Data in Brief 7 (2016) 201–205

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A first dataset toward a standardized community-driven global mapping of the human immunopeptidome

Pouya Faridi [a], Ruedi Aebersold [b,c], Etienne Caron [b]

[a] Department of Phytopharmaceuticals, School of Pharmacy and Pharmaceutical Sciences Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[b] Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
[c] Faculty of Science, University of Zurich, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 5 October 2015
Received in revised form
16 October 2015
Accepted 4 February 2016
Available online 12 February 2016

Keywords:
Immunology
HLA
Peptidomics
Mass spectrometry

ABSTRACT

We present the first standardized HLA peptidomics dataset generated by the immunopeptidomics community. The dataset is composed of native HLA class I peptides as well as synthetic HLA class II peptides that were acquired in data-dependent acquisition mode using multiple types of mass spectrometers. All laboratories used the spiked-in landmark iRT peptides for retention time normalization and data analysis. The mass spectrometric data were deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD001872. The generated data were used to build HLA allele-specific peptide spectral and assay libraries, which were stored in the SWATHAtlas database. Data presented here are described in more detail in the original *eLife* article entitled ‘An open-source computational and data resource to analyze digital maps of immunopeptidomes’.

E-mail address: caron@imsb.biol.ethz.ch (E. Caron).

http://dx.doi.org/10.1016/j.dib.2016.02.016

Specifications table

Subject area	*Biology*
More specific subject area	*Immunology, Peptidomics*
Type of data	*HLA peptidomics, Tandem mass spectrometry*
How data was acquired	*Mass spectrometry: TripleTOF 5600+ (AB Sciex), LTQ-Orbitrap ELITE and LTQ-Orbitrap XL (Thermo Scientific)*
Data format	*Raw/wiff, centroid mzXML, pepXML, peptide spectral libraries (SpectraST format), assay libraries (CSV, TraML)*
Experimental factors	*N/A*
Experimental features	*Native HLA class I peptide complexes were enriched by immunoaffinity purification. The isolated peptides were used for mass spectrometry analysis using data-dependent acquisition (DDA).*
Data source location	*1- Monash University; Australia 2- University of Oxford; United Kingdom 3-Spanish National Biotechnology Center; Spain 4- ETH-Zurich, Switzerland*
Data accessibility	• Mass spectrometry discovery peptidomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD001872. • HLA allele-specific peptide spectral libraries (SpectraST format) and assay libraries (CSV, TraML) have been deposited to the SWATHAtlas database.

Value of the data

- Public repository of standardized HLA allele-specific peptide assay libraries generated by the immunopeptidomics community.
- The public repository serves as an initial framework to further collect, store and share immunopeptidomics data.
- The public repository can be used by the community to extract highly reproducible quantitative information from HLA peptidomics data acquired in data-independent acquisition (DIA)/SWATH mode.

1. Data, experimental design, materials and methods

1.1. Sample preparation

In this study, native HLA class I peptides and synthetic HLA class II peptides were analyzed. Native HLA class I peptides were isolated from primary cells (i.e. PBMCs) as well as cell lines (i.e. JY, Jurkat and C1R). Synthetic HLA class II peptides were obtained from a collection of 20,176 MTB peptides that were synthesized by Mimotopes (Victoria, Australia) [1]. PBMCs from healthy donors were isolated by density gradient centrifugation. HLA typing was carried out by the Department of Hematology and Oncology, Tübingen, Germany. JY, Jurkat and C1R cells were cultured in RPMI supplemented with 10% fetal bovine serum, 50 IU/mL penicillin, and 50 μg/mL streptomycin (Invitrogen, Life Technologies Europe BV, Zug, Switzerland). C1R cells were stably transfected with -B2705, -B3901 and -B4002 constructs. Native HLA class I peptide complexes were isolated by immunoaffinity purification. In brief, snap-frozen cell pellets were lysed in 10 mM CHAPS/PBS (3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate/phosphate-buffered saline) (AppliChem, St. Louis, MO, USA) containing 1 × complete protease inhibitor

(Roche, Basel, Switzerland). HLA molecules were single-step purified using the pan-HLA class I-specific mAb W6/32, covalently linked to CNBr-activated sepharose (GE Healthcare, Chalfont St. Giles, UK). HLA–peptide complexes were eluted by repeated addition of 0.2% trifluoroacetic acid. Elution fractions E1–E8 were pooled and free HLA ligands were isolated by ultrafiltration using Amicon centrifugal filter units (Millipore, Billerica, MA, USA). HLA ligands were extracted and desalted from the filtrate using Milipore ZipTip C18 pipette tips (Millipore, Billerica, MA, USA). Extracted peptides were eluted in 35 µl of 80% acetonitrile/0.2% trifluoroacetic acid, centrifuged to complete dryness and resuspended in 25 µl of 1% acetonitrile/0.05% trifluoroacetic acid. Samples were stored at −20 °C until analysis by LC–MS/MS. For RT normalization and analysis, iRT peptides (Biognosys AG, Schlieren, Switzerland) were added to samples prior to MS injection [2].

1.2. DDA mass spectrometry

Both naturally presented and synthetic HLA peptides were analyzed on an Eksigent nanoLC system coupled with a SWATH-MS-enabled TripleTOF 5600+ System. The mass spectrometer was operated in DDA top20 mode, with 500 and 150 ms acquisition time for the MS1 and MS2 scans respectively, and 20 s dynamic exclusion. Rolling collision energy with a collision energy spread of 15 eV was used for fragmentation.

Mtb synthetic peptides were analyzed on an Eksigent LC system coupled to an LTQ-Orbitrap ELITE mass spectrometer. Full mass spectra were acquired with the Orbitrap analyser operated at a resolving power of 30,000 (at *m*/*z* 400). Mass calibration used an internal lock mass (protonated (Si $(CH_3)_2O$))6; *m*/*z* 445.120029) and mass accuracy of peptide measurements was within 5 ppm. MS/MS spectra were acquired in CID and HCD mode with a normalized collision energy of 35%. Up to ten precursor ions were accumulated to a target value of 50,000 with a maximum injection time of 300 ms and fragment ions were transferred to the Orbitrap analyser operating at a resolution of 15,000 at *m*/*z* 400.

Naturally presented HLA class I peptides from PBMC samples were analyzed by reversed-phase liquid chromatography coupled with an LTQ-Orbitrap XL hybrid mass spectrometer. Samples were analyzed in five technical replicates. Sample volumes of 5 µl (sample shares of 20%) were injected onto a 75 µmx2 cm trapping column (Acclaim PepMap RSLC; Thermo Fisher) at 4 µl/min for 5.75 min. Eluting peptides were ionized by nanospray ionization and analyzed in the mass spectrometer implementing a top five CID method generating fragment spectra for the five most abundant precursor ions in the survey scans. Resolution was set to 60,000. For HLA class I ligands, the mass range was limited to 400–650*m*/*z* with charge states 2 and 3 permitted for fragmentation.

1.3. Database search engines and statistical validation

All raw instrument data were centroided and processed as described previously [3,4]. The datasets were searched individually using X!tandem [5], MS-GF+ [6] and Comet [7] against the full non-redundant, canonical human genome as annotated by the UniProtKB/Swiss-Prot (2014_02) with 20,270 ORFs and appended iRT peptide and decoy sequence. Oxidation (*M*) was the only variable modification. Parent mass error was set to ±5 ppm, fragment mass error was set to ±0.5 Da. The search identifications were then combined and statistically scored using PeptideProphet and iProphet within the TPP (4.7.0) [8·, 9]. All peptides with an iProbability/iProphet score above 0.7 were exported in Excel. Assumed charges were also exported, as this information is needed in SpectraST [10]. Length considered was 8−12 residues for class I HLA peptides. FDR was manually estimated based on the target-decoy approach [11]. Peptides (1% and 5% peptide-level FDR) were then exported to a.txt file for annotation to their respective HLA allele.

1.4. HLA allele annotation using python and R scripts

Annotation of the identified peptides (1% and 5% peptide-level FDR) to their respective HLA allele was performed automatically by integrating the stand-alone software package of NetMHC 3.4 with our in-house software tools [12,13]. Our in-house software tools include python and R scripts, which are described and available in Supplementary file 1 (see http://elifesciences.org/content/4/e07661/DC8) and Source Code 1 (see http://elifesciences.org/content/4/e07661/DC9) of the original eLife paper [13]. In brief, the in-house software tools perform three main tasks: (1) prediction of HLA binding affinities for individual peptides, (2) calculation of annotation scores, and (3) clustering analysis and data visualization. To perform these tasks, we first used the python script 'matrix_netMHC_MHCI.py' from a list of identified peptides and we predict their binding affinities to several pre-defined sample-specific HLA alleles. The length of the peptides has to be between 8 and 12 amino acids, otherwise they will be filtered out. The output matrix table generated (e.g. peptides_test_matrix_netMHC.txt) is then pipelined through the R script 'matrix_netMHC_analysis_batch.R', which calls functions from two other R scripts: 'allele_distributions.R' and 'sequence_motiv.R'. These R scripts score and annotate peptides, cluster data, and generate heatmaps for data visualization. The annotation score is calculated as the fold change between the best and the second best HLA binding affinity value. The higher the annotation score is, the more confident the annotation of a peptide to an allele will be. A predefined cutoff score of 3 was used in this study to annotate each peptide to their respective HLA allele. A cutoff value of 3 was selected because > 90% of the identified peptides with an annotation score above 3 have a predicted IC50 below 1000 nM. Finally, the lists of annotated HLA-allele specific peptides were exported into a.txt file and used in SpectraST for library generation.

1.5. *Generation of standardized HLA allele-specific peptide spectral and assay libraries*

The parameters below were used for Spectrast. Exact meaning of each parameter can be found in the following link: http://tools.proteomecenter.org/wiki/index.php?title=Software:SpectraST.

Spectrast was used in library generation mode with CID-QTOF settings (-cICID-QTOF) for the Triple-TOF 5600+ or CID (default) settings for the Orbitrap-XL and Orbitrap-ELITE. Retention times were normalized against the iRT Kit peptide sequences (-c_IRTiRT.txt -c_IRR) [13,14]. Only HLA-allele specific peptide ions were included for library generation (-cT).

Below are the command lines that were used to build the standardized HLA allele-specific peptide spectral and assay libraries, which were deposited in the SWATHAtlas database (http://www.swathatlas.org/):

- spectrast -cNSpecLib_celltype_allele_fdr_iRT -cICID-QTOF -cTReference_celltype_allele_fdr.txt -cP0.7 -c_IRTiRT.txt -c_IRR iprophet.pep.xml
- A consensus library was then generated: spectrast -cNSpecLib_cons_celltype_allele_fdr_iRT -cICID-QTOF -cAC SpecLib_celltype_allele_fdr_iRT.splib
- Optionally, HLA-allele specific consensus libraries were merged: spectrast -cNSpecLib_cons_celltype_alleles_fdr_iRT -cJU -cAC SpecLib_celltype_allele1_fdr_iRT.splib SpecLib_celltype_allele2_fdr_iRT.splib SpecLib_celltype_allele3_fdr_iRT.splib SpecLib_celltype_allele4_fdr_iRT.splib
- The script spectrast2tsv.py (msproteomicstools 0.2.2; https://pypi.python.org/pypi/msproteomicstools) was then used to generate the HLA-allele specific peptide assay library with the following recommended settings: spectrast2tsv.py -l 350,2000 -s b,y -x 1,2 -o 6 -n 6 -p 0.05 -d -e -w swaths.txt -k openswath -a SpecLib_cons_celltype_alleles_fdr_iRT_openswath.csv SpecLib_cons_celltype_alleles_fdr_iRT.sptxt
- The _openswath.csv file was then converted into a.tsv file and opened in Excel. Reference coordinates for the 11 iRT peptides were confirmed and any remaining decoy sequences were removed. The file was then saved in.txt format and then converted back in.csv format. The OpenSWATH tool ConvertTSVToTraML converted the TSV/CSV file to TraML: ConvertTSVToTraML -in SpecLib_cons_celltype_alleles_fdr_iRT_openswath.csv -out SpecLib_cons_celltype_alleles_fdr_iRT.TraML
- Decoys were appended to the TraML assay library with the OpenSWATH tool OpenSwathDecoyGenerator in reverse mode with a similarity threshold of 0.05 Da and an identity threshold of 1: OpenSwathDecoyGenerator -in SpecLib_cons_celltype_alleles_fdr_iRT.TraML -out SpecLib_cons_celltype_alleles_fdr_iRT_decoy.TraML -method shuffle -append -exclude_similar

Acknowledgement

E.C. is supported by a Marie Curie Intra-European Fellowship (326937). R.A. acknowledges the following grant support: TBVAC2020 (2-73838-14), ERC grant Proteomics v3.0 (ERC-2008-AdG_20080422), ERC Proteomics 4D (670821) and the Swiss National Science Foundation (3100A0-688 107679).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.02.016.

References

[1] C.S. Lindestam Arlehamn, A. Gerasimova, F. Mele, R. Henderson, J. Swann, J.A. Greenbaum, Y. Kim, J. Sidney, E.A. James, R. Taplitz, D.M. McKinney, W.W. Kwok, H. Grey, F. Sallusto, B. Peters, A. Sette, Memory T cells in latent *Mycobacterium tuberculosis* infection are directed against three antigenic islands and largely contained in a CXCR3+CCR6+ Th1 subset, PLoS Pathog. 9 (2013) e1003130.

[2] C. Escher, L. Reiter, B. MacLean, R. Ossola, F. Herzog, J. Chilton, M.J. MacCoss, O. Rinner, Using iRT, a normalized retention time for more targeted measurement of peptides, Proteomics 12 (2012) 1111–1121.

[3] B.C. Collins, L.C. Gillet, G. Rosenberger, H.L. Röst, A. Vichalkovski, M. Gstaiger, R. Aebersold, Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system, Nat. Methods 10 (2013) 1246–1253.

[4] G. Rosenberger, C.C. Koh, T. Guo, H.L. Röst, P. Kouvonen, B.C. Collins, M. Heusel, Y. Liu, E. Caron, A. Vichalkovski, M. Faini, O. T. Schubert, P. Faridi, H.A. Ebhardt, M. Matondo, H. Lam, S.L. Bader, D.S. Campbell, E.W. Deutsch, R.L. Moritz, S. Tate, R. Aebersold, A repository of assays to quantify 10,000 human proteins by SWATH-MS, Sci. Data 1 (2014) 140031.

[5] R. Craig, J.P. Cortens, R.C. Beavis, Open source system for analyzing, validating, and storing protein identification data, J. Proteome Res. 3 (2004) 1234–1242.

[6] S. Kim, P.A. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics, Nat. Commun. 5 (2014) 5277.

[7] J.K. Eng, T.A. Jahan, M.R. Hoopmann, Comet: an open-source MS/MS sequence database search tool, Proteomics 13 (2012) 22–24.

[8] D. Shteynberg, E.W. Deutsch, H. Lam, J.K. Eng, Z. Sun, N. Tasman, L. Mendoza, R.L. Moritz, R. Aebersold, A.I. Nesvizhskii, iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates, Mol. Cell. Proteom. 10 (2011), M111.007690.

[9] D. Shteynberg, A.I. Nesvizhskii, R.L. Moritz, E.W. Deutsch, Combining results of multiple search engines in proteomics, Mol. Cell. Proteom. 12 (2013) 2383–2393.

[10] H. Lam, R. Aebersold, Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics, Nat. Methods 54 (2011) 424–431.

[11] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat. Methods 4 (2007) 207–214.

[12] C. Lundegaard, K. Lamberth, M. Harndahl, S. Buus, O. Lund, M. Nielsen, NetMHC−3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8−11, Nucleic Acids Res. 36 (2008) W509–W512.

[13] E. Caron, L. Espona, D.J. Kowalewski, H. Schuster, N. Ternette, A. Alpízar, R.B. Schittenhelm, S.H. Ramarathinam, C. S. Lindestam Arlehamn, C.C. Koh, L.C. Gillet, A. Rabsteyn, P. Navarro, S. Kim, H. Lam, T. Sturm, M. Marcilla, A. Sette, D. S. Campbell, E.W. Deutsch, R.L. Moritz, A.W. Purcell, H.G. Rammensee, S. Stevanovic, R. Aebersold, An open-source computational and data resource to analyze digital maps of immunopeptidomes, eLife 4 (2015) e07661.

[14] O.T. Schubert, L.C. Gillet, B.C. Collins, P. Navarro, G. Rosenberger, W.E. Wolski, H. Lam, D. Amodei, P. Mallick, B. MacLean, R. Aebersold, Building high-quality assay libraries for targeted analysis of SWATH MS data, Nat. Protoc. 10 (2015) 426–441.