report. The lists are then subset to peptides 8 to 12 or 9 to 22 mers in length for class I or class II peptides, respectively. Binding predictions for these subsets are made using NetMHCpan4.0 or 4.1 or NetMHCIIpan4.0 for each user-indicated allele, and the EL percent rank scores are used to generate the “Annotation Results” and “Binding Heatmaps” sections in the HTML report. LF and BF scores are also calculated to provide numerical values regarding the overall quality of the samples. LF score is the fraction of all peptides that are 8 to 12 mers in length for class I peptides or 9 to 22 mers in length for class II peptides; BF score is the fraction of all peptides within the appropriate length range, which are predicted to be either WBs or SBs by NetMHCpan or NetMHCIIpan (Table 1). Peptide grouping and alignment is then performed by GibbsCluster twice, once using the complete subset of peptides (yielding the results in the “Unsupervised GibbsCluster” tab) and then again for the sets of peptides predicted to bind the different alleles (yielding the results in the “Allele-Specific GibbsCluster” tab). Logos of the most prominent motifs identified by GibbsCluster (according to KLD scores) are generated using Plotly-Logo. Note that the goal of this clustering approach is not to provide new biological insights or highly accurate annotations of peptide subgroups to their alleles but to rather provide an overall assessment of the quality and MHC specificity of the data.

#### MVP Annotates Peptides Using the EL Method

NetMHCpan (version 4.1) and NetMHCIIpan (version 4.0) are critical components of the MVP software for scoring class I and II peptides, respectively (Fig. 1A). The predicted MHC BA computed by these algorithms defines SBs (%Rank <0.5), WBs (2.0 < %Rank >0.5), or NBs (%Rank >2.0) in the immunopeptidome dataset. Those predicted values are used to calculate the BF scores, which serve as an index of MHC specificity for the samples.

The methods that are available for peptide scoring in NetMHCpan and NetMHCIIpan are EL and BA (41). Here, we

compared both methods using four publicly available immunopeptidomic datasets (see [Experimental procedures](#) section) to rationalize which method to apply and integrate as part of the MVP software. For class I peptides, we found that EL and BA performed equally well ([supplemental Fig. S1](#)). Indeed, the two methods resulted in very similar numbers of SBs and WBs, both in mouse ([supplemental Fig. S1, A and B](#)) and human ([supplemental Fig. S1, C and D](#)). In contrast, for class II peptides, substantial differences were observed between EL and BA, in both mouse ([supplemental Fig. S2, A and B](#)) and human ([supplemental Fig. S2, C and D](#)). Specifically, we found that EL resulted in a nearly 4-fold increase in annotated peptides compared with BA. Furthermore, we found that the peptide groups identified by GibbsCluster contained a much higher proportion of SBs when using EL (up to 75%) *versus* BA (up to 14%) ([supplemental Fig. S2, B and D](#)). These results are in agreement with a recent study indicating that prediction scores based on EL were more accurate than those using BA, for class II peptides in particular (56). Thus, the EL method was chosen and implemented in MVP.

#### Content of the HTML Report

The HTML report generated from the MVP GUI contains three main sections: (1) sample overview, (2) annotation results and binding heatmaps, and (3) sequence motifs. Examples of HTML reports are provided in [supplemental Data S1](#) and [S2](#) for mouse and human class I peptides, respectively; and in [supplemental Data S3](#) and [S4](#) for mouse and human class II peptides, respectively.

The first section of the HTML report (Sample overview) contains (i) a peptide length distribution graph (up to 30-mers), (ii) a descriptive table indicating the total number of peptides per sample, the number of peptides corresponding to the expected length related to the MHC class selected, and LF and BF scores for each sample, and (iv) an “UpSet” plot (48) showing the number of unique and shared peptides between multiple samples.

TABLE 1  
LF and BF scores calculated from immunopeptidomic data generated from eight different tumor biopsies

Sample	Cancer type	Sample weight (mg)	Total peptides	Peptides (8–12 mers)	LF score	BF score
Biopsy 1	Head and neck (adnexal adenocarcinoma)	<13	352	284	0.81	0.62
Biopsy 2	Bile duct (cholangiocarcinoma)	60	144	68	0.47	0.07
Biopsy 3	Lung (N/A)	120	486	207	0.43	0.44
Biopsy 4	Gastric (carcinoma)	100	969	452	0.47	0.32
Biopsy 5	Vascular (hemangioendothelioma)	<13	691	585	0.85	0.93
Biopsy 6	Head and neck (squamous cell carcinoma)	<13	1018	969	0.95	0.98
Biopsy 7	Bladder (sarcomatoid carcinoma)	<13	391	323	0.83	0.92
Biopsy 8	Pancreatic (adenocarcinoma)	150	981	897	0.7	0.89

Abbreviation: N/A, not available.

Immunopeptidomic data were obtained from the study by Rijensky *et al.* Eight different biopsies and cancer types were analyzed. The sample weight, total number of peptides identified, fraction of peptides between 8 and 12 mers, and LF and BF scores are indicated. LF and BF scores are color coded to illustrate high-quality (green), middle-quality (brown-red), and low-quality (red) immunopeptidomic data.



In the second section of the HTML report (*annotation results and binding heatmaps*), MVP uses the EL scoring and annotation method described previously to generate histograms and heatmaps illustrating the proportion of immunoaffinity-purified peptides predicted to bind MHC molecules. Those graphs provide a bird's eye view on the quality and MHC specificity of the samples and are a way to visualize the BS scores calculated in the "Sample overview" section.

The third section of the HTML report focuses on the identification and visualization of sequence motifs that are enriched in the analyzed samples using GibbsCluster and our in-house Plotly-Logo algorithm, respectively (see the [Experimental procedures](#) section). The sequence motifs from the GibbsCluster analysis are presented in two ways: (i) the unsupervised tab displays the most prominent motif(s) represented by the subset of peptides in the sample along with the percent of peptides associated with each allele and (ii) the allele-specific cluster tab displays peptide motif(s) generated from SBs and WBs for each allele, as well as motifs from peptides predicted to be nonassigned MHC binders. Together, MVP generates in a few clicks a complete HTML report for assessing the quality and MHC specificity of immuno-peptidomic data generated by MS.

#### QC Analysis of Tumor Biopsies

The general quality and MHC specificity of immuno-peptidomic data are assessed using LF and BF scores, as mentioned previously. Such numerical values are useful to rapidly detect abnormalities to determine if samples are of high, middle, or low quality. To show the utility of this scoring approach, we selected a small immuno-peptidomic dataset generated from eight representative tumor biopsies, recently published in the study by Rijensky *et al.* (55). This dataset is composed of biopsies of different cancer types, including head and neck, bile duct, lung, gastric, vascular, bladder, and pancreatic cancer. Sample weight of those biopsies ranged from <13 to 150 mg, and number of peptides detected ranged from 144 peptides (bile duct; cholangiocarcinoma) to 1018 peptides (head and neck; squamous cell carcinoma) (Table 1). To calculate the LF scores, the number of peptides between 8 and 12 mers per biopsy sample was divided by the total number of peptides detected. To calculate the BF score, the number of peptides predicted to be SB or WB (NetMHCpan %Rank) were divided by the number of peptides between 8 and 12 mers. Our data show that the LF scores varied from 0.43 to 0.95 and the BF scores varied from 0.06 to 0.98 (Table 1)—a numerical value of 1 being the highest score. Overall, those samples were qualified of low to high quality. For instance, biopsies 5, 6, and 7 yielded high-quality immuno-peptidomic data, with an average combined score (LF and BF) of 0.91. Indeed, the proportions of peptides between 8 and 12 mers that were predicted to bind the HLAs expressed in these biopsy samples were relatively high, as visualized in Figure 3A. In contrast, biopsy 2 yielded low-quality immuno-peptidomic data, with an average combined score (LF and BF)

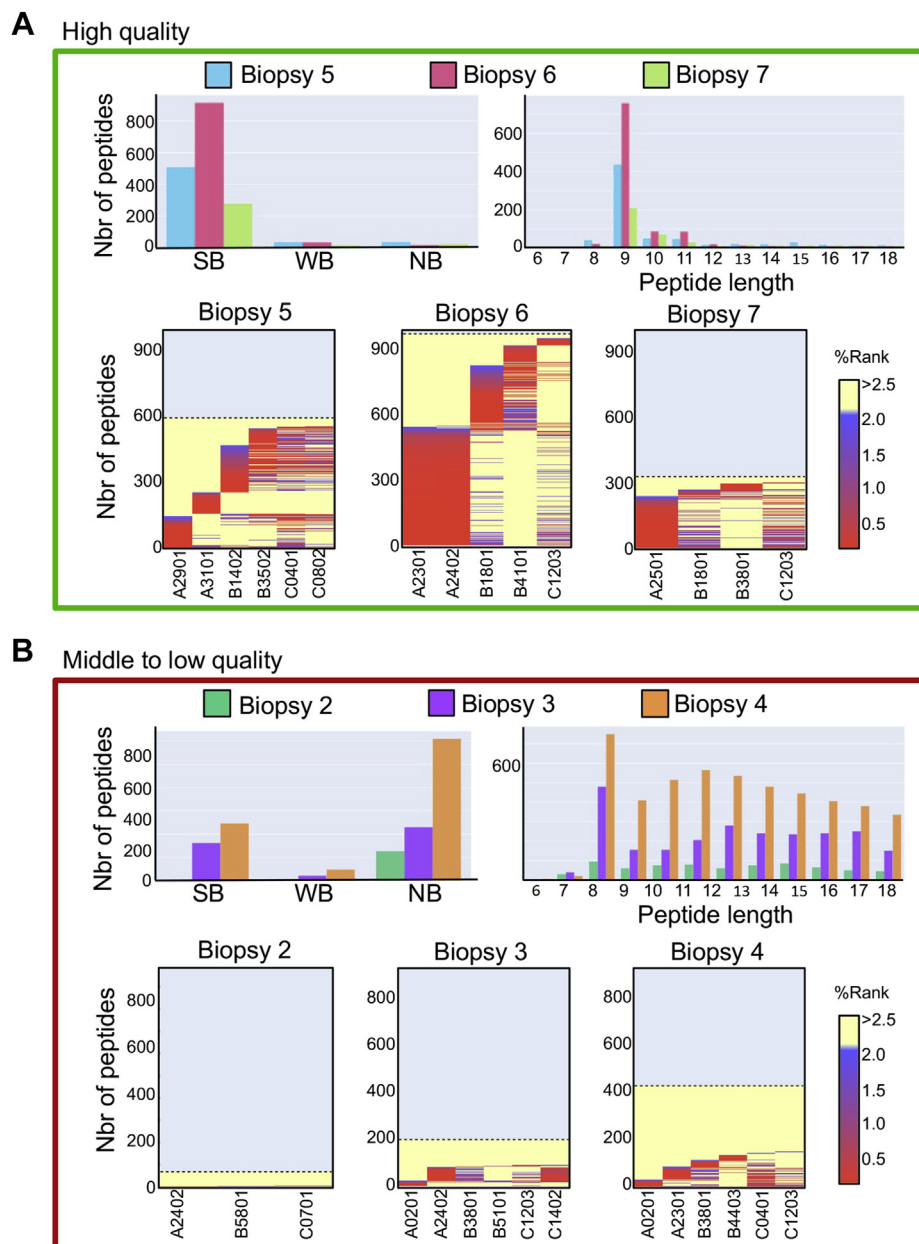
of 0.27 (Table 1). This relatively low score can be explained by the mere absence of 8 to 12 mers predicted as SB or WB (Fig. 3C). Finally, biopsies 3 and 4 yielded middle-quality immuno-peptidomic data, with an average combined score (LF and BF) of 0.42 (Table 1). This score can be explained by the noticeable presence of SB and WB within a relatively large proportion of peptides that were observed to be longer than 12 mers (likely contaminant peptides), and therefore, not predicted to bind the HLAs expressed in these biopsy samples (Fig. 3B). Thus, our analysis show that LF and BF is a simple QC scoring approach that is integrated within the MVP software tool to determine the overall quality and MHC specificity of HLA peptidomes isolated from human biopsies of various cancer types.

#### High-speed Performance of MVP

A major advantage of processing immuno-peptidomic data through MVP is the speed at which HTML reports are generated for one or multiple samples. To benchmark the speed performance of MVP, we measured the precise time estimates to assess data quality of five different datasets using (i) the human/manual-based approach, as currently done in many laboratories or (ii) the computer/MVP-based approach (see the [Experimental procedures](#) section) (Fig. 4A). Briefly, the commonly applied human-based approach includes uploading peptide lists and running multiple web interfaces, downloading the results, editing them in Excel, making multiple figures, and then copying them into a reporting document. The computer/MVP-based approach includes uploading peptide lists on the GUI, selecting the appropriate alleles, and clicking the "GO!" button (Fig. 2). The datasets that were selected to benchmark the software included a range of different samples per dataset (Fig. 4A). By comparing the two approaches, our results show that the measured time estimates varied from ~20 min (for one sample) to ~400 min (for 20 mouse tissue samples) using the conventional human/manual-based approach. Notably, the time estimates for the exact same datasets varied from 1 min (for one sample) to 18 min (for 20 mouse tissue samples) using the computer-/MVP-based approach, thereby accelerating the analysis and figure/table generation process by ~22-fold on average (Fig. 4A). Thus, MVP represents a QC software package in MS-based immuno-peptidomics and provides unprecedented speed for fast visualization and quality assessment of immuno-peptidomic data. Given its high-speed performance, MVP should find utility in QC analysis of large immuno-peptidome sample cohorts.

#### QC Analysis of a Large Immuno-peptidomic Sample Cohort

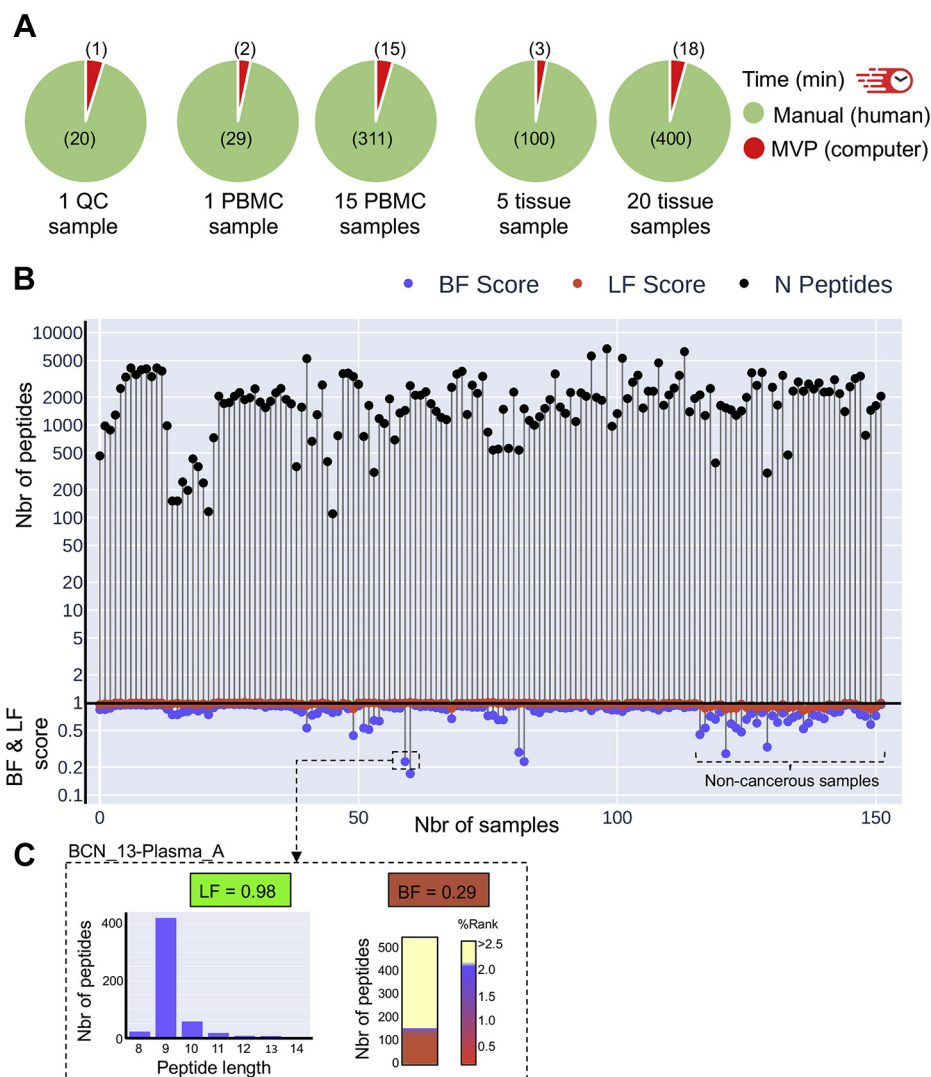
To assess the performance of MVP to run "batch" analysis, we selected a large immuno-peptidomic dataset recently published in the study by Shraibman *et al.* (46). In this study, Shraibman *et al.* (46) reported the analysis of the HLA immuno-peptidome of 152 samples, including (i) plasma-soluble HLA molecules of 142 plasma samples from both glioblastoma and noncancerous patients and (ii) membranal HLA of 10



**FIG. 3. Visualization of the QC scores (LF and BF) is calculated in Table 1.** A and B, histograms and heatmaps illustrating high-quality (A) and middle- to low-quality (B) immunopeptidomic data generated by MS from various tumor biopsies (Table 1). Histograms showing the distribution of peptides according to their length (right panel) and predicted MHC binding affinity (left panel). Heatmaps are automatically generated by MVP and provided in the HTML report. For each HLA allele, NetMHCpan provides a %Rank score for individual peptides, which are color coded on the heatmap in red, blue, and yellow for SB, WB, and NB, respectively, with a linear gradient between the colors. BF, binding fraction; HLA, human leukocyte antigen; LF, length fraction; MHC, major histocompatibility complex; MVP, MhcVizPipe; NB, nonbinder; QC, quality control; SB, strong binder; WB, weak binder.

glioblastoma tumor tissues. This analysis covers 52 different HLA allotypes, more than 35,000 different HLA-associated peptides, and represents, to the best of our knowledge, the largest immunopeptidomic dataset ever generated from plasma samples. The “batch analysis” of the 152 samples was run automatically using the CLI and a simple Python script (<https://github.com/CaronLab/MhcVizPipe/wiki/Command->

[line-interface](#)). The run time was approximately 81 min (running in Ubuntu 20.04 on an 8-core desktop). The metrics files generated when running MVP from the command line (sample\_metrics.txt) were used to generate a QC plot showing three points/scores per sample (i.e., the total number of peptides detected and the LF and BF scores) and illustrating the variability in these scores between all the samples



**FIG. 4. Speed performance of the MVP software for small- to large-scale QC analyses.** **A**, Pie charts showing the time estimates to process and analyze the selected datasets manually (human-based in green) or using MVP (computer-based in red). The following datasets were tested (from left to right): one QC sample (JY), one PBMC sample, 15 PBMC samples, five mouse tissue samples, and 20 mouse tissue samples. **B**, large-scale QC analysis of the HLA peptidome of 152 samples. Graph showing the total number of peptides and the two scores calculated based on peptide lengths (LF) and the number of peptides predicted to bind HLAs (BF) (y-axis). Each sample has three points, one for each of the values (shown in the legend). The points for each sample are connected by a vertical line. **C**, zoom in on the sample “BCN\_13-Plasma\_A.” Histogram and heat map showing the length distribution and predicted HLA-binding affinity (%Rank) of the detected peptides in this sample. LF and BF scores are indicated. BF, binding fraction; HLA, human leukocyte antigen; LF, length fraction; MVP, MhcVizPipe; PBMC, peripheral blood mononuclear cell; QC, quality control.

(Fig. 4B). In this dataset, the average LF score was 0.96 and the average BF score was 0.80, highlighting the overall high quality of this large dataset. In addition, the generated QC plot was particularly useful to point out potentially problematic samples. For instance, among the lowest BF scores in this dataset was from the sample BCN\_13-Plasma\_A (Fig. 4C). In fact, visualization of the data from this sample indicates that 151 peptides were predicted to be SB or WB, whereas 376 peptides were predicted to be NB, hence, resulting in a relatively low BF score of 0.29 (Fig. 4C). Thus, MVP can run “batch” analysis relatively rapidly to provide an assessment of

the quality of individual samples within a large immunopeptidomic dataset. Together, MVP is a generic software tool, can be applied to any HLA immunopeptidomes, and given its high-speed performance and scalability, will find utility to rapidly warn the user about problematic samples in large-scale clinical immunopeptidomic studies.

#### DISCUSSION

The development of QC software is common practice for high-quality research in the life sciences. Here, we present

MVP, a GUI-based QC approach, to rapidly and simultaneously assess the quality and MHC specificity of small to large immunopeptidomic datasets generated by MS. Although MVP can be perceived as a relatively basic software solution by experts in the field, this is, to the best of our knowledge, one of the first reports focusing on the development of QC software in MS-based immunopeptidomics. Indeed, very little has been described in the literature with regard to the development of QC approaches, despite the fact that QC will become critically important in MS-based immunopeptidomics to comply with pharmaceutical regulations for real-life clinical applications (37).

For various reasons, the development of QC software tools has not been prioritized in immunopeptidomics over the last recent years. The main reason is most likely because most immunopeptidomic studies still have to deal with relatively small-scale datasets; and the quality and MHC specificity of such datasets can still be manually assessed using the currently available peptide sequence clustering and MHC peptide-binding algorithms. However, with the recent development of automated technologies for robust and high-throughput MHC immunoprecipitation protocols, the production of immunopeptidomic data on the scale of multiple gigabytes per instrument per day can be anticipated to accelerate the pace of capturing therapeutically relevant data (57). In this regard, at least two large-scale immunopeptidomics studies have recently been reported: (1) the profiling of immunopeptidomes from 152 clinical biospecimens, which were used in this study (46) and (2) the mapping of the first draft of the human immunopeptidome from 227 benign tissue samples, for a total of nearly 3500 raw files generated by MS (47). With this perspective in mind, one can envision that other large-scale immunopeptidomic cohort studies will be conducted in the future, as recently discussed in the context of the human immunopeptidome project (14). Therefore, we believe that automated QC software solutions specialized in MS-based immunopeptidomics are inevitable. The creation of MVP represents a very first step in this direction and should stimulate the development of additional QC software solutions and products specialized in MS-based immunopeptidomics in the future.

The current version of MVP has obvious limitations. MVP was not designed to provide new biological insights. MVP is intended so far for situations where instrument operators or scientists in charge of downstream data analysis want a quick and consolidated view of the general quality, composition, and MHC specificity of one or multiple immunopeptidomic samples in parallel. Whether the user should continue using the evaluated data or not (e.g., with low-quality data) will very much depend on the goal of the project or the specific scientific question(s) asked. Therefore, we think that it is up to the user to judge what is best, if he and/or she should drop low-quality samples or keep troubleshooting them to perform additional experiments and analyses.

MVP is also subject to the accuracy of the available MHC peptide prediction algorithms and clustering tools. In this regard, new algorithms continue to be released for the prediction of HLA class I peptides (42, 43, 58) and HLA class II peptides (45, 59) and will be integrated in future versions of MVP in a yearly basis to improve the overall performance of the QC software. The predictors used in the future to upgrade MVP will also improve its accessibility. In the current version of MVP, NetMHCpan suite tools were used, although they remain poorly compatible with Windows. Therefore, installing MVP on Windows is possible, as explained in detailed in GitHub (<https://github.com/CaronLab/MhcVizPipe/wiki/Windows-installation>) but is more challenging for a noncomputer expert. Integration of Windows-compatible HLA peptide-binding algorithms, such as MHCFlurry, will help to improve the accessibility of future versions of MVP.

Over the last recent years, considerable advances in HLA ligand purification protocols, MS instrumentations, and computational methods have pushed forward the boundaries of MS-based immunopeptidomics. For instance, state-of-the-art data-independent acquisition methods have been applied (39, 60, 61), specialized computational workflows have been created (62, 63), and new automated and semiautomated HLA ligand purification platforms have been developed to increase the throughput, speed, sensitivity, and reproducibility of immunopeptidome analysis by MS (57, 64, 65). In the context of this progressive technological development, the current version of MVP will inevitably need to be further developed to reach its full QC capabilities and potential for the field. Currently, MVP is best suited for peptide datasets generated by data-dependent acquisition-MS. However, a long-term goal in immunopeptidomics is to scale to large and quantitative clinical datasets generated by data-independent acquisition-MS (39, 60, 61), and subject to stringent QC, as is performed in large-scale clinical proteomics studies (66–70). In this regard, generation of high-quality and quantitative immunopeptidomic data in large-scale studies is challenging because of issues induced by the samples themselves because of collection, storage, and processing. To help overcome this challenge, MVP could be further developed to evaluate immunopeptidomic data in near real-time—as described in proteomics (27)—allowing for interventions as soon as deviations in data quality of specific immunopeptidomic samples are detected.

Beside QC software, we also believe that future community-driven immunopeptidomic studies that focus on the evaluation and establishment of QC samples/standards are needed. In fact, QC samples in immunopeptidomics remain poorly documented to date (37). Nevertheless, such samples are critical in the long run to generate reproducible clinical datasets across multiple laboratories (13). If successful in doing so, robust and high-quality immunopeptidomic data would be consistently generated, stored, and shared through



specialized public repositories—for example, SysteMHC Atlas (71, 72) and caAtlas (73)—thereby providing a global and sustainable mechanism to foster collaborations among computational scientists, biostatisticians, immunologists, and clinical investigators to improve T-cell-based immunotherapies (13).

In conclusion, MVP is a semiautomated GUI-based QC software for rapid assessment of multiple immunopeptidomic datasets generated by MS. The current version of MVP parallelizes the use of well-established immunopeptidomic algorithms as well as new in-house software tools to assess data quality and MHC specificity at unprecedented speed and will therefore be of immediate utility for any expert and nonexpert in the field. We envision that further development of this software tool will facilitate QC for upcoming large-scale immunopeptidomic cohort studies.

#### DATA AVAILABILITY

All the data generated or analyzed during this study are included in this published article and/or the [supplemental data](#). Created datasets and code are publicly available. Immunopeptidomic data visualized in [supplemental Data S2](#) have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the dataset identifier PXD028633. All the codes are available at github: <https://github.com/CaronLab/MhcVizPipe>.

The MVP source code is open source and freely available at: <https://github.com/CaronLab/MhcVizPipe>. The version of MVP used in this publication is 0.7.9 (<https://github.com/CaronLab/MhcVizPipe/releases/tag/v0.7.9>).

**Supplemental data**—This article contains [supplemental data](#) (39, 52, 53).

**Acknowledgments**—This work was supported by funding from the Fonds de recherche du Québec-Santé, the Cole Foundation, CHU Sainte-Justine and the Charles-Bruneau Foundations, Canada Foundation for Innovation, the National Sciences and Engineering Research Council (grant no.: RGPIN-2020-05232), and the Canadian Institutes of Health Research (grant no.: 174924). Institute for Research in Immunology and Cancer proteomics facility is a Genomics Technology platform funded in part by the Canadian Government through Genome Canada.

**Funding and additional information**—K. A. K. is a recipient of the postdoctoral scholarship of the Institute for Data Valorization (grant no.: 4879287150).

**Author contributions**—K. A. K., E. C., and I. S. conceptualization; E. C., K. A. K., Q. M., L. W., M. C., and I. S. formal analysis; K. A. K., Q. M., L. W., F. S., J. D. D., P. K., D. J. H.,

P. F., C. L., A. W. P., A. J., E. P., M. T., L. R., R. B., J. L., E. B., M. C., P. T., E. C., and I. S. investigation; E. C., K. A. K., Q. M., L. W., M. C., and I. S. data curation; K. A. K., I. S., and E. C. writing—original draft; K. A. K., J. D. D., P. K., D. J. H., P. F., C. L., A. W. P., A. J., E. P., M. T., L. R., R. B., J. L., E. B., M. C., P. T., E. C., and I. S. writing—review & editing; E. C. and I. S. supervision; E. C. funding acquisition.

**Conflict of interest**—A. J. and E. P. are employees of Cell-Carda (Montreal, Canada); M. T., L. R., and R. B. are employees of Biognosys (Zürich, Switzerland). All other authors declare no competing interests.

**Abbreviations**—The abbreviations used are: BA, binding affinity; BF, binding fraction; CLI, command line interface; EL, eluted ligand; GUI, graphical user interface; HLA, human leukocyte antigen; KLD, Kullback-Leibler Distance; LF, length fraction; MHC, major histocompatibility complex; MVP, MhcVizPipe; NB, nonbinder; QC, quality control; SB, strong binder; WB, weak binder; WSL, Windows Subsystem for Linux.

Received May 21, 2021, and in revised form, October 28, 2021. Published, MCPRO Papers in Press, November 17, 2021, <https://doi.org/10.1016/j.mcpro.2021.100178>

#### REFERENCES

1. Wan, X., Vomund, A. N., Peterson, O. J., Chervonsky, A. V., Lichti, C. F., and Unanue, E. R. (2020) The MHC-II peptidome of pancreatic islets identifies key features of autoimmune peptides. *Nat. Immunol.* **21**, 455–463
2. Gonzalez-Duque, S., Azoury, M. E., Colli, M. L., Afonso, A. J., Turatsinze, J.-V., Nigi, L., Lalanne, A. I., Sebastiani, G., Carré, A., Pinto, S., Culina, S., Corcos, N., Bugliani, M., Marchetti, P., Armanet, M., *et al.* (2018) Conventional and neo-antigenic peptides presented by  $\beta$  cells are targeted by circulating Naive CD8<sup>+</sup> T cells in type 1 diabetic and healthy donors. *Cell Metab.* **28**, 946–960.e6
3. Gubin, M. M., Zhang, X., Schuster, H., Caron, E., Ward, J. P., Noguchi, T., Ivanova, Y., Hundal, J., Arthur, C. D., Krebber, W.-J., Mulder, G. E., Toebes, M., Vesely, M. D., Lam, S. S. K., Korman, A. J., *et al.* (2014) Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515**, 577–581
4. Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R., Busch, D. H., Peschel, C., *et al.* (2016) Direct identification of clinically relevant neopeptides presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404
5. Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., Modrusan, Z., Mellman, I., Lill, J. R., and Delamarre, L. (2015) Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576
6. Paes, W., Leonov, G., Partridge, T., Chikata, T., Murakoshi, H., Frangou, A., Brackenridge, S., Nicastrì, A., Smith, A. G., Learn, G. H., Li, Y., Parker, R., Oka, S., Pellegrino, P., Williams, I., *et al.* (2019) Contribution of proteasome-catalyzed peptide cis-splicing to viral targeting by CD8<sup>+</sup> T cells in HIV-1 infection. *Proc. Natl. Acad. Sci. USA* **116**, 24748–24759
7. Croft, N. P., Smith, S. A., Pickering, J., Sidney, J., Peters, B., Faridi, P., Witney, M. J., Sebastian, P., Flesch, I. E. A., Heading, S. L., Sette, A., Gruta, N. L. L., Purcell, A. W., and Tschärke, D. C. (2019) Most viral peptides displayed by class I MHC on infected cells are immunogenic. *Proc. Natl. Acad. Sci. USA* **116**, 3112–3117
8. Wu, T., Guan, J., Handel, A., Tschärke, D. C., Sidney, J., Sette, A., Wakim, L. M., Sng, X. Y. X., Thomas, P. G., Croft, N. P., Purcell, A. W., and Gruta,

- N. L. L. (2019) Quantification of epitope abundance reveals the effect of direct and cross-presentation on influenza CTL responses. *Nat. Commun.* **10**, 2846
9. Ternette, N., Yang, H., Partridge, T., Llano, A., Cedeño, S., Fischer, R., Charles, P. D., Dudek, N. L., Mothe, B., Crespo, M., Fischer, W. M., Korber, B. T. M., Nielsen, M., Borrow, P., Purcell, A. W., et al. (2015) Defining the HLA class I-associated viral antigen repertoire from HIV-1 infected human cells. *Eur. J. Immunol.* **46**, 60–69
  10. Bettencourt, P., Müller, J., Nicastri, A., Cantillon, D., Madhavan, M., Charles, P. D., Fotso, C. B., Wittenberg, R., Bull, N., Pinpathomrat, N., Waddell, S. J., Stylianou, E., Hill, A. V. S., Ternette, N., and McShane, H. (2020) Identification of antigens presented by MHC for vaccines against tuberculosis. *NPJ Vaccines* **5**, 2
  11. Parker, R., Partridge, T., Wormald, C., Kawahara, R., Stalls, V., Aggelakopoulou, M., Parker, J., Doherty, R. P., Morejon, Y. A., Lee, E., Saunders, K., Haynes, B. F., Acharya, P., Thaysen-Andersen, M., Borrow, P., et al. (2021) Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. *Cell Rep.* **35**, 109179
  12. Weingarten-Gabbay, S., Klaeger, S., Sarkizova, S., Pearlman, L. R., Chen, D.-Y., Gallagher, K. M. E., Bauer, M. R., Taylor, H. B., Dunn, W. A., Tarr, C., Sidney, J., Rachimi, S., Conway, H. L., Katsis, K., Wang, Y., et al. (2021) Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. *Cell* **184**, 3962–3980
  13. Caron, E., Aebersold, R., Banaei-Esfahani, A., Chong, C., and Bassani-Sternberg, M. (2017) A case for a human immuno-peptidome project Consortium. *Immunity* **47**, 203–208
  14. Vizcaino, J. A., Kubiniok, P., Kovalchik, K., Ma, Q., Duquette, J. D., Mongrain, I., Deutsch, E. W., Peters, B., Sette, A., Sirois, I., and Caron, E. (2020) The human immunopeptidome project: A roadmap to predict and treat immune diseases. *Mol. Cell Proteomics* **19**, 31–49
  15. Admon, A., and Bassani-Sternberg, M. (2011) The Human Immunopeptidome Project, a suggestion for yet another postgenome next big thing. *Mol. Cell Proteomics* **10**, O111.011833
  16. Kaiser, J., and Couzin-Frankel, J. (2018) Cancer immunotherapy sweeps Nobel for medicine. *Science* **362**, 13
  17. Hardy, M.-P., Vincent, K., and Perreault, C. (2019) The genomic landscape of antigenic targets for T cell-based leukemia immunotherapy. *Front. Immunol.* **10**, 2934
  18. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonnell, É., Laverdure, J.-P., Gendron, P., Courcelles, M., Hardy, M.-P., Côté, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., et al. (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516
  19. Ehx, G., Larouche, J.-D., Durette, C., Laverdure, J.-P., Hesnard, L., Vincent, K., Hardy, M.-P., Thériault, C., Rulleau, C., Lanoix, J., Bonnell, E., Feghaly, A., Apavaloaei, A., Noronha, N., Laumont, C. M., et al. (2021) Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity* **54**, 737–752
  20. Smith, C. C., Selitsky, S. R., Chai, S., Armistead, P. M., Vincent, B. G., and Serody, J. S. (2019) Alternative tumour-specific antigens. *Nat. Rev. Cancer* **19**, 465–478
  21. Minati, R., Perreault, C., and Thibault, P. (2020) A roadmap toward the definition of actionable tumor-specific antigens. *Front. Immunol.* **11**, 583287
  22. Haen, S. P., Löffler, M. W., Rammensee, H.-G., and Brossart, P. (2020) Towards new horizons: Characterization, classification and implications of the tumour antigenic repertoire. *Nat. Rev. Clin. Oncol.* **17**, 595–610
  23. Caron, E., Kowalewski, D. J., Koh, C. C., Sturm, T., Schuster, H., and Aebersold, R. (2015) Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol. Cell Proteomics* **14**, 3105–3117
  24. Leko, V., and Rosenberg, S. A. (2020) Identifying and targeting human tumor antigens for T cell-based immunotherapy of solid tumors. *Cancer Cell* **38**, 454–472
  25. National Academies of Sciences, Engineering, and Medicine. (2019) *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC. <https://doi.org/10.17226/25303>
  26. Bittremieux, W., Tabb, D. L., Impens, F., Staes, A., Timmerman, E., Martens, L., and Laukens, K. (2018) Quality control in mass spectrometry-based proteomics. *Mass Spectrom. Rev.* **37**, 697–711
  27. Stanfill, B. A., Nakayasu, E. S., Bramer, L. M., Thompson, A. M., Ansong, C. K., Clauss, T., Gritsenko, M. A., Monroe, M. E., Moore, R. J., Orton, D. J., Piehowski, P. D., Schepmoes, A. A., Smith, R. D., Webb-Robertson, B.-J., Metz, T. O., & TEDDY Study Group. (2018) QC-ART: A tool for real-time quality control assessment of mass spectrometry-based proteomics data. *Mol. Cell Proteomics* **17**, 1824–1836
  28. Bielow, C., Mastrobuoni, G., and Kempa, S. (2016) Proteomics quality control: Quality control software for MaxQuant results. *J. Proteome Res.* **15**, 777–787
  29. Kim, T., Chen, I. R., Parker, B. L., Humphrey, S. J., Crossett, B., Cordwell, S. J., Yang, P., and Yang, J. Y. H. (2019) Qcmap: An interactive web-tool for performance diagnosis and prediction of LC-MS systems. *Proteomics* **19**, e1900068
  30. Kovalchik, K. A., Colborne, S., Spencer, S. E., Sorensen, P. H., Chen, D. D. Y., Morin, G. B., and Hughes, C. S. (2018) RawTools: Rapid and dynamic interrogation of Orbitrap data files for mass spectrometer system management. *J. Proteome Res.* **18**, 700–708
  31. Gallien, S., Bourmaud, A., and Domon, B. (2014) A simple protocol to routinely assess the uniformity of proteomics analyses. *J. Proteome Res.* **13**, 2688–2695
  32. Chiva, C., Olivella, R., Borràs, E., Espadas, G., Pastor, O., Solé, A., and Sabido, E. (2018) QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLoS One* **13**, e0189209
  33. Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L., Carr, S. A., Clauser, K. R., Jaffe, J. D., Kowalski, K. A., Neubert, T. A., et al. (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **9**, 761–776
  34. Guo, Y., Ye, F., Sheng, Q., Clark, T., and Samuels, D. C. (2014) Three-stage quality control strategies for DNA re-sequencing data. *Brief Bioinform.* **15**, 879–889
  35. Guo, Y., Zhao, S., Sheng, Q., Ye, F., Li, J., Lehmann, B., Pietenpol, J., Samuels, D. C., and Shyr, Y. (2014) Multi-perspective quality control of Illumina exome sequencing data using QC3. *Genomics* **103**, 323–328
  36. Morgenstern, D., Barzilay, R., and Levin, Y. (2021) RawBeans: A simple, vendor-independent, raw-data quality-control tool. *J. Proteome Res.* **20**, 2098–2104
  37. Ghosh, M., Gauger, M., Marcu, A., Nelde, A., Denk, M., Schuster, H., Rammensee, H.-G., and Stevanović, S. (2020) Guidance document: Validation of a high-performance liquid chromatography-tandem mass spectrometry immunopeptidomics assay for the identification of HLA class I ligands suitable for pharmaceutical therapies. *Mol. Cell Proteomics* **19**, 432–443
  38. Fritsche, J., Kowalewski, D. J., Backert, L., Gwinner, F., Dörner, S., Priemer, M., Tsou, C.-C., Hoffgaard, F., Römer, M., Schuster, H., Schoor, O., and Weinschenk, T. (2021) Pitfalls in HLA ligandomics – how to catch a lie? *Mol. Cell Proteomics* **20**, 100110
  39. Caron, E., Espona, L., Kowalewski, D. J., Schuster, H., Ternette, N., Alpizar, A., Schittenhelm, R. B., Ramarathnam, S. H., Arlehamn, C. S. L., Koh, C. C., Gillet, L. C., Rabsteyn, A., Navarro, P., Kim, S., Lam, H., et al. (2015) An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife* **4**
  40. Jurtz, V., Paul, S., Andreatta, M., Marcotili, P., Peters, B., and Nielsen, M. (2017) NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368
  41. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454
  42. O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018) MHCflurry: Open-Source class I MHC binding affinity prediction. *Cell Syst.* **7**, 129–132.e4
  43. O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020) MHCflurry 2.0: Improved Pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.* **11**, 42–48.e7
  44. Andreatta, M., Alvarez, B., and Nielsen, M. (2017) GibbsCluster: Unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* **45**, W458–W463

45. Racle, J., Michaux, J., Rockinger, G. A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., Bassani-Sternberg, M., and Gfeller, D. (2019) Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286
46. Shraibman, B., Barnea, E., Kadosh, D. M., Haimovich, Y., Slobodin, G., Rosner, I., López-Larrea, C., Hilf, N., Kutruff, S., Song, C., Britten, C., Castle, J., Kreiter, S., Frenzel, K., Tatagiba, M., *et al.* (2019) Identification of tumor antigens among the HLA peptidomes of glioblastoma tumors and plasma. *Mol. Cell Proteomics* **18**, 1255–1268
47. Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D. J., Freudenmann, L. K., Backert, L., Mühlenbruch, L., Szolek, A., Lübke, M., Wagner, P., Engler, T., Matovina, S., Wang, J., Hauri-Hohl, M., Martin, R., *et al.* (2021) HLA ligand atlas: A benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J. Immunother. Cancer* **9**, e002071
48. Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., and Pfister, H. (2014) UpSet: Visualization of intersecting sets. *IEEE T Vis. Comput. Gr.* **20**, 1983–1992
49. Tareen, A., and Kinney, J. B. (2019) Logomaker: Beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274
50. Thomsen, M. C. F., and Nielsen, M. (2012) Seq2Logo: A method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40**, W281–W287
51. Lund, O., Nielsen, M., Lundegaard, C., Kesmir, C., and Brunak, S. (2005) *Immunological Bioinformatics*. The MIT Press, Cambridge, MA; London, UK
52. Schuster, H., Shao, W., Weiss, T., Pedrioli, P. G. A., Roth, P., Weller, M., Campbell, D. S., Deutsch, E. W., Moritz, R. L., Planz, O., Rammensee, H.-G., Aebersold, R., and Caron, E. (2018) A tissue-based draft map of the murine MHC class I immunopeptidome. *Sci. Data* **5**, 180157
53. Sofron, A., Ritz, D., Neri, D., and Fugmann, T. (2015) High-resolution analysis of the murine MHC class II immunopeptidome. *Eur. J. Immunol.* **46**, 319–328
54. Ritz, D., Sani, E., Debiec, H., Ronco, P., Neri, D., and Fugmann, T. (2018) Membranal and blood-soluble HLA class II peptidome analyses using data-dependent and independent acquisition. *Proteomics* **18**, e1700246
55. Rijensky, N. M., Shraga, N. R. B., Barnea, E., Peled, N., Rosenbaum, E., Popovtzer, A., Stemmer, S. M., Livoff, A., Shlapobersky, M., Moskovits, N., Perry, D., Rubin, E., Haviv, I., and Admon, A. (2020) Identification of tumor antigens in the HLA peptidome of patient-derived xenograft tumors in mouse. *Mol. Cell Proteomics* **19**, 1360–1374
56. Garde, C., Ramarathnam, S. H., Jappe, E. C., Nielsen, M., Kringelum, J. V., Trolle, T., and Purcell, A. W. (2019) Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics* **71**, 445–454
57. Zhang, L., McAlpine, P. L., Heberling, M. L., and Elias, J. E. (2020) Automated ligand purification platform accelerates immunopeptidome analysis by mass spectrometry. *J. Proteome Res.* **20**, 393–408
58. Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachiredy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., *et al.* (2019) A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209
59. Chen, B., Khodadoust, M. S., Olsson, N., Wagar, L. E., Fast, E., Liu, C. L., Muftuoglu, Y., Sworder, B. J., Diehn, M., Levy, R., Davis, M. M., Elias, J. E., Altman, R. B., and Alizadeh, A. A. (2019) Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* **37**, 1332–1343
60. Ritz, D., Kinzi, J., Neri, D., and Fugmann, T. (2017) Data-independent acquisition of HLA class I peptidomes on the Q exactive mass spectrometer platform. *Proteomics* **17**. <https://doi.org/10.1002/pmic.201700177>
61. Pak, H., Michaux, J., Huber, F., Chong, C., Stevenson, B. J., Müller, M., Coukos, G., and Bassani-Sternberg, M. (2021) Sensitive immunopeptidomics by leveraging available large-scale multi-HLA spectral libraries, data-independent acquisition and MS/MS prediction. *Mol. Cell Proteomics* **20**, 100080
62. Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., Mohr, C., Peltzer, A., Kuchenbecker, L., Sachsenberg, T., Walz, J. S., Stevanović, S., Rammensee, H.-G., and Kohlhauser, O. (2019) MHCquant: Automated and reproducible data analysis for immunopeptidomics. *J. Proteome Res.* **18**, 3876–3884
63. Courcelles, M., Durette, C., Daouda, T., Laverdure, J.-P., Vincent, K., Lemieux, S., Perreault, C., and Thibault, P. (2020) Mapdp: A cloud-based computational platform for immunopeptidomics analyses. *J. Proteome Res.* **19**, 1873–1881
64. Marino, F., Chong, C., Michaux, J., and Bassani-Sternberg, M. (2019) High-throughput, fast, and sensitive immunopeptidomics sample processing for mass spectrometry. *Methods Mol. Biol.* **1913**, 67–79
65. Chong, C., Marino, F., Pak, H.-S., Racle, J., Daniel, R. T., Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017) High-throughput and sensitive immunopeptidomics platform reveals profound IFN $\gamma$ -mediated remodeling of the HLA ligandome. *Mol. Cell Proteomics* **17**, 533–548
66. Bruderer, R., Muntel, J., Müller, S., Bernhardt, O. M., Gandhi, T., Cominetti, O., Macron, C., Carayol, J., Rinner, O., Astrup, A., Saris, W. H. M., Hager, J., Valsesia, A., Dayon, L., and Reiter, L. (2019) Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance. *Mol. Cell Proteomics* **18**, 1242–1254
67. Midha, M. K., Campbell, D. S., Kapil, C., Kusebauch, U., Hoopmann, M. R., Bader, S. L., and Moritz, R. L. (2020) DIALib-QC an assessment tool for spectral libraries in data-independent acquisition proteomics. *Nat. Commun.* **11**, 5251
68. Poulos, R. C., Hains, P. G., Shah, R., Lucas, N., Xavier, D., Manda, S. S., Anees, A., Koh, J. M. S., Mahboob, S., Wittman, M., Williams, S. G., Sykes, E. K., Hecker, M., Dausmann, M., Wouters, M. A., *et al.* (2020) Strategies to enable large-scale proteomics for reproducible research. *Nat. Commun.* **11**, 3793
69. Winter, S. V., Karayel, O., Strauss, M. T., Padmanabhan, S., Surface, M., Merchant, K., Alcalay, R. N., and Mann, M. (2021) Urinary proteome profiling for stratifying patients with familial Parkinson's disease. *EMBO Mol. Med.* **13**, e13257
70. Coscia, F., Doll, S., Bech, J. M., Schweizer, L., Mund, A., Lengyel, E., Lindbjerg, J., Madsen, G. I., Moreira, J. M., and Mann, M. (2020) A streamlined mass spectrometry-based proteomics workflow for large-scale FFPE tissue analysis. *J. Pathol.* **251**, 100–112
71. Shao, W., Pedrioli, P. G. A., Wolski, W., Scurtescu, C., Schmid, E., Vizcaino, J. A., Courcelles, M., Schuster, H., Kowalewski, D., Marino, F., Arlehamn, C. S. L., Vaughan, K., Peters, B., Sette, A., Ottenhoff, T. H. M., *et al.* (2017) The SystemMHC atlas project. *Nucleic Acids Res.* **46**, D1237–D1247
72. Shao, W., Caron, E., Pedrioli, P., and Aebersold, R. (2020) The SystemMHC atlas: A computational pipeline, a website, and a data repository for immunopeptidomic analyses. *Methods Mol. Biol.* **2120**, 173–181
73. Yi, X., Liao, Y., Wen, B., Li, K., Dou, Y., Savage, S. R., and Zhang, B. (2021) caAtlas: an immunopeptidome atlas of human cancer. *iScience* **24**, 103107