Published online 29 July 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D1237–D1247***
doi: 10.1093/nar/gkx664

The SysteMHC Atlas project

Wenguang Shao[1,†], Patrick G.A. Pedrioli[1,†], Witold Wolski[2], Cristian Scurtescu[3], Emanuel Schmid[3], Juan A. Vizcaíno[4], Mathieu Courcelles[5], Heiko Schuster[6,7], Daniel Kowalewski[6,7], Fabio Marino[8,9,10], Cecilia S.L. Arlehamn[11], Kerrie Vaughan[11], Bjoern Peters[11], Alessandro Sette[11], Tom H.M. Ottenhoff[12], Krista E. Meijgaarden[12], Natalie Nieuwenhuizen[13], Stefan H.E. Kaufmann[13], Ralph Schlapbach[2], John C. Castle[14], Alexey I. Nesvizhskii[15,16], Morten Nielsen[17,18], Eric W. Deutsch[19], David S. Campbell[19], Robert L. Moritz[19], Roman A. Zubarev[20], Anders Jimmy Ytterberg[20,21], Anthony W. Purcell[22], Miguel Marcilla[23], Alberto Paradela[23], Qi Wang[24], Catherine E. Costello[24], Nicola Ternette[25], Peter A. van Veelen[26], Cécile A.C.M. van Els[27], Albert J.R. Heck[9,10], Gustavo A. de Souza[28,29], Ludvig M. Sollid[28], Arie Admon[30], Stefan Stevanovic[6,7], Hans-Georg Rammensee[6,7], Pierre Thibault[5], Claude Perreault[5], Michal Bassani-Sternberg[8], Ruedi Aebersold[1,31,*] and Etienne Caron[1,*]

[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland, [2]Functional Genomics Center Zurich, ETH Zurich and University of Zurich, Zurich 8057, Switzerland, [3]Scientific IT Services (SIS), ETH Zurich, Zurich 8093, Switzerland, [4]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, [5]Institute for Research in Immunology and Cancer, Université de Montréal, Montreal, H3T 1J4, Canada, [6]Department of Immunology, Interfaculty Institute for Cell Biology, University of Tübingen, Tübingen, 72076, Germany, [7]German Cancer Consortium (DKTK), DKFZ partner site Tübingen, Tübingen, 72076, Germany, [8]Ludwig Institute for Cancer Research, University Hospital of Lausanne, Lausanne 1011, Switzerland, [9]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, 3584 CH, The Netherlands, [10]Netherlands Proteomics Centre, Utrecht, 3584 CH, The Netherlands, [11]La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA, [12]Department of Infectious Diseases, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands, [13]Department of Immunology, Max Planck Institute for Infection Biology, Berlin 10117, Germany, [14]Vaccine Research and Translational Medicine, Agenus Switzerland Inc., 4157 Basel, Switzerland, [15]Department of Pathology, BRCF Metabolomics Core, University of Michigan, Ann Arbor, MI 48109, USA, [16]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA, [17]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, 1650, Argentina, [18]Department of Bio and Health Informatics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, [19]Institute for Systems Biology, Seattle, WA 98109, USA, [20]Division of Physiological Chemistry I, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, SE-171 77, Sweden, [21]Rheumatology Unit, Department of Medicine, Solna, Karolinska Institutet, Stockholm, SE-171 77, Sweden, [22]Infection and Immunity Program, Department of Biochemistry and Molecular Biology, Monash Biomedicine Discovery Institute, Monash University, Clayton 3800, Australia, [23]Proteomics Unit, Spanish National Biotechnology Centre, Madrid 28049, Spain, [24]Center for Biomedical Mass Spectrometry, Department of Biochemistry, Boston University School of Medicine, Boston, MA 02118, USA, [25]The Jenner Institute, Target Discovery Institute Mass Spectrometry Laboratory, University of Oxford, Oxford, OX3 7FZ, UK, [26]Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands, [27]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, 3720 BA, The Netherlands, [28]Centre for Immune

[*]To whom correspondence should be addressed. Tel: +41 44 633 26 97; Fax: +41 44 633 10 51; Email: caron@imsb.biol.ethz.ch
Correspondence may also be addressed to Ruedi Aebersold. Tel: +41 44 633 31 70; Fax: +41 44 633 10 51; Email: aebersold@imsb.biol.ethz.ch
[†]These authors contributed equally to the paper as first authors.

Regulation, Department of Immunology, University of Oslo and Oslo University Hospital-Rikshospitalet, Oslo 0372, Norway, [29]The Brain Institute, Universidade Federal do Rio Grande do Norte, 59056–450, Natal-RN, Brazil, [30]Department of Biology, Technion, Israel Institute of Technology, Haifa 3200003, Israel and [31]Faculty of Science, University of Zurich, 8006 Zurich, Switzerland

Received June 13, 2017; Revised July 16, 2017; Editorial Decision July 18, 2017; Accepted July 21, 2017

ABSTRACT

Mass spectrometry (MS)-based immunopeptidomics investigates the repertoire of peptides presented at the cell surface by major histocompatibility complex (MHC) molecules. The broad clinical relevance of MHC-associated peptides, e.g. in precision medicine, provides a strong rationale for the large-scale generation of immunopeptidomic datasets and recent developments in MS-based peptide analysis technologies now support the generation of the required data. Importantly, the availability of diverse immunopeptidomic datasets has resulted in an increasing need to standardize, store and exchange this type of data to enable better collaborations among researchers, to advance the field more efficiently and to establish quality measures required for the meaningful comparison of datasets. Here we present the SysteMHC Atlas (https://systemhcatlas.org), a public database that aims at collecting, organizing, sharing, visualizing and exploring immunopeptidomic data generated by MS. The Atlas includes raw mass spectrometer output files collected from several laboratories around the globe, a catalog of context-specific datasets of MHC class I and class II peptides, standardized MHC allele-specific peptide spectral libraries consisting of consensus spectra calculated from repeat measurements of the same peptide sequence, and links to other proteomics and immunology databases. The SysteMHC Atlas project was created and will be further expanded using a uniform and open computational pipeline that controls the quality of peptide identifications and peptide annotations. Thus, the SysteMHC Atlas disseminates quality controlled immunopeptidomic information to the public domain and serves as a community resource toward the generation of a high-quality comprehensive map of the human immunopeptidome and the support of consistent measurement of immunopeptidomic sample cohorts.

INTRODUCTION

T cells have the ability to eliminate abnormal cells through recognition of short peptides presented at the cell surface by major histocompatibility complex (MHC) molecules (human leukocyte antigen [HLA] molecules in human). In mammals, cells are decorated by thousands of such peptides, which are collectively referred to as the MHC class I and class II immunopeptidome (1–3). The MHC class I immunopeptidome is composed predominantly of peptides of 8–12 amino acids in length that are presented at the surface of virtually any cell- and tissue-type in the body. The MHC class II immunopeptidome is composed of peptides of 10–25 amino acids in length that are mainly found on a subset of professional antigen presenting cells, reviewed in (4,5). The amino acid sequence of those peptides is not random. In fact, individual peptides have the ability to bind MHC molecules via specific anchor residues that define a MHC binding motif (6). Such motifs are generally MHC allele-specific, thereby limiting the pool of peptides that can be presented on the surface of a specific cell for scrutiny by T cells. In humans, this limitation is counteracted by the very high diversity of HLA alleles. In fact, each individual can express up to six different HLA class I allotypes and typically eight different HLA class II allotypes, and more than 16 700 allelic forms are expressed at the human population level (http://www.ebi.ac.uk/ipd/imgt/hla/stats.html; May 2017). Thus, the composition of the human immunopeptidome is tremendously complex (7). Describing and understanding the complexity of the immunopeptidome and its functional implications is a central and fundamental challenge of immunology with important clinical implications in precision medicine (8).

Mass spectrometry (MS) is a powerful unbiased method to explore the composition of the immunopeptidome (9). Following pioneering work by Hans-Georg Rammensee (10) and Donald Hunt (11) in the early 90's, the analytical performance of this technique has rapidly evolved and currently enables the identification of thousands of HLA-associated peptides from a single MS measurement (12–22). Notably, the use of MS techniques to conduct 'immunopeptidomic' studies has become increasingly popular over recent years, thanks to technical advances and breakthroughs in the field of immuno-oncology (23). As a consequence, huge amounts of immunopeptidomic data have been and continue to be generated at significant expense.

Immunopeptidomics is an expanding field driven by a rapidly growing community of researchers and deep technology platforms. In 2015, a Human Immuno-Peptidome Project (HIPP; https://www.hupo.org/Human-Immuno-Peptidome-Project) was created as a new initiative of the Biology/Disease-Human Proteome Project (B/D-HPP)—a program under the umbrella of the Human Proteome Organization (HUPO) (24). The long-term goal of this initiative is to make the robust analysis of immunopeptidomes accessible to any immunologist, clinical investigator and other researchers by the generation and dissemination of new methods/technologies and informational resources (25–27). Participants in this initiative recognized the need

for an open immunopeptidomics repository in which output files of mass spectrometric measurements of immunopeptidome samples would be annotated, stored and shared without restriction. Here, we introduce the SysteMHC Atlas project, the first public repository devoted to MS-based immunopeptidomics. In brief, the SysteMHC Atlas uploads raw immunopeptidomics MS data originally deposited into the PRIDE database along with the metadata associated with the experiment (Figure 1) (28). Each project is labeled with the HIPP tag as a B/D-HPP subproject. Raw MS data are then processed through a uniform computational pipeline for MHC peptide identification, annotation (29) and statistical validation (30,31) (Figure 1B). Lists of MHC peptide ligands as well as sample/context- and allele-specific peptide spectral libraries (32) are generated and presented in the Atlas in a way that they can be searched and browsed by researchers via a web interface. Allele-specific peptide spectral libraries can be further converted into formats that are compatible for uploads into the SWATHAtlas database in order to support immunopeptidomic analyses by SWATH-MS/DIA (Data-Independent Acquisition) methods. Importantly, the SysteMHC Atlas aims to be an open and active repository in which raw MS data can be periodically reprocessed with more advanced informatics tools for peptide identification, statistical validation, HLA peptide annotation and library generation, as these become available to the community—a procedure that has been successfully applied in the field of proteomics to ensure high-quality peptide identification with well-understood false discovery rates (FDR) and quality controls (33). The community is expected to benefit from the SysteMHC Atlas at various levels: (i) basic scientists and clinicians can navigate within a large catalog of high-quality context-specific HLA-associated peptides to gain new insights into the composition of the immunopeptidome, (ii) computational scientists find a rich source of data to develop or test new algorithms for immunopeptidomic analyses and (iii) access to HLA peptide assay spectral libraries facilitates next-generation MS analysis of immunopeptidomes (i.e. SWATH-MS/DIA) (34).

CONTENT AND FUNCTIONALITIES OF THE ATLAS

The first version of the SysteMHC Atlas (February 2017) contains raw and processed MS data derived from 16 published human immunopeptidomics projects/datasets (Figure 2). It also contains information from seven unpublished datasets that were released by the data producers. The number of MS output files per project ranges between 1 and 192 for a total of 1184 raw files. All datasets were generated in data-dependent acquisition (DDA) mode using different instruments and fragmentation methods, including collision-induced dissociation (CID), higher energy collisional dissociation (HCD), electron transfer dissociation (ETD) and electron transfer and higher energy collision dissociation (EThcD) (21). Several laboratories used the spiked-in landmark indexed Retention Time (iRT) peptides for retention time normalization (35,36). Each dataset is labeled with a unique and permanent SYSMHC number. Direct links to PubMed, PRIDE and Immune Epitope Database (IEDB) are also provided if applicable (Figure 2). We briefly describe below the content and functionalities of the SysteMHC Atlas.

A catalog of context- and allele-specific MHC class I and class II peptides

The SysteMHC Atlas contains mainly naturally presented human MHC class I and class II peptides. Natural MHC-associated peptides were extracted by immunoaffinity purification or mild acid elution from cell lines, primary cells and tissues—i.e. peripheral blood mononuclear cells (PBMCs), T cells, B cells, dendritic cells, macrophages, fibroblasts, colon carcinoma, breast cancer and glioblastoma. All biological sources were HLA typed and peptides from 67 HLA class I and 27 HLA class II alleles are represented in the current version of the database (February 2017). A full listing of all the samples and corresponding metadata (i.e. organism, tissue and cell type, culture conditions, disease state, HLA type, antibody used for immunoaffinity purification, LC-MS/MS parameters etc.) can be found next to the raw MS files at the project website.

In May 2017, ∼29.5 million MS/MS spectra were searched using a uniform and well-tested computational pipeline and yielded 250, 768 and 1458, 698 distinct peptides with iProphet probability $P \geq 0.9$ and $P > 0.0$, respectively. After applying strict confidence filters for the identification of class I and class II peptides, 119 073 high-confidence HLA class I peptides (peptide FDR 1%, 8–12 amino acids) were identified and annotated to specific HLA-A, -B or -C alleles using an automated annotation strategy as described (34) (see Supplementary Figure S1 for statistics). For class II molecules, 73 465 high-confidence peptides were identified (peptide FDR 1%, 10–25 amino acids, belonging to groups of two or more peptides with an overlap of at least four amino acids). Of note, the assignment of peptides to specific HLA class II alleles will be considered in the future as soon as robust bioinformatics tools for class II peptide annotation become openly available (26). The high-confidence class I and class II peptides were mapped onto 13, 132 and 7704 of the human UniProtKB/Swiss-Prot proteins, respectively.

An important goal in MS-based immunopeptidomics is to assess the size of the human immunopeptidome at the population level. To answer this question, we plotted the cumulative number of distinct HLA class I peptides as a function of the addition of identified spectra at FDR 1% (Figure 3). Each data point on the curve represents an added experiment, and the experiments are presented in chronological order of data acquisition. When looking at the combined data from all HLA class I alleles in the Atlas, our analysis suggests that for the presently available technology the saturation level might already be reached (Figure 3A). However, this observation might be biased given the limited number of HLA alleles as well as the limited number of cell and tissue types that were sampled until now. In addition, when individual HLA class I alleles were considered, the number of distinct peptides continued to steeply increase for several HLA class I allele such as -A02, -C02 and C16 (Figure 3B) indicating that for these alleles, saturation had not yet been achieved as the curves are expected to reach saturation only

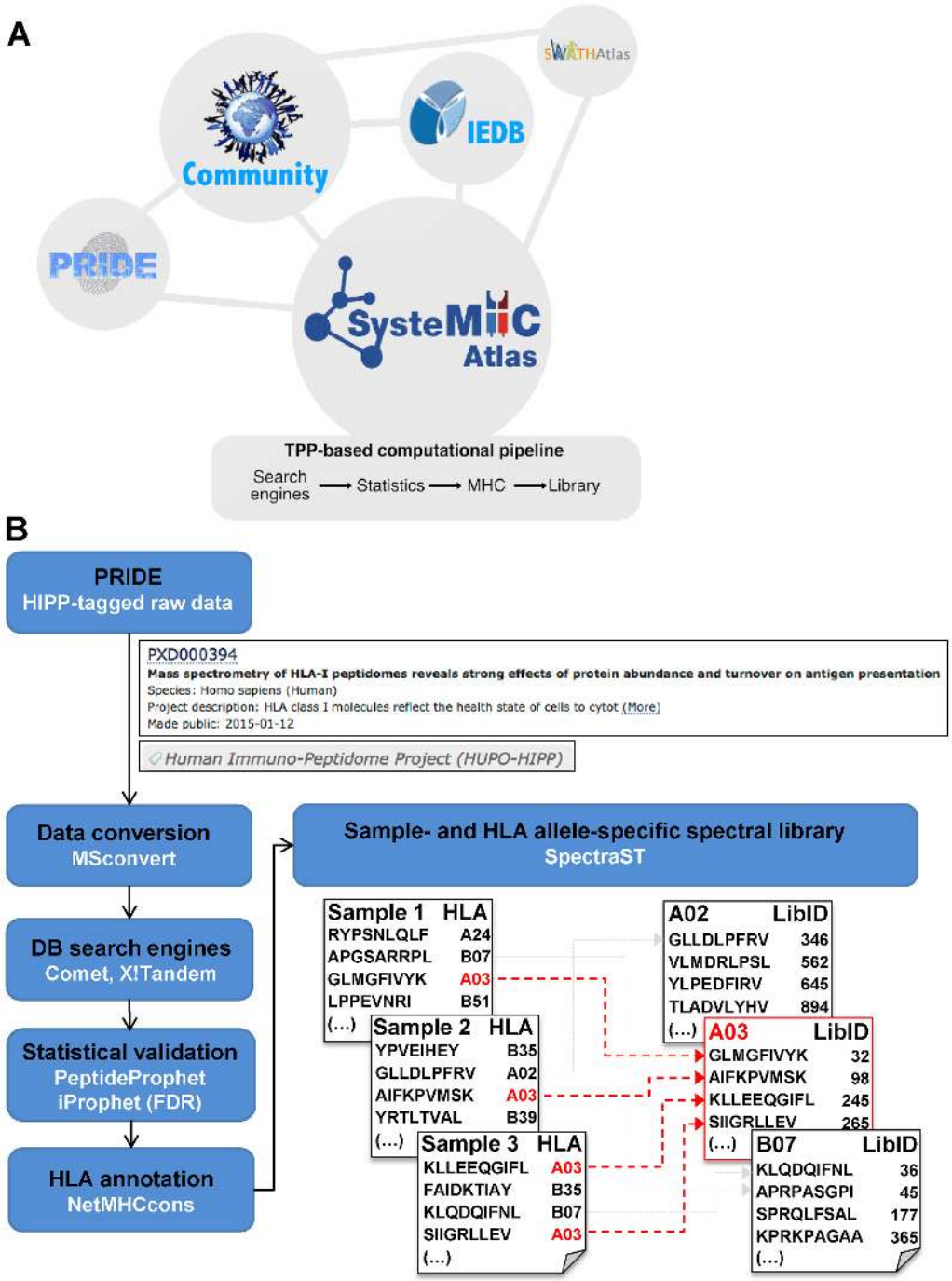

Figure 1. Overview of the SysteMHC Atlas project. (**A**) The SysteMHC Atlas aims to be a long-term data-driven project that serves the community. It is linked to other repositories of proteomic data and consists of two main components: (i) a uniform computational pipeline for processing raw MS files and (ii) a web interface with storage, searching and browsing capabilities. First, shotgun/DDA-MS experimental data generated for specific projects are submitted by the data producers to PRIDE. Raw MS data are then uploaded into the SysteMHC Atlas and processed through a consistent and open computational pipeline (**B**) that controls the quality of peptide identification and peptide annotation to specific HLA alleles. Spectral libraries are generated and can be converted into high-quality HLA allele-specific peptide assay libraries, also available at SWATHAtlas. All the results generated by the computational pipeline are made available to the public domain via the SysteMHC Atlas web-based interface, which provides links to the Immune Epitope Database (IEDB) for accessing lists of peptides originally identified and published by the data producers. (B) Current computational pipeline used for generating the immunopeptidome- and spectral database for different HLA allotypes. MS output files generated from several types of instruments are first converted into mzXML file format and then searched using several open-source database search engines. The resulting peptide identifications are combined and statistically scored using PeptideProphet and iProphet within the Trans-Proteomic Pipeline (TPP) (30,31). The identified peptides are next annotated to their respective HLA allele in a fully automated fashion using the stand-alone software package of NetMHCcons 1.1 (29). Spectral libraries are generated using SpectraST (32). Allele-specific peptide spectral libraries are generated from multiple samples—an example for HLA-A03 is highlighted in red. Each HLA peptide is labeled with a unique and permanent library identifier (LibID). Details regarding the computational pipeline and how the data were processed are available at the SysteMHC Atlas website in the 'ABOUT' section.

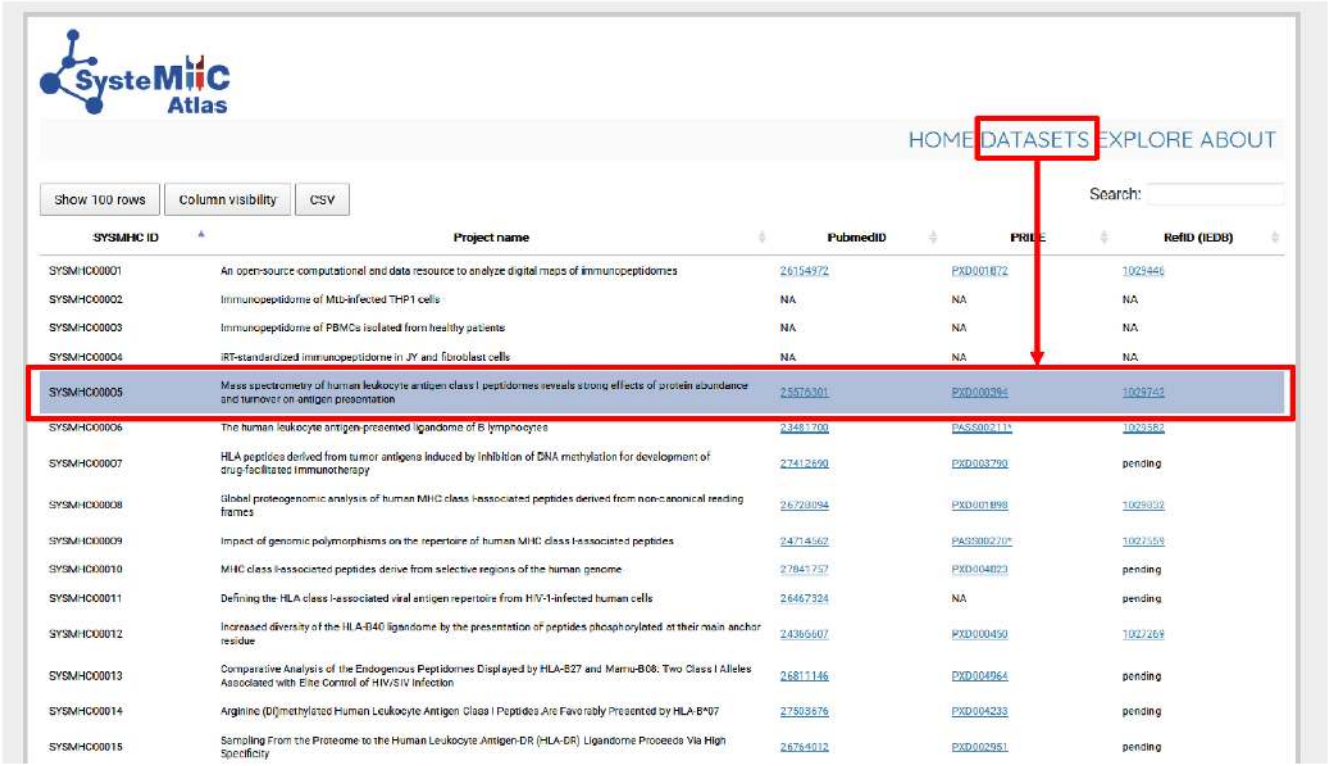

Figure 2. Immunopeptidomics datasets used for building the first version of the SysteMHC Atlas. Data from 23 projects that collectively generated 1184 raw MS files constitute the initial contents of the SysteMHC Atlas. Each project is labeled with a unique SYSMHC identifier and linked to its corresponding PubMed, PRIDE and IEDB ID. For unpublished projects, IDs are not applicable (NA).

when most peptides observable with the applied technology will have been cataloged. Altogether, our current analysis suggests that the Atlas data are not yet comprehensive. In the future, collecting additional MS/MS data from new experiments—including from new HLA alleles, new cell origins, new experimental conditions, new protocols and new MS technologies—will be necessary to properly assess the size and complexity of the human immunopeptidome.

In additional to naturally processed ligands, the SysteMHC Atlas also contains data for synthetic peptides predicted to bind specific HLA alleles. Datasets generated from synthetic peptides might be particularly useful for benchmarking new software tools (37) and to extend the contents of libraries derived from native peptides for targeted analysis of immunopeptidomes (3,38,39). To date, SysteMHC Atlas contains four datasets composed of synthetic peptides: SYSMHC00001 contains data generated from a large collection of synthetic HLA class II peptides encoded by *Mycobacterium tuberculosis* (Mtb) (34,40); SYSMHC00020, SYSMHC00021 and SYSMHC00022 contain data obtained from synthetic HLA class I peptides encoded by Mtb (41,42), Epstein–Barr virus (EBV) and Homo sapiens, respectively.

The SysteMHC Atlas user interface

Researchers can browse, search and download information using query interfaces available at the website (https://systemhcatlas.org). In particular, the 'EXPLORE' link leads to a page where immunopeptidomic data are searchable on numerous levels, including peptide sequence, source protein, as well as HLA class and type. For instance, the user can query the data to specifically identify (i) all class I peptides derived from a specific source protein (e.g. BIRC6 in Figure 4), (ii) the repertoire of peptides presented by a specific HLA type and/or (iii) in which tissues or experimental conditions have specific peptides been observed etc. Thus, the SysteMHC user interface enables large immunopeptidomics datasets to be explored in a user-specifiable fashion.

An important function of the SysteMHC Atlas is to serve as a repository devoted to immunopeptidomics MS-related data at several levels of processing. Specifically, we provide raw and converted mzXML files, iProphet results and HLA peptide spectral libraries, all available for download at the website (Figure 5). In the current version of the Atlas, a total of 539 sample/context- and 37 HLA allele-specific peptide spectral libraries were made available and can be visualized using the respective links from the web interface. Three new allele-specific spectral libraries (i.e. HLA-B15, -C03 and -C07) were also converted into TraML files for SWATH-MS/DIA analysis of immunopeptidomes, as previously described (34,36). These standardized libraries contained the iRT peptides for retention time normalization and data analysis. TraML files are directly available for download at SWATHAtlas.

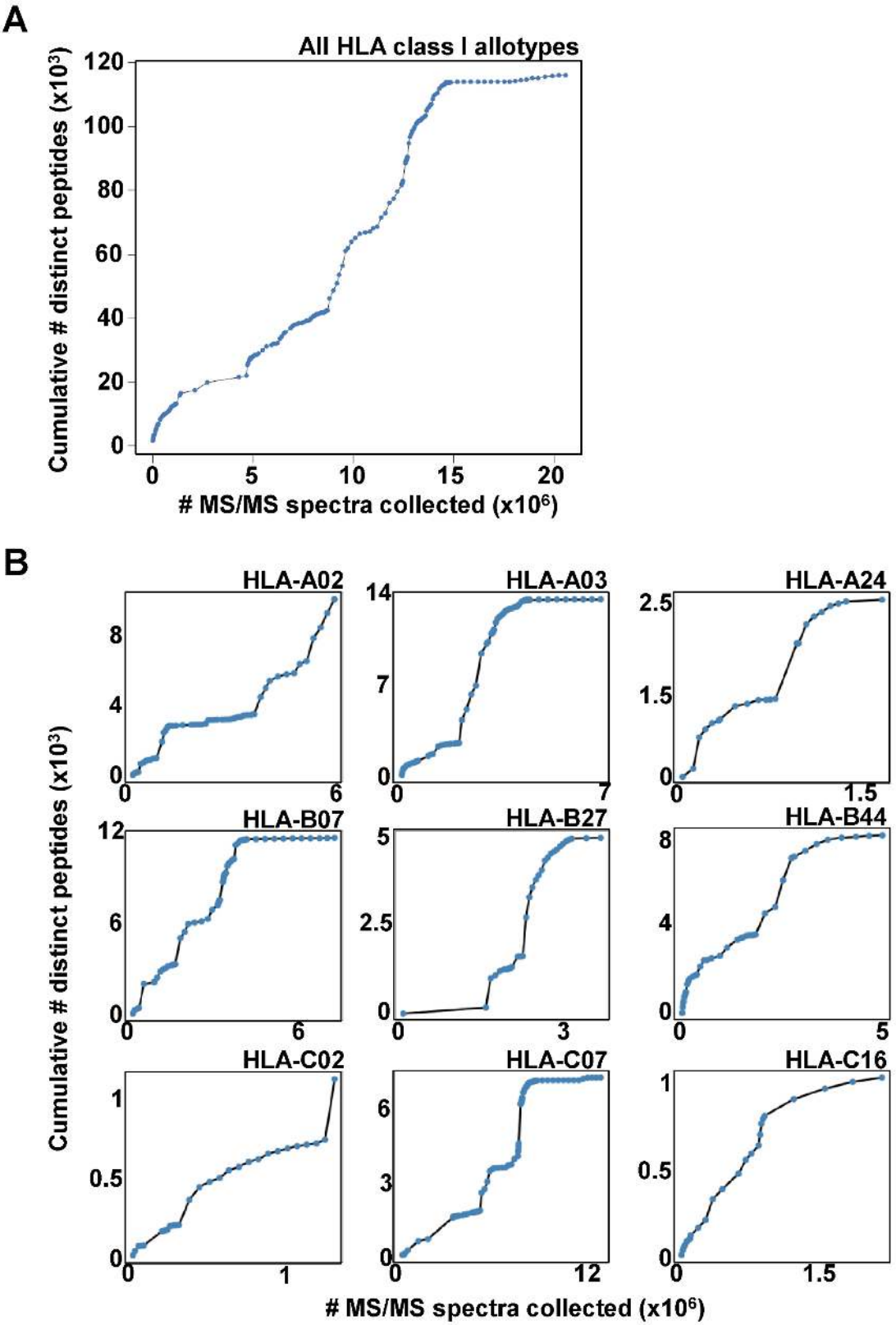

Figure 3. Cumulative number of MS/MS spectra versus cumulative number of distinct peptides for HLA class I alleles at FDR 1%. (**A**) All HLA class I peptides were combined. (**B**) HLA class I alleles that were frequently found in various datasets. Eventually, the curves are expected to reach saturation when most observable peptides will have been cataloged at 1% peptide FDR.

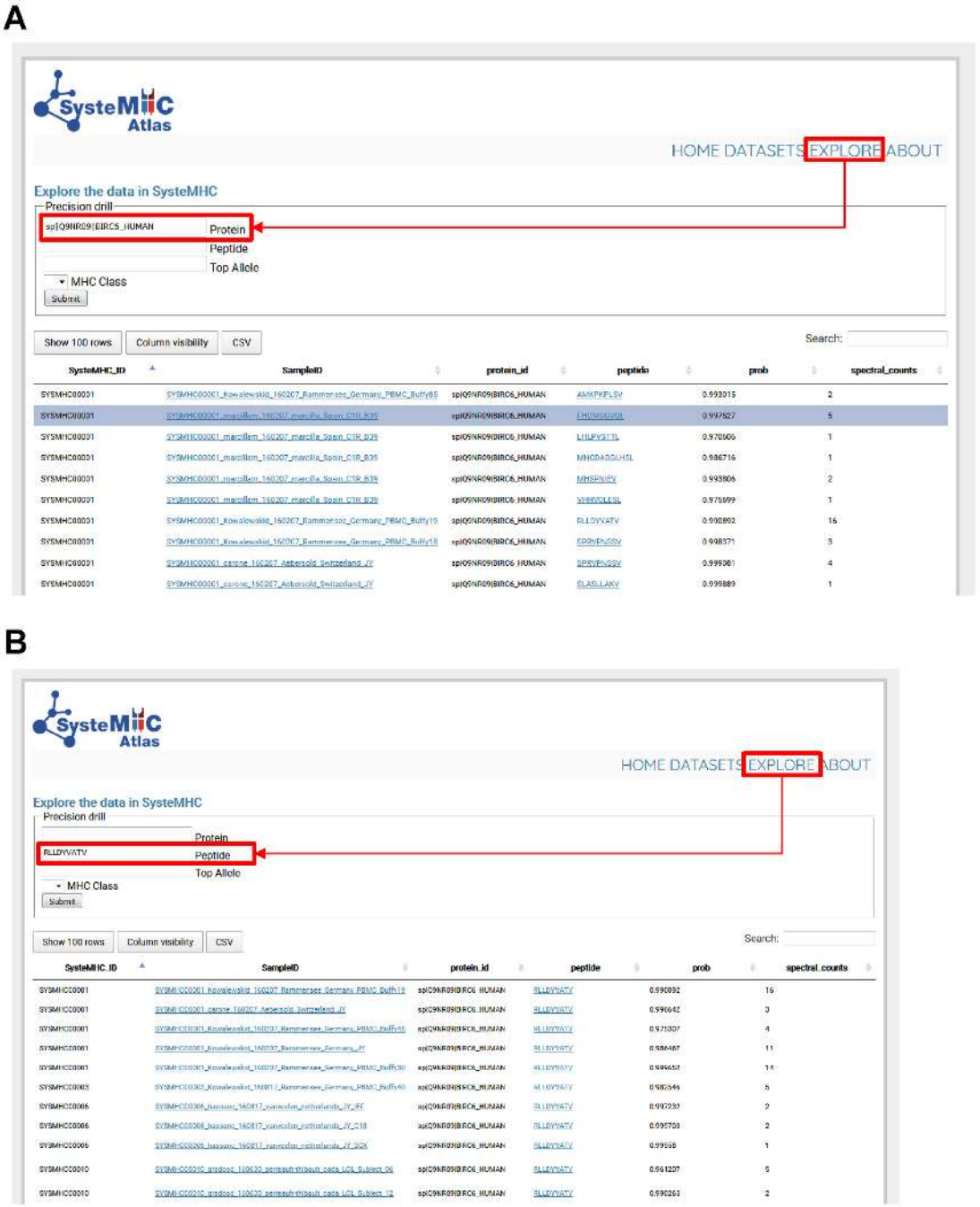

Figure 4. Explore page in the SysteMHC Atlas web-based interface. HLA allele-specific peptide spectral libraries can be downloaded here. The web interface can also be used to query the SysteMHC Atlas and find specific information. (**A**) As an example the source protein BIRC6 was searched and the Atlas returned back all HLA-associated peptides originating from this protein as well as the context (i.e. SysteMHC ID, Sample ID, iProphet score, HLA annotation score, spectral counts, assigned HLA type and class) in which this peptide was observed. Then, the user can click on a specific Sample ID hyperlink and be redirected to the corresponding raw MS files and metadata (e.g. tissue type, cell type, culture condition, purification method, antibody used, mass spectrometer used etc). (**B**) The peptide RLLDYVATV was searched and the Atlas returned back the datasets in which this peptide was observed. By clicking on the peptide sequence hyperlink, the user is redirected to a new page in which the LibID information is available for MS/MS spectra visualization. Information can be downloaded as .csv files for further analysis.

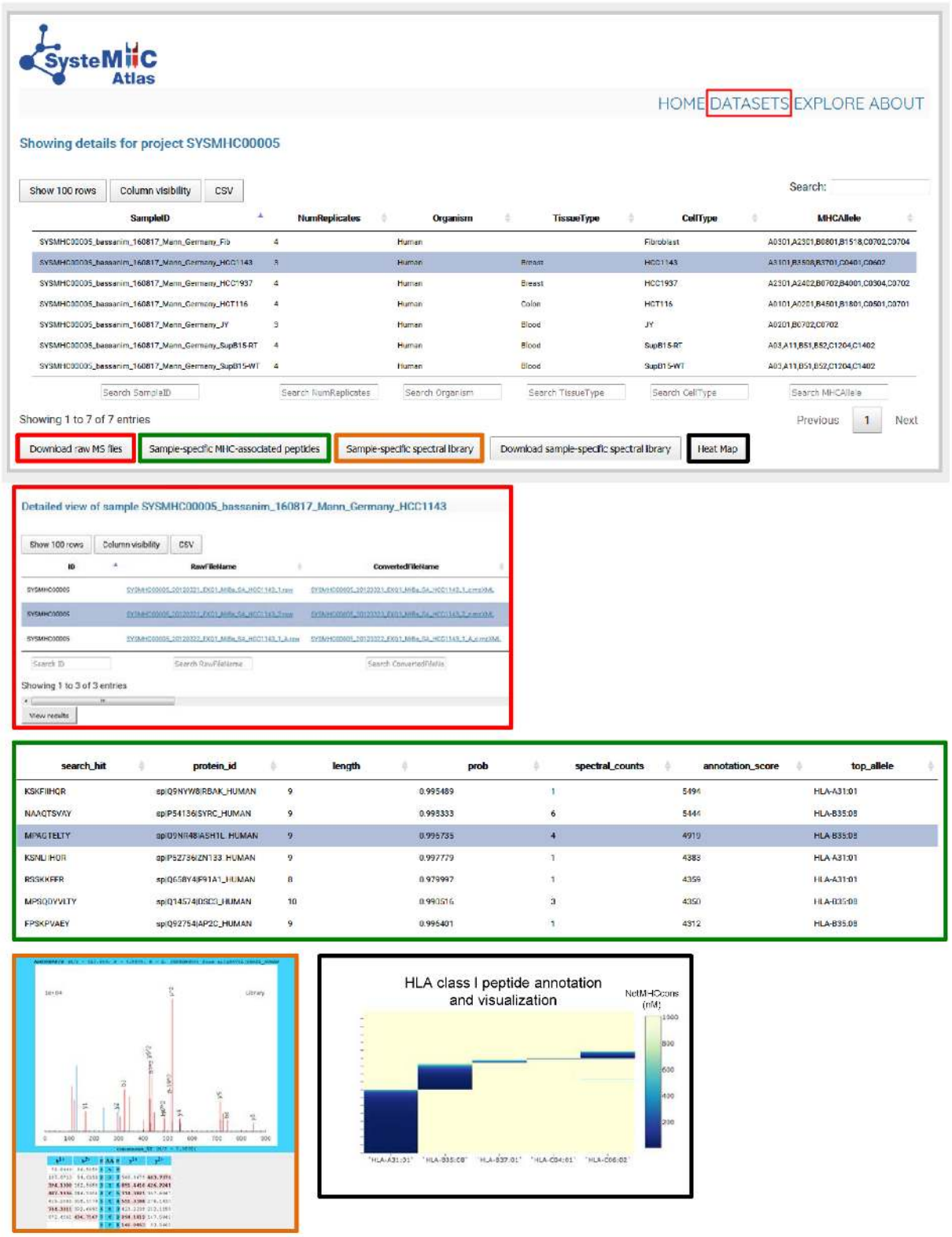

Figure 5. Data storage and visualization. To access information about specific datasets, the user selects a specific SYSMHC ID/Project name (e.g. SYSMHC00005) and clicks on 'view dataset' at the bottom left of the screen. The samples related to this project are then listed and linked to the number of replicates, organism, tissue and cell type of origin as well as the HLA typing information (upper panel). The user can then click on a specific Sample ID to visualize the metadata and to download the raw or converted mzXML MS files (red squares). A list of sample-specific HLA-associated peptides can be visualized at 1% peptide-level FDR (green squares). Sample-specific spectral libraries, including consensus fragment ion spectra, can be visualized and downloaded (orange and blue squares). Heat maps (black squares) are used to visualize the annotation of individual peptides to their respective HLA allele (dark blue peptides are predicted to be strong HLA binders according to NetMHCcons).

FUTURE DIRECTIONS

Data sharing, public resources and large-scale/community projects are growing in popularity and necessity in life sciences (43–48), and specifically in proteomics (49,50) where public data sharing is growing exponentially in recent years. Along this line, the SysteMHC Atlas represents the first community-driven resource devoted to collect, store, organize and share large immunopeptidomics datasets generated by MS methods—an important contribution to the Human Immuno-Peptidome Project (25,27). The SysteMHC Atlas will be further developed and enhanced to enable public dissemination of uniform and high-quality immunopeptidome data generated by an open and ever-improving computational pipeline. To this end, raw MS data will be reprocessed periodically using novel high-performance software tools as they are made available to the community. Future software tools are expected to outperform current algorithms for (i) MHC peptide identification, (ii) MHC peptide FDR estimation in large immunopeptidomic datasets and (iii) class I and class II peptide annotation to specific HLA alleles, as described (http://www.biorxiv.org/content/early/2017/05/13/098780) (51). In the near future, we aim at providing the necessary tools to retrieve information on post-translationally modified MHC-associated peptides: phosphopeptides, Arg-methylated peptides and proteasome-generated spliced peptides in particular, as those might be of particular relevance for the rational design of immunotherapeutic interventions (52–56). We also plan to identify the potential for large-scale integration and interoperability of all immunopeptidomic data with PRIDE (28), IEDB (57) and SWATHAtlas (34). Thus, we intend the SysteMHC Atlas to become a growing community-driven database and an interoperable, high-performance infrastructure for systematic analysis of terabytes of immunopeptidomic big data. If successful in longer term, we anticipate that the SysteMHC Atlas project will provide key insights to the immunology community and will foster the development of vaccines and immunotherapies against various immune-related diseases such as autoimmunity, allergies, infectious diseases and cancers.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Layla Lang (www.laylalang.com) for illustrating the immunopeptidomic landscape of cells on the website home page and An Guo for designing the logo of the SysteMHC Atlas. We thank Lorenz Blum and Pascal Kägi for computational assistance. We also thank all members of the Aebersold and Wollscheid laboratories for discussions.

FUNDING

TBVAC2020 [2–73838-14 to R.A.]; ERC grant Proteomics v3.0 [ERC-2008-AdG_20080422 to R.A.]; ERC Proteomics 4D [670821 to R.A.]; Swiss National Science Foundation [3100A0–688 107679 to R.A.]; Netherlands Organization for Scientific Research (NWO) (Roadmap Initiative Proteins@Work (project number 184.032.201)) (to F.M., A.J.R.H.); EC HOR2020 project TBVAC2020 [643381 to T.H.M.O.]; Wellcome Trust [WT101477MA to J.A.V.]; National Institute of Allergy and Infectious Diseases [HHSN272201200010C to A.S.]; National Health and Medical Council of Australia Senior Research Fellowship [APP1044215 to A.W.P.]; ERC [AdG339842 to H.-G.R.]; Mutaediting (to H.-G.R.); NIH, National Institute of General Medical Sciences [R01 GM087221 to E.W.D., R.L.M.]; 2P50 GM076547/Center for Systems Biology (to R.L.M.); Research Council of Norway [179573/V40 to A.G.deS., L.M.S., in part]; NIH National Institute of General Medical Sciences [P41 GM104603 to C.E.C.]. Funding for open access charge: TBVAC2020 [2-73838-14].
Conflict of interest statement. None declared.

REFERENCES

1. Istrail,S., Florea,L., Halldórsson,B.V., Kohlbacher,O., Schwartz,R.S., Yap,V.B., Yewdell,J.W. and Hoffman,S.L. (2004) Comparative immunopeptidomics of humans and their pathogens. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 13268–13272.
2. Caron,E., Vincent,K., Fortier,M.-H., Laverdure,J.-P., Bramoullé,A., Hardy,M.-P., Voisin,G., Roux,P.P., Lemieux,S., Thibault,P. *et al.* (2011) The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol. Syst. Biol.*, **7**, 533–533.
3. 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.
4. Neefjes,J., Jongsma,M.L.M., Paul,P. and Bakke,O. (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.*, **11**, 823–836.
5. Rock,K.L., Reits,E. and Neefjes,J. (2016) Present yourself! by MHC class I and MHC class II molecules. *Trends Immunol.*, **37**, 724–737.
6. Falk,K., Rötzschke,O., Stevanovic,S., Jung,G. and Rammensee,H.-G. (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature*, **351**, 290–296.
7. Cole,D.K. (2015) The ultimate mix and match: making sense of HLA alleles and peptide repertoires. *Immunol. Cell Biol.*, **93**, 515–516.
8. Bassani-Sternberg,M. and Coukos,G. (2016) Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr. Opin. Immunol.*, **41**, 9–17.
9. Mann,M. (2016) Origins of mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.*, **17**, 678.
10. Rötzschke,O., Falk,K., Deres,K., Schild,H., Norda,M., Metzger,J., Jung,G. and Rammensee,H.G. (1990) Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells. *Nature*, **348**, 252–254.
11. Hunt,D.F., Henderson,R.A., Shabanowitz,J., Sakaguchi,K., Michel,H., Sevilir,N., Cox,A.L., Appella,E. and Engelhard,V.H. (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science*, **255**, 1261–1263.
12. Hassan,C., Kester,M.G.D., de Ru,A.H., Hombrink,P., Drijfhout,J.W., Nijveen,H., Leunissen,J.A.M., Heemskerk,M.H.M., Falkenburg,J.H.F. and van Veelen,P.A. (2013) The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol. Cell. Proteomics*, **12**, 1829–1843.
13. Bergseng,E., Dørum,S., Arntzen,M.Ø., Nielsen,M., Nygård,S., Buus,S., de Souza,G.A and Sollid,L.M. (2014) Different binding motifs of the celiac disease-associated HLA molecules DQ2.5, DQ2.2, and DQ7.5 revealed by relative quantitative proteomics of endogenous peptide repertoires. *Immunogenetics*, **67**, 73–84.
14. Pearson,H., Daouda,T., Granados,D.P., Durette,C., Bonneil,E., Courcelles,M., Rodenbrock,A., Laverdure,J.-P., Côté,C., Mader,S. *et al.* (2016) MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.*, **126**, 4690–4701.
15. Abelin,J.G., Keskin,D.B., Sarkizova,S., Hartigan,C.R., Zhang,W., Sidney,J., Stevens,J., Lane,W., Zhang,G.L., Eisenhaure,T.M. *et al.* (2017) Mass spectrometry profiling of HLA-associated peptidomes in

mono-allelic cells enables more accurate epitope prediction. *Immunity*, **46**, 315–326.
16. Bassani-Sternberg,M., Bräunlein,E., Klar,R., Engleitner,T., Sinitcyn,P., Audehm,S., Straub,M., Weber,J., Slotta-Huspenina,J., Specht,K. *et al.* (2016) Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.*, **7**, 13404.
17. Khodadoust,M.S., Olsson,N., Wagar,L.E., Haabeth,O.A.W., Chen,B., Swaminathan,K., Rawson,K., Liu,C.L., Steiner,D., Lund,P. *et al.* (2017) Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature*, **543**, 723–727.
18. Kowalewski,D.J., Schuster,H., Backert,L., Berlin,C., Kahn,S., Kanz,L., Salih,H.R., Rammensee,H.-G., Stevanovic,S. and Stickel,J.S. (2014) HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E116–E175.
19. Mommen,G.P.M., Marino,F., Meiring,H.D., Poelen,M.C.M., van Gaans-van den Brink,J.A.M., Mohammed,S., Heck,A.J.R. and van Els,C.A.C.M. (2016) Sampling from the proteome to the human leukocyte antigen-DR (HLA-DR) ligandome proceeds via high specificity. *Mol. Cell. Proteomics*, **15**, 1412–1423.
20. Schellens,I.M.M., Hoof,I., Meiring,H.D., Spijkers,S.N.M., Poelen,M.C.M., van Gaans-van den Brink,J.A.M., van der Poel,K., Costa,A.I., van Els,C.A.C.M., van Baarle,D. *et al.* (2015) Comprehensive analysis of the naturally processed peptide repertoire: differences between HLA-A and B in the immunopeptidome. *PLoS One*, **10**, e0136417.
21. Mommen,G.P.M., Frese,C.K., Meiring,H.D., van Gaans-van den Brink,J., de Jong,A.P.J.M., van Els,C.A.C.M. and Heck,A.J.R. (2014) Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 4507–4512.
22. Wang,Q., Drouin,E.E., Yao,C., Zhang,J., Huang,Y., Leon,D.R., Steere,A.C. and Costello,C.E. (2016) Immunogenic HLA-DR-presented self-peptides identified directly from clinical samples of synovial tissue, synovial fluid, or peripheral blood in patients with rheumatoid arthritis or lyme arthritis. *J. Proteome Res.*, **16**, 122–136.
23. Pardoll,D.M. (2012) The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer*, **12**, 252–264.
24. van Eyk,J.E., Corrales,F.J., Aebersold,R., Cerciello,F., Deutsch,E.W., Roncada,P., Sanchez,J.-C., Yamamoto,T., Yang,P., Zhang,H. *et al.* (2016) Highlights of the Biology and Disease-driven Human Proteome Project, 2015–2016. *J. Proteome Res.*, **15**, 3979–3987.
25. Caron,E. and Aebersold,R. (2016) The Human Immuno-Peptidome Project: a new initiative of B/D-HPP Program. In: *15th Human Proteome Organization World Congress*. Taipei.
26. Sette,A., Schenkelberg,T.R. and Koff,W.C. (2015) Deciphering the human antigenome. *Expert Rev. Vaccines*, **15**, 167–171.
27. Admon,A. and Bassani-Sternberg,M. (2011) The Human Immunopeptidome Project, a suggestion for yet another postgenome next big thing. *Mol. Cell. Proteomics*, **10**, 1–4.
28. Vizcaíno,J.A., Csordas,A., del-Toro,N., Dianes,J.A., Griss,J., Lavidas,I., Mayer,G., Perez-Riverol,Y., Reisinger,F., Ternent,T. *et al.* (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.*, **44**, D447–D456.
29. Karosiene,E., Lundegaard,C., Lund,O. and Nielsen,M. (2012) NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*, **64**, 177–186.
30. Shteynberg,D., Nesvizhskii,A.I., Moritz,R.L. and Deutsch,E.W. (2013) Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics*, **12**, 2383–2393.
31. Shteynberg,D., Deutsch,E.W., Lam,H., Eng,J.K., Sun,Z., Tasman,N., Mendoza,L., Moritz,R.L., Aebersold,R. and Nesvizhskii,A.I. (2011) iProphet: Multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics*, **10**, 1–15.
32. Lam,H., Deutsch,E.W., Eddes,J.S., Eng,J.K., Stein,S.E. and Aebersold,R. (2008) Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods*, **5**, 873–875.
33. Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N. and Aebersold,R. (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–D658.
34. Caron,E., Espona,L., Kowalewski,D.J., Schuster,H., Ternette,N., Alpízar,A., Schittenhelm,R.B., Ramarathinam,S.H., Lindestam Arlehamn,C.S., Chiek Koh,C. *et al.* (2015) An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife*, **4**, e07661.
35. Escher,C., Reiter,L., MacLean,B., Ossola,R., Herzog,F., Chilton,J., MacCoss,M.J. and Rinner,O. (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics*, **12**, 1111–1121.
36. Faridi,P., Aebersold,R. and Caron,E. (2016) A first dataset toward a standardized community-driven global mapping of the human immunopeptidome. *Data Brief.*, **7**, 201–205.
37. Röst,H.L., Rosenberger,G., Navarro,P., Gillet,L., Miladinović,S.M., Schubert,O.T., Wolski,W., Ben C,Collins, Malmström,J., Malmström,L. *et al.* (2014) OpenSWATH enables automated, targeted analysis of data- independent acquisition MS data. *Nat. Biotechnol.*, **32**, 219–223.
38. Croft,N.P., Purcell,A.W. and Tscharke,D.C. (2015) Quantifying epitope presentation using mass spectrometry. *Mol. Immunol.*, **68**, 77–80.
39. Gubin,M.M., Zhang,X., Schuster,H., Caron,E., Ward,J.P., Noguchi,T., Ivanova,Y., Hundal,J., Arthur,C.D., Krebber,W.-J. *et al.* (2014) Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*, **515**, 577–581.
40. Lindestam Arlehamn,C.S., Gerasimova,A., Mele,F., Henderson,R., Swann,J., Greenbaum,J.A., Kim,Y., Sidney,J., James,E.A., Taplitz,R. *et al.* (2013) 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**, e1003130.
41. Tang,S.T., van Meijgaarden,K.E., Caccamo,N., Guggino,G., Klein,M.R., van Weeren,P., Kazi,F., Stryhn,A., Zaigler,A., Sahin,U. *et al.* (2011) Genome-based in silico identification of new mycobacterium tuberculosis antigens activating polyfunctional CD8+ T cells in human tuberculosis. *J. Immunol.*, **186**, 1068–1080.
42. Joosten,S.A., van Meijgaarden,K.E., van Weeren,P.C., Kazi,F., Geluk,A., Savage,N.D.L., Drijfhout,J.W., Flower,D.R., Hanekom,W.A., Klein,M.R. *et al.* (2010) Mycobacterium tuberculosis peptides presented by HLA-E molecules are targets for human CD8+ T-cells with cytotoxic as well as regulatory activity. *PLoS Pathog.*, **6**, e1000782.
43. Aebersold,R., Bader,G.D., Edwards,A.M., van Eyk,J.E., Kussmann,M., Qin,J. and Omenn,G.S. (2013) The Biology/Disease-driven Human Proteome Project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.*, **12**, 23–27.
44. Uhlén,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
45. Kusebauch,U., Campbell,D.S., Deutsch,E.W., Chu,C.S., Spicer,D.A., Brusniak,M.-Y., Slagel,J., Sun,Z., Stevens,J., Grimes,B. *et al.* (2016) Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. *Cell*, **166**, 766–778.
46. Whiteaker,J.R., Halusa,G.N., Hoofnagle,A.N., Sharma,V., MacLean,B., Yan,P., Wrobel,J.A., Kennedy,J., Mani,D.R., Zimmerman,L.J. *et al.* (2014) CPTAC Assay Portal: a repository of targeted proteomic assays. *Nat. Methods*, **11**, 703–704.
47. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
48. McKiernan,E.C., Bourne,P.E., Brown,C.T., Buck,S., Kenall,A., Lin,J., McDougall,D., Nosek,B.A., Ram,K., Soderberg,C.K. *et al.* (2016) How open science helps researchers succeed. *Elife*, **5**, e16800.
49. Vaudel,M., Verheggen,K., Csordas,A., Raeder,H., Berven,F.S., Martens,L., Vizcaíno,J.A. and Barsnes,H. (2016) Exploring the potential of public proteomics data. *Proteomics*, **16**, 214–225.
50. Martens,L. and Vizcaíno,J.A. (2017) A golden age for working with public proteomics data. *Trends Biochem. Sci.*, **42**, 333–341.
51. Bassani-Sternberg,M. and Gfeller,D. (2016) Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide–HLA interactions. *J. Immunol.*, **197**, 2492–2499.
52. Zarling,A.L., Polefrone,J.M., Evans,A.M., Mikesh,L.M., Shabanowitz,J., Lewis,S.T., Engelhard,V.H. and Hunt,D.F. (2006)

Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 14889–14894.
53. Cobbold,M., De La Peña,H, Norris,A., Polefrone,J.M., Qian,J., English,A.M., Cummings,K.L., Penny,S., Turner,J.E., Cottine,J. *et al.* (2013) MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia. *Sci. Transl. Med.*, **5**, 203ra125.
54. Marino,F., Mommen,G.P.M., Jeko,A., Meiring,H.D., van Gaans-van den Brink,J.A.M., Scheltema,R.A., van Els,C.A.C.M. and Heck,A.J.R. (2016) Arginine (Di)methylated Human Leukocyte Antigen Class I Peptides Are Favorably Presented by HLA-B*07. *J. Proteome Res.*, **16**, 34–44.
55. Liepe,J., Marino,F., Sidney,J., Jeko,A., Bunting,D.E., Sette,A., Kloetzel,P.M., Stumpf,M.P., Heck,A.J. and Mishto,M. (2016) A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science*, **354**, 354–358.
56. Delong,T., Wiles,T.A., Baker,R.L., Bradley,B., Barbour,G., Reisdorph,R., Armstrong,M., Powell,R.L., Reisdorph,N., Kumar,N. *et al.* (2016) Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion. *Science*, **351**, 711–714.
57. Vita,R., Overton,J.A., Greenbaum,J.A., Ponomarenko,J., Clark,J.D., Cantrell,J.R., Wheeler,D.K., Gabbard,J.L., Hix,D., Sette,A. *et al.* (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, **43**, D405–D412.