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

Kevin A. Kovalchik, Qing Ma, Laura Wessling, Frederic Saab, Jérôme Despault, 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 Bonneil, Mathieu Courcelles, Pierre Thibault, Etienne Caron, Isabelle Sirois

PII: S1535-9476(21)00150-X
DOI: https://doi.org/10.1016/j.mcpro.2021.100178
Reference: MCPRO 100178

To appear in: Molecular & Cellular Proteomics

Received Date: 21 May 2021
Revised Date: 28 October 2021
Accepted Date: 1 November 2021


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 THE AUTHORS. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology.
Sample cohorts → Immunocapture of HLA/MHC ligands → Peptide identification

Rapid analysis of multiple datasets

MhcVizPipe

Data visualization and QC

MhcVizPipe (v0.7.8) - Analysis report

Sample cohorts

Immunocapture of HLA/MHC ligands

Peptide identification

RYPSNLQLF → VYMDWYEKF → RYAPMVAFF

10000 5000 2500 1000 500 250 100 50 1 0.5 0.2 0.1

# Peptides

BF Score

Journal Pre-proof
MhcVizPipe: A Quality Control Software for Rapid Assessment of Small-to Large-Scale Immunopeptidome Data Sets

Authors

Kevin A. Kovalchik¹, Qing Ma², Laura Wessling¹, Frederic Saab¹, Jérôme Despault¹, Peter Kubiniok¹, David J. Hamelin¹, Pouya Faridi³, Chen Li³, Anthony W. Purcell³, Anne Jang⁴, Eustache Paramithiotis⁴, Marco Tognetti⁵, Lukas Reiter⁵, Roland Bruderer⁵, Joël Lanoix⁶, Éric Bonneil⁶, Mathieu Courcelles⁶, Pierre Thibault⁶, Etienne Caron¹, Isabelle Sirois¹

Affiliations

¹CHU Sainte-Justine Research Center, Montreal, QC H3T 1C5, Canada
²School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, ON K1N 6N5, Canada
³Infection and Immunity Program and Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia
⁴CellCarta, Montreal, QC H2X 3Y7, Canada
⁵Biognosys, Wagistrasse 21, 8952 Schlieren, Switzerland
⁶Institute of Research in Immunology and Cancer, Montreal, QC H3T 1J4, Canada
⁷Department of Chemistry, Université de Montréal, Montreal, QC H3T 1J4, Canada
⁸Department of Pathology and Cellular Biology, Faculty of Medicine, Université de Montréal, QC H3T 1J4, Canada

*Corresponding authors: Etienne Caron (etienne.caron@umontreal.ca) and Isabelle Sirois (isabelle.sirois.phd@gmail.com)
List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BA</td>
<td>Binding affinity</td>
</tr>
<tr>
<td>BF</td>
<td>Binding fraction</td>
</tr>
<tr>
<td>DDA</td>
<td>Data-dependent acquisition</td>
</tr>
<tr>
<td>DIA</td>
<td>Data-independent acquisition</td>
</tr>
<tr>
<td>EL</td>
<td>Eluted ligand</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical user interface</td>
</tr>
<tr>
<td>HLA</td>
<td>Human leukocyte antigen</td>
</tr>
<tr>
<td>LF</td>
<td>Length fraction</td>
</tr>
<tr>
<td>MHC</td>
<td>Major histocompatibility complex</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>MVP</td>
<td>MhcVizPipe</td>
</tr>
<tr>
<td>QA</td>
<td>Quality assurance</td>
</tr>
<tr>
<td>QC</td>
<td>Quality control</td>
</tr>
</tbody>
</table>
ABSTRACT

Mass spectrometry (MS)-based immunopeptidomics is maturing into an automatized, high-throughput technology, producing small- to large-scale datasets of clinically relevant MHC class I- and II-associated peptides. Consequently, the development of quality control (QC) and quality assurance (QA) 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 semi-automated QC software tool that enables rapid and simultaneous assessment of multiple MHC class I and II immunopeptidomic 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 immunopeptidomic 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 immunopeptidomic laboratory and proteomics core facility that provides immunopeptidomic services to the community.
INTRODUCTION

The importance of mass spectrometry (MS)-based immunopeptidomics 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 immunopeptidome project (13–15). In fact, the growing interest for clinical immunopeptidomics 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 (TSA) in various cancer types (17–22). In its simplest form, clinical immunopeptidomics involves the isolation of HLA class I- 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 immunopeptidomic data generated by MS is of utmost importance. In genomics and MS-based proteomics, the importance of quality control (QC) and quality assurance (QA) 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 immunopeptidomics, QC samples and QC software tools specialized for MHC class I- and II-associated peptides 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. (2020) describes 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 the limit of detection of the HLA ligands of interest, and v) the technical and biological reproducibility (37). 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. (2021) presents 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 fragmentations 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 (38). Akin to MS-based proteomics, such QC approaches are useful to show the overall quality of immunopeptidomic datasets in additional 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, the length of the detected peptides, the number of strong, weak and non-MHC binders per sample, the number of strong, weak and non-MHC binders per allele per sample, the number of peptides making up each sequence motif and the 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 semi-automated 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, 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 above (NetMHC suite tools and GibbsCluster) and, once installed, provides a semi-automated and fast post-data acquisition system to assess and control sample quality and MHC-specificity through effective visualization of one or multiple immunopeptidomic datasets in a HTML report (Figure 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 is also available for large-scale / ‘batch’ analyses (https://github.com/CaronLab/MhcVizPipe/wiki/Command-line-interface). Below, we describe i) the installation procedure of MVP, ii) the MVP graphical user interface 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- 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.

**METHODS**

**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 GibbsCluster. 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 CPU requirements for MVP. However, because MVP utilizes multi-threading, performance increases on systems with higher numbers of CPUs. 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 GibbsCluster with Python visualization libraries to create portable HTML reports for interpretation of immunopeptidomics MS data. In addition to NetMHCpan, NetMHCIIpan, and GibbsCluster, it makes use of the following third-party Python libraries: Plotly, PlotlyDash, Dash Bootstrap Components, Pandas, Numpy, Dominate, UpsetPlotly, Waitress and PlotlyLogo (PlotlyLogo and
UpsetPlotly developed in-house and describe below). The graphical user interface of MVP is built using the PlotlyDash library and runs as a local web application (i.e. it runs in a web browser).

**Peptide list pre-processing**

Prior to analysis, peptide lists are stripped of chemical modifications (e.g. oxidation of methionine and carbamidomethylation), any peptides containing non-standard amino acids are removed, and flanking amino acids are removed if found (e.g. P and V in P.KAPDNRETL.V). It is assumed 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 ≤ 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 percent rank) used to annotate the peptides as follows: for class I, strong binders rank is ≤ 0.5, weak binders rank is ≤ 2.0 and non-binders rank is > 2.0; for class II, strong binders rank is ≤ 2.0, weak binders rank is ≤ 10 and non-binders rank is > 10. The results are presented in tabular format as well as with a bar
plot and heat map. 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 a strong binder). The heatmap is sorted by percent rank eluted ligand score from left-to-right (i.e. the peptides are ordered first by the left-most column, then the second-left-most, etc.). 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 strong binders, blue approximately represents weak binders, and yellow represents non binders with a linear gradient between the colors.

Length fraction and binding fraction score

The length fraction (LF) score is the fraction of all peptides which are 8-14 mers in length for class I peptides, or 8-25 mers in length for class II peptides. The binding fraction (BF) score is the fraction of all peptides within the above length range which are predicted to be either weak or strong binders 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 above. 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 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 strong or weak binders for the respective allele. An additional subset is created in which
peptides predicted to be non-binders for all the alleles are combined (i.e. the peptides in this set are
not predicted binders for any allele). Any subset containing less that 20 peptides is discarded. Each
of the remaining subsets is run in GibbsCluster using the above parameters with -g set to 1, forcing
GibbsCluster to look for only one peptide group. An exception is the subset of non-binders, in
which -g is set to 1-5, allowing GibbsCluster to look for multiple groups in these unannotated
peptides. As above, 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 5 motifs for the non-binding peptides. All the peptide groups shown
in the “Sequence Motifs” section are visualized using the PlotlyLogo library as described below.

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, etc. With a single sample this doesn’t 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](https://github.com/kevinkovalchik/UpSetPlotly) and it can be installed as 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](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 v0.0.2) 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, 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 I clustering and pseudo count 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 BA in annotating peptide for specific MHC alleles, the following datasets were used: H2-K^b and H2-D^b class I peptides extracted from mouse liver tissue (PXD008733) (52), HLA-A and -B class I peptides extracted from PBMC (PXD001872) (39), H2-IA^d and H2-IE^d class II peptides extracted from the mouse A20 cell line (53), and HLA-DQA10101-DQB10501 and HLA-DQA10103-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 from multiple PBMC samples (PXD001872) (39), and H2-K^b/H2-D^b class I peptides extracted from various mouse tissues (PXD008733) (52). Supplementary materials in Rijensky et al. (55) and Shraibman et al. (46) was used to create Table 1, Figure 3 and Figure 4.

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 4 cores and 16GB 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 a 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 binding affinity 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 command line interface (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.

**Code availability**

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).
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 which 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 MhcVizPipe directory (Figure 1B). MVP can then be run by double-clicking the MhcVizPipe 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 MhcVizPipe 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 above. MVP can be installed in an existing Python environment (≥3.7) from the Python Package Index by invoking “pip install MhcVizPipe” from the terminal. As above, 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 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.
The MVP graphical user interface

MVP provides a simple and intuitive graphical user interface (GUI) in any web browser (tested in Firefox, Safari, Chrome, Chromium and Edge) (Figure 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 multi-column file is opened (e.g. database search results in .csv format), the user is asked by MVP to select the header of the column which contains the peptide sequences. There is no limit to the number of files which 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” section of the HTML report. The lists are then subset to peptides 8-12 or 9-22 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 eluted ligand
percent rank scores are used to generate the “Annotation Results” and “Binding Heatmaps” sections in the HTML report. Length fraction (LF) and binding fraction (BF) scores are also calculated to provide numerical values regarding the overall quality of the samples. LF score is the fraction of all peptides which are 8-12 mers in length for class I peptides, or 9-22 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 weak or strong binders 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 Elution Ligand (EL) method**

NetMHCpan (v4.1) and NetMHCIIpan (v4.0) are critical components of the MVP software for scoring class I and II peptides, respectively (Figure 1A). The predicted MHC binding affinity computed by these algorithms defines strong binders (%Rank < 0.5), weak binders (2.0 < %Rank > 0.5) or non-binders (%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 for peptide scoring are Eluted Ligand (EL) and Binding Affinity (BA) (41). Here, we compared both
methods using four publicly available immunopeptidomic datasets (see Methods) 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 Figure 1). Indeed, the two methods resulted in very similar numbers of strong and weak binders, both in mouse (Supplementary Figure 1A, B) and human (Supplementary Figure 1C, D). In contrast, for class II peptides, substantial differences were observed between EL and BA, in both mouse (Supplementary Figure 2A, B) and human (Supplementary Figure 2C, D). Specifically, we found that EL resulted in a nearly 4-fold increase in annotated peptides compared to BA. Furthermore, we found that the peptide groups identified by GibbsCluster contained a much higher proportion of strong binders when using EL (up to 75%) versus BA (up to 14%) (Supplementary Figure 2B, 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 graphic user interface 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 Supplementary 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.

In the second section of the HTML report (Annotation Results and Binding Heatmaps), MVP uses the EL scoring and annotation method described above 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 is a way 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 Methods). 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 % of peptides associated with each allele, and ii) the allele-specific cluster tab displays peptide motif(s) generated from strong and weak MHC binders for each allele, as well as motifs from peptides predicted to be non-assigned MHC binders. Together, MVP generates in a few clicks a complete HTML report for assessing the quality and MHC-specificity of immunopeptidomic data generated by MS.

QC analysis of tumor biopsies

The general quality and MHC-specificity of immunopeptidomic data is assessed using LF and BF scores, as mentioned above. 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 immunopeptidomic dataset generated from eight representative tumor biopsies, recently published in 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 mg to 150 mg, and number of peptides detected ranged from 144 peptides (bile duct; cholangiocarinoma) 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 were 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 8-12 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 immunopeptidomic data, with an average combined score (LF and BF) of 0.91. Indeed, the proportion of peptides between 8-12 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 immunopeptidomic 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 8-12 mers predicted as SB or WB (Figure 3C). Finally, biopsies 3 and 4 yielded middle quality immunopeptidomic 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 (Figure 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 immunopeptidomic 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 Methods) (Figure 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 (Figure 2). The datasets that were selected to benchmark the software included a range of different samples per dataset (Figure 4A). By comparing the two approaches, our results show that the measured time estimates varied from ~20 minutes (for 1 sample) to ~400 minutes (for 20 mouse tissue samples) using the conventional human/manual-based approach. Notably, the time estimates for the exact same datasets varied from 1 minute (for 1 sample) to 18 minutes (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 (Figure 4A). Thus, MVP represents a QC software package in MS-based immunopeptidomics and provides unprecedented speed for fast visualization and quality assessment of immunopeptidomic data. Given its high-speed performance, MVP should find utility in QC analysis of large immunopeptidome sample cohorts.

QC analysis of a large immunopeptidomic sample cohort
To assess the performance of MVP to run ‘batch’ analysis, we selected a large immunopeptidomic dataset recently published in Shraibman et al. (46). In this study, Shraibman et al. report the analysis of the HLA immunopeptidome of 152 samples, including i) plasma-soluble HLA molecules (sHLA) of 142 plasma samples from both glioblastoma and non-cancerous patients, and ii) membranal HLA (mHLA) of 10 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 (46). The ‘batch analysis’ of the 152 samples was run automatically using the command line interface and a simple Python script (https://github.com/CaronLab/MhcVizPipe/wiki/Command-line-interface). The run time was approximaetly 81 minutes (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, the LF and the BF scores) and illustrating the variability in these scores between all the samples (Figure 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 (Figure 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, resuling in a relatively low BF score of 0.29 (Figure 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 3,500 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 of 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/she should drop low quality samples or keep troubleshooting them to perform additional experiments and analyzes.

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 non-computer 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 (DIA) methods have been applied (39, 60, 61), specialized computational workflows have been created (62, 63), and new automated and semi-automated 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 (DDA)-MS. However, a long-term goal in immunopeptidomics is to scale to large and quantitative clinical datasets generated by DIA-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, quantitative immunopeptidomic data in large-scale studies is challenging because of issues induced by the samples themselves due to 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—e.g. 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 semi-automated, 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 non-expert 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 of the data generated or analyzed during this study are included in this published article and/or the supplementary materials. Created datasets and code are publicly available. Immunopeptidomic data visualized in Supplemental Data S2 have been deposited to the ProteomeXchange Consortium (http://proteomcentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD028633. All the codes are available at github: https://github.com/CaronLab/MhcVizPipe.

ACKNOWLEDGMENTS
This work was supported by funding from the Fonds de recherche du Québec – Santé (FRQS), the Cole Foundation, CHU Sainte-Justine and the Charles-Bruneau Foundations, Canada Foundation for Innovation, the National Sciences and Engineering Research Council (NSERC) (#RGPIN-2020-05232) and the Canadian Institutes of Health Research (CIHR) (#174924). IRIC proteomics facility is a Genomics Technology platform funded in part by the Canadian Government through Genome Canada. KK is a recipient of IVADO’s postdoctoral scholarship (#4879287150).
REFERENCES


abundance reveals the effect of direct and cross-presentation on influenza CTL responses. *Nat Commun* 10, 2846


865
Figure 1. Overview and installation of the MhcVizPipe (MVP) software. **(A)** Illustration showing the experimental workflow for the isolation and identification of MHC-associated peptides by LC-MS/MS (i), the software tools integrated into the MVP pipeline for rapid data processing (ii), and the HTML report generated to support (or not) downstream immunopeptidomic data interpretation through quality control (QC) and quality assurance (QA) (iii). 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 http://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 (i) and the installation of MVP (ii).

Figure 2. Overview of the 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.

Figure 3. Visualization of the QC scores (LF & BF) calculated in Table 1. **(A,B)** Histograms and heatmaps illustrating high quality (A) and middle to low quality (B) immunopeptidomic data.
generated by MS from various tumor biopsies (see Table 1). Histograms showing the distribution of peptides according to their length (right panel) and predicted MHC binding affinity (left panel). SB: Strong Binder, WB: Weak Binder, NB: Non-Binder. Heatmaps are automatically generated by MVP and are 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.

Figure 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).Datasets tested were: (from left to right) 1 QC sample (JY), 1 PBMC sample, 15 PBMC samples, 5 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.

Table 1. Length Fraction (LF) and Binding Fraction (BF) scores calculated from immunopeptidomic data generated from eight different tumor biopsies. Immunopeptidomic data were obtained from Rijensky et al. (2020). Eight different biopsies and cancer types were analyzed. The sample weight, the total number of peptides identified, the fraction of peptides
between 8 and 12 mers and the 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.
Table 1. Length Fraction (LF) and Binding Fraction (BF) scores calculated from immunopeptidomic data generated from eight different tumor biopsies. Immunopeptidomic data were obtained from Rijensky et al. (2020). Eight different biopsies and cancer types were analyzed. The sample weight, the total number of peptides identified, the fraction of peptides between 8 and 12 mers and the 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.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Cancer type</th>
<th>Sample weight (mg)</th>
<th>Total peptides</th>
<th>Peptides (8-12 mers)</th>
<th>LF Score</th>
<th>BF Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biopsy 1</td>
<td>Head and neck (adnexal adenocarcinoma)</td>
<td>&lt;13</td>
<td>352</td>
<td>284</td>
<td>0.81</td>
<td>0.62</td>
</tr>
<tr>
<td>Biopsy 2</td>
<td>Bile duct (Cholangiocarcinoma)</td>
<td>60</td>
<td>144</td>
<td>68</td>
<td>0.47</td>
<td>0.07</td>
</tr>
<tr>
<td>Biopsy 3</td>
<td>Lung (N/A)</td>
<td>120</td>
<td>486</td>
<td>207</td>
<td>0.43</td>
<td>0.44</td>
</tr>
<tr>
<td>Biopsy 4</td>
<td>Gastric (Carcinoma)</td>
<td>100</td>
<td>969</td>
<td>452</td>
<td>0.47</td>
<td>0.32</td>
</tr>
<tr>
<td>Biopsy 5</td>
<td>Vascular (Hemangioendothelioma)</td>
<td>&lt;13</td>
<td>691</td>
<td>585</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>Biopsy 6</td>
<td>Head and neck (Squamous cell carcinoma)</td>
<td>&lt;13</td>
<td>1018</td>
<td>969</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>Biopsy 7</td>
<td>Bladder (Sarcomatoid carcinoma)</td>
<td>&lt;13</td>
<td>391</td>
<td>323</td>
<td>0.83</td>
<td>0.92</td>
</tr>
<tr>
<td>Biopsy 8</td>
<td>Pancreatic (Adenocarcinoma)</td>
<td>150</td>
<td>981</td>
<td>897</td>
<td>0.7</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Figure 2

MhcVizPipe

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

Click here for help and resources

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)

Loaded data:

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Description (optional)</th>
<th>Alleles (required)</th>
</tr>
</thead>
<tbody>
<tr>
<td>biopsy_F2</td>
<td></td>
<td>HLA-A2402, HLA-B5801, HLA-C0701</td>
</tr>
<tr>
<td>biopsy_F4</td>
<td></td>
<td>HLA-A2301, HLA-A9201, HLA-B4403, HLA-B3801, HLA-C0401, HLA-C1201</td>
</tr>
</tbody>
</table>

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!
A  High quality

B  Middle to low quality
Figure 4

A

(1) (20) 1 QC sample
(2) (29) 1 PBMC sample
(15) (311) 15 PBMC samples
(3) (100) 5 tissue sample
(18) (400) 20 tissue samples

Time (min)
- Manual (human)
- MVP (computer)

B

Nbr of peptides

Non-cancerous samples

BF Score
LF Score
N Peptides

BF & LF score

C

BCN_13-Plasma_A

LF = 0.98
BF = 0.29

Nbr of peptides

Peptide length

%Rank

Nbr of peptides
Highlights

- MhcVizPipe is a software tool for quality control (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
AUTHOR CONTRIBUTIONS

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 semi-automated, GUI-based QC software tool that enables ultra-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.