Identifying neoantigens for cancer vaccines by personalized deep learning of individual immunopeptidomes

Ngoc Hieu Tran[†,1], Rui Qiao[†,2], Lei Xin[3], Xin Chen[3], Baozhen Shan[3], Ming Li[1,*]

[1] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.

[2] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

[3] Bioinformatics Solutions Inc., Waterloo, Ontario, Canada.

[†] Joint-first authors.

* Corresponding author. Email: mli@uwaterloo.ca

Abstract. Tumor-specific neoantigens play the main role for developing personal vaccines in cancer immunotherapy. We propose, for the first time, a personalized *de novo* sequencing workflow to identify HLA-I and HLA-II neoantigens directly and solely from mass spectrometry data. Our workflow trains a specific deep learning model on the immunopeptidome of an individual patient and then use it to predict mutated neoantigens of that patient. This personalized learning and mass spectrometry-based approach enables comprehensive and accurate identification of neoantigens. We applied the workflow to datasets of eight melanoma patients and seven other tumor samples. Our workflow successfully identified neoantigens of both HLA-I and HLA-II, including those with validated T cell responses and those novel neoantigens that had not been reported in previous studies.

INTRODUCTION

Neoantigens are tumor-specific mutated peptides that are brought to the surface of tumor cells by major histocompatibility complex (MHC) proteins for T cell recognition. As neoantigens carry tumor-specific mutations and are not found in normal tissues, they represent ideal targets for the immune system to distinguish cancer cells from non-cancer ones [1-3]. The potential of neoantigens for personalized cancer vaccines were reinforced by evidences of neoantigen-specific T cells found in some cancer patients treated with the immune checkpoint inhibitor therapy [4, 5]. Indeed, three independent studies have further demonstrated successful clinical trials of personalized neoantigen vaccines for patients with melanoma [6-8]. And so began the "gold rush" for neoantigen mining [1-3]. The current prevalent approach to identify candidate neoantigens often includes two major phases: (i) exome sequencing of cancer and normal tissues to find somatic mutations and (ii) predicting which mutated peptides are most likely to be presented by MHC proteins for T cell recognition. The first phase is strongly backed by high-throughput sequencing technologies and bioinformatics pipelines that have been well established through several genome sequencing projects during the past decade. The second phase, however, is still facing challenges due to our lack of knowledge of the MHC antigen processing pathway: how mutated proteins are processed into peptides; how those peptides are delivered to the endoplasmic reticulum by the transporter associated with antigen processing; and how they bind to MHC proteins. To make it further complicated, human leukocyte antigens (HLA), those genes that encode MHC proteins, are located among the most genetically variable regions and their alleles basically change from one individual to another. The problem is especially more challenging for HLA class II (HLA-II) peptides than HLA class I (HLA-I), because the formers are longer, their motifs have greater variations, and very limited data is available.

Current *in silico* methods focus on predicting which peptides bind to MHC proteins given the HLA alleles of a patient, e.g. NetMHC [9, 10]. However, only very few from a large number of predicted candidates were confirmed to be presented on the surface of tumor cells and even less were found to trigger T cell responses [1, 2]. Several efforts have been made to improve the MHC binding prediction, including using mass spectrometry data in addition to binding affinity data for more accurate prediction of MHC antigen presentation [11, 12]. Recently, proteogenomic approaches have been proposed to combine mass spectrometry and exome sequencing to identify neoantigens directly isolated from MHC proteins, thus overcoming the limitations of MHC binding prediction [13, 14]. In those approaches, exome sequencing was performed to build a customized protein database that included all normal and mutated protein sequences. The database was further used by a search engine to identify endogenous peptides, including neoantigens, that were obtained by immunoprecipitation assays and mass spectrometry. Existing database search engines, however, are not designed for HLA peptides and are biased towards tryptic peptides [15, 16]. Furthermore, they have sensitivity and specificity issues when dealing with a very large search space created by (i) all mRNA isoforms from exome sequencing and (ii) unknown digestion rules for HLA peptides. Exome sequencing errors and biases could also be accumulated into the database search process.

In this study, we propose, for the first time, a personalized *de novo* sequencing workflow to identify HLA-I and HLA-II neoantigens directly and solely from mass spectrometry data. *De novo* sequencing is invented for the purpose of discovering novel peptides and proteins, genetic variants or mutations. Thus, its application to identify neoantigens is a perfect match. We bring

de novo sequencing to the "personalized" level by training a specific machine learning model for each individual patient using his/her own data. In particular, we use the collection of normal HLA peptides, i.e. the immunopeptidome, of a patient to train a *de novo* sequencing model and then use it to predict mutated HLA peptides of that patient. Learning an individual's immunopeptidome is made possible by our recent deep learning model, DeepNovo [17, 18], which uses a long-short-term memory (LSTM) recurrent neural network (RNN) to capture sequence patterns in peptides or proteins, in a similar way to natural languages [19]. This personalized learning workflow can overcome the challenge of numerous variations in HLA alleles. The accuracy of *de novo* sequencing can also be significantly boosted for comprehensive and reliable identification of neoantigens. Furthermore, this mass spectrometry-based approach can bypass the questions of MHC antigen processing pathway that are yet fully understood. We applied the workflow to datasets of eight melanoma patients and seven other human tumor samples from three previous studies [8, 13, 14]. Our workflow successfully identified neoantigens of both HLA-I and HLA-II, including those with validated T cell responses and those novel neoantigens that had not been reported in earlier studies.

RESULTS

Personalized *De novo* Sequencing of Individual Immunopeptidomes

Here we describe four steps of personalized *de novo* sequencing from the LC-MS/MS data of HLA peptides of an individual patient: (1) build the immunopeptidome of the patient; (2) train personalized machine learning model; (3) personalized *de novo* sequencing; and (4) quality control of *de novo* peptides (Figure 1). The details of each step on an example dataset, patient Mel-15 from [13], are also provided for illustration (Supplementary Table S1). Step 5, neoantigen selection, is described in a later section.

To build the immunopeptidome of the patient, we search the LC-MS/MS data against the standard UniProt/Swiss-Prot human protein database. As digestion rules for HLA peptides are unknown, a search engine that supports no-enzyme-specific digestion is needed (we used PEAKS X [18]). Other common settings include precursor mass tolerance 15 ppm, fragment mass tolerance 0.05 Da, variable modifications (oxidation and deamidation), and false discovery rate (FDR) 1.0%. We identify normal HLA peptides and their peptide-spectrum matches (PSMs) which represent the patient's immunopeptidome and its spectral library. Mutated HLA peptides do not exist in the database and their spectra remain unlabeled. For example, we identified 341,216 PSMs of 35,551 HLA-I peptides and 67,021 PSMs of 9,664 HLA-II peptides from patient Mel-15 dataset. The numbers of unlabeled spectra were 596,915 and 135,490, respectively.

Next, we use the identified PSMs to train DeepNovo, a neural network model for *de novo* peptide sequencing [17, 18]. In addition to capturing fragment ions in MS/MS spectra, DeepNovo also learns sequence patterns of peptides by modelling them as a special language with an alphabet of 20 amino acid letters. This unique advantage allows DeepNovo to adapt to a specific immunopeptidome of an individual patient. At the same time, it is essential to apply counter-overfitting techniques to avoid "memorizing" the training peptide sequences. We partition the PSMs into training, validation, and test sets so that they do not share common peptide sequences; and then stop the training process if there is no improvement on the validation set. As a result, the model can achieve very high accuracy on an individual immunopeptidome and can detect mutations that it has not seen during training. This approach

is particularly useful for missense mutations, the primary source of neoantigens. The model can be trained from scratch or re-trained from an existing one if the training data is limited. For patient Mel-15 dataset, we trained HLA-I model from scratch and HLA-II model on top of HLA-I model.

After training DeepNovo model, we perform *de novo* peptide sequencing on both labeled spectra (i.e., the identified PSMs) and unlabeled spectra. Results from labeled spectra are needed for accuracy evaluation and quality control of *de novo* sequencing, e.g. calibrating prediction confidence scores in the next step*. De novo* peptides identified from unlabeled spectra contain mutated peptides, and among them, candidate neoantigens of interest.

Next, a quality control procedure is designed to select high-confidence *de novo* peptides and to estimate their false discovery rate (FDR). We first calculate the sequencing accuracy on the test set of PSMs. As it does not share common peptides with the training set, the performance on the test set is approximately what we could expect for mutated peptides that the model has not seen during training. For instance, we achieved an overall precision of 86.68% for HLA-I and 69.86% for HLA-II on Mel-15 test sets (precision: the total number of correctly predicted amino acids divided by the total number of predicted amino acids).The accuracy can be further boosted substantially by filtering predicted peptides with DeepNovo confidence score. We analyze the distribution of score versus precision on the test set to determine a score threshold and use it to filter *de novo* peptides of unlabeled spectra. In addition, we also apply a majority-voting technique to correct common errors of *de novo* sequencing. For patient Mel-15 dataset, we found 25,665 HLA-I and 3,867 HLA-II high-confidence *de novo* peptides with an expected precision of 95%. Finally, to estimate the FDR of those *de novo* peptides, we perform a second-round PEAKS X search of all MS/MS spectra against a combined list of *de novo* peptides and database peptides (from step 1). Only *de novo* peptides identified at 1% FDR are retained. Thus, our stringent procedure of quality control guarantees that each *de novo* peptide is supported by solid evidences from two independent tools, DeepNovo and PEAKS X. We obtained 5,320 HLA-I and 863 HLA-II *de novo* peptides as candidates for neoantigen selection for patient Mel-15 (Figure 1, Supplementary Table S1).

Evaluation of *De novo* HLA peptides

Figure 2 presents several evaluation measures on 5,320 HLA-I *de novo* peptides that we identified from patient Mel-15 dataset. In particular, Figures 2a and 2b explain the distribution of DeepNovo confidence score versus accuracy and the majority-voting technique to correct *de novo* sequencing errors. Figure 2c shows an example neoantigen PSM; Figure 2d shows the distribution of PEAKS X scores of *de novo* PSMs against those of database and decoy PSMs together with 1% FDR threshold. The distributions confirm that *de novo* peptides have strong supporting PSMs as database peptides and are clearly distinguishable from decoy peptides. Figure 2e shows that both *de novo* and database peptides have the same characteristic length distribution of HLA-I peptides. We also downloaded previously reported human HLA-I peptides from the Immune Epitope Database (IEDB) [20] for the four alleles HLA-A03:01, HLA-A68:01, HLA-B27:05, and HLA-B35:03 of patient Mel-15. The Venn diagram in Figure 2f shows that 56 *de novo* peptides were reported in earlier studies. The small overlap between database and IEDB peptides is due to the large variation of HLA peptides and further emphasizes the importance of our personalized approach. Finally, we used NetMHCpan [10] to verify the binding affinity of *de novo*, database, and IEDB peptides to MHC proteins of the respective alleles. Figures 2g and 2h show that *de novo* peptides have similar levels of binding affinity as

database and IEDB peptides. Furthermore, majority of *de novo* peptides were predicted as good binders by different criteria: 74.0% (3,938/5,320) with affinity less than 500 nM, 79.3% (4,220/5,320) with rank less than 2% (weak-binding), and 51.8% (2,757/5,320) with rank less than 0.5% (strong-binding). Figure 2i shows an example binding motif of *de novo* peptides for allele HLA-A68:01, which is very close to the motif of NetMHCpan database [10]. Overall, the evaluation results in Figure 2 confirm the correctness and HLA characteristics of our *de novo* peptides.

Neoantigen Selection

The last step of our workflow focuses on criteria to select or prioritize candidate neoantigens from *de novo* peptides (Figure 1, step 5). We consider several features that have been widely used in previous studies for neoantigen selection [6-8, 13, 14, 21, 22]. Specifically, we check whether a peptide carries one amino acid substitution by aligning its sequence to the UniProt/Swiss-Prot human protein database and whether that substitution involves one single nucleotide difference in the codons. In this paper, we refer to those substitutions as “missense-like” mutations. In addition, we record whether the corresponding wild-type is detected and the mutation is at a flanking position. Next, we calculate expression level information for each peptide, including the number of PSMs, their total score, and their total abundance. We also calculate the binding affinity of each peptide to MHC proteins using NetMHCpan [10] and the patient’s HLA alleles. The results for 5,320 HLA-I and 863 HLA-II peptides of patient Mel-15 are provided in Supplementary Tables S2 and S3.

The problem of neoantigen selection can be addressed by a machine learning approach with the above features as input. However, such approach requires enough data of immunogenicity-validated neoantigens. Here we apply a pre-defined set of criteria to filter candidate neoantigens. First, we only choose peptides that carry one single missense-like mutation. We found that the number of peptides was reduced significantly, e.g., from 5,320 to 328 HLA-I and from 863 to 154 HLA-II peptides of patient Mel-15. We further filter out peptides with mutations at flanking positions or with only one supporting PSM because they are error-prone and less stable to be effective neoantigens. We also remove peptides that belong to MHC or antibody proteins in the immunoprecipitation assays. Those criteria considerably improve the quality of candidate neoantigens. Our final list of candidate neoantigens for patient Mel-15 includes 158 HLA-I peptides, with 88.6% (140/158) weak-binding and 64.6% (102/158) strong-binding, and 37 HLA-II peptides (Supplementary Tables S1, S2, S3). Binding affinity and expression level information can be further used to prioritize candidate neoantigens, we avoid using them as hard filters.

Evaluation of Candidate Neoantigens against RNA-Seq

We cross-checked our candidate neoantigens of patient Mel-15 against the nucleotide mutations and mRNA transcripts reported by RNA-Seq in the original publication [13]. Our workflow successfully recovered four of eight HLA-I neoantigens reported in that study (Table 1), including the one that triggered superior T-cell responses, “GRIAF**F**LKY” (gene name: SYTL4; protein name: Synaptotagmin-like protein 4). All four neoantigens have strong-binding affinity. However, the expression level of “GRIAF**F**LKY”, including number of PSMs, total score, total abundance, are at least one order of magnitude higher than the others (Supplementary Figure S1, Supplementary Table S2). This observation confirms the critical role of protein/peptide abundance for effective immune responses. We also investigated the missing

four neoantigens. Three of them are not supported by good PSMs, and in fact, DeepNovo and PEAKS X identified alternative peptides that better matched the corresponding spectra (Supplementary Figure S2). Those three were also originally reported in [13] at 5.0% FDR instead of 1.0%. The remaining neoantigen was filtered out because the mutation is at a flanking position.

More importantly, our workflow identified an extra neoantigen that matched a nucleotide mutation reported by RNA-Seq: “SL**S**SALRPSTSRSLY” (gene name: VIM; protein name: Vimentin). Even more interesting is that the neoantigen is HLA-II but it showed up in HLA-I dataset, probably due to purification limitations (Table 1). This neoantigen would have been missed if a filter of binding affinity had been imposed (Supplementary Table S2). Our analysis of 37 HLA-II candidate neoantigens further found ten peptides that covered this mutation, and among them, “SL**S**SALRPSTSRSLY” had the best abundance level and binding affinity (Table 2, Supplementary Table S3). The results in Table 2 also illustrate the complicated digestion of HLA-II peptides and emphasizes the essential of our personalized *de novo* sequencing workflow.

Identifying Neoantigens for Eight Melanoma Patients and Seven Other Tumor Samples

We applied the workflow to datasets of eight melanoma patients and seven other tumor samples from three previous studies [8, 13, 14]. The results are summarized in Supplementary Table S4 and are discussed in Supplementary Note 1.

Acknowledgements

This work was funded in part by the NSERC OGP0046506 grant, China's Research and Development Program grants 2016YFB100090 and 2018YFB1003202, NSFC grant 61832019, and the Canada Research Chair program. N.H.T. was supported by the Mitacs Elevate Fellowship. The authors thank Nicole Keshav and Dr. Kwok Pui Choi for proofreading the manuscript.

Author Contributions

M.L. and B.S. conceived the research idea. N.H.T. designed the neoantigen discovery workflow. N.H.T and R.Q. implemented the software and analyzed the results. X.C. and L.X. contributed to the model design, software development, and data analysis. N.H.T, M.L., and R.Q. wrote the manuscript. M.L., B.S., and L.X. supervised the research project.

Competing Interests Statement

The workflow in Figure 1 has been applied for patent in the USPTO Provisional Application number 62/798,830, filed by Bioinformatics Solutions Inc., Waterloo, Canada, on 30 January 2019. The authors are named inventors in the patent application. L.X., X.C., and B.S. are employees of Bioinformatics Solutions Inc.

Software Availability

DeepNovo and the workflow are implemented in Python. The latest version is provided to the editor and reviewers during the review process and shall be made available on GitHub (https://github.com/nh2tran/DeepNovo) upon publication.

References

1. Hu, Z., Ott, P.A., & Wu, C.J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.* **18**, 168-182 (2018).
2. Editorial. The problem with neoantigen prediction. *Nat. Biotechnol.* **35**, 97 (2017).
3. Vitiello, A. & Zanetti M. Neoantigen prediction and the need for validation. *Nat. Biotechnol.* **35**, 815-817 (2017).
4. Van Allen, E.M. et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207-211 (2015).
5. Rizvi, N.A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124-128 (2015).
6. Ott, P.A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217-221 (2017).
7. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222-226 (2017).
8. Carreno, B.M. *et al.* A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* **348**, 803-808 (2015).

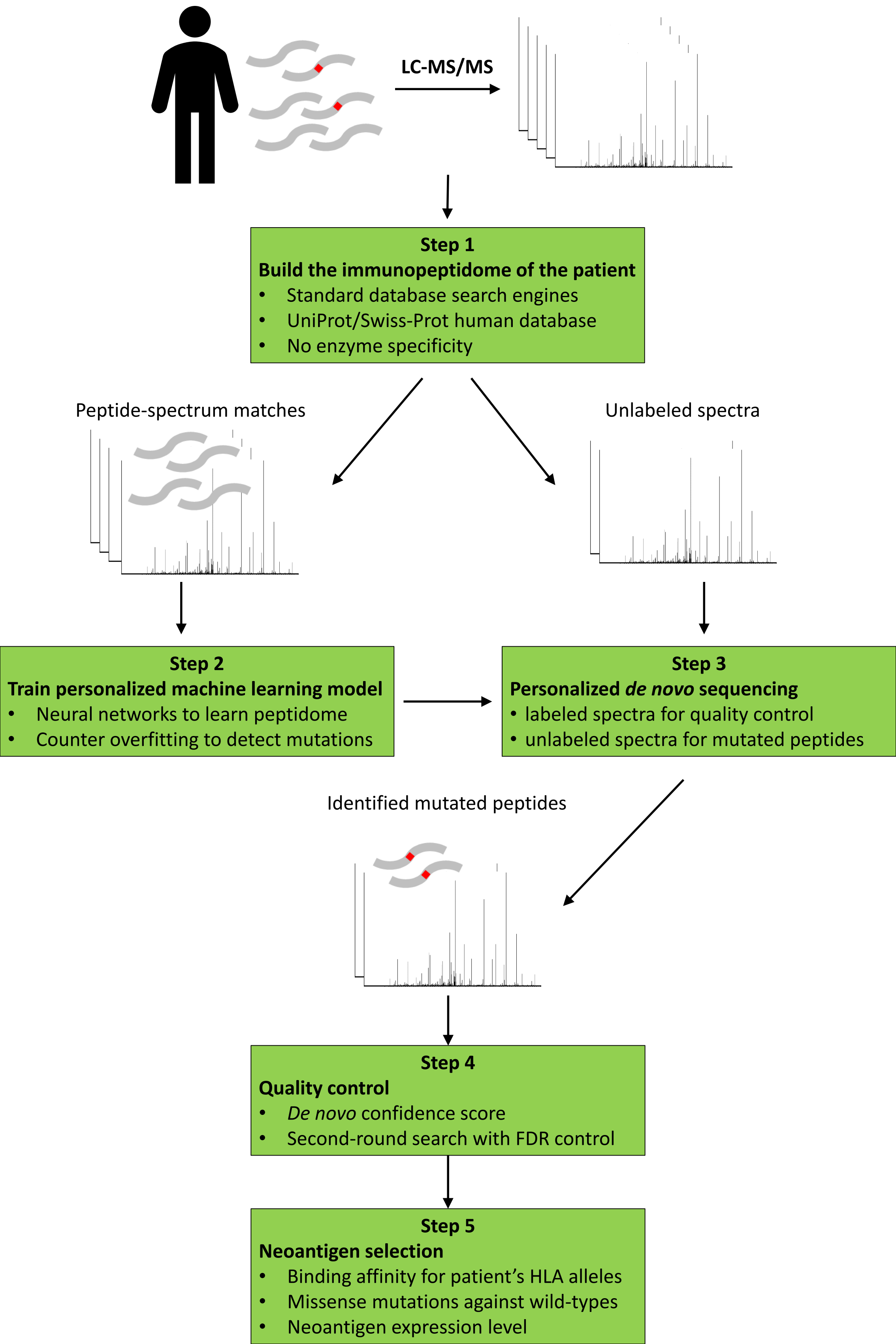
9. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511-517 (2016).
10. Jurtz, V. *et al.* NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360-3368 (2017).
11. Abelin J.G. *et al.* Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315-326 (2017).
12. Bulik-Sullivan, B. *et al.* Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* **37**, 55-63 (2019).
13. Bassani-Sternberg M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404 (2016).
14. Laumont, C.M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516 (2018).
15. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-1372 (2008).
16. Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111.010587 (2012).
17. Tran, N.H, Zhang, X., Xin, L., Shan, B. & Li, M. *De novo* peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8247-8252 (2017).
18. Tran, N.H *et al.* Deep learning enables *de novo* peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* **16**, 63-66 (2019).
19. Sutskever, I., Vinyals, O. & Le, Q. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **27**, 3104-3112 (2014).
20. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339-D343 (2018).
21. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen L.J., & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **14**, 658-673 (2015).
22. Keskin, D.B. *et al.* Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **565**, 234-239 (2019).

Figure Legends

Figure 1. The personalized *de novo* sequencing workflow for neoantigen discovery.

Figure 2. Evaluation of *de novo* HLA-I peptides from patient Mel-15 dataset. (a) Distribution of amino acid accuracy versus DeepNovo confidence score. (b) Majority-voting technique to correct *de novo* sequencing errors. *De novo* peptides are assigned to groups of the same mass and the peptide with the highest total score is used to correct other peptides with Hamming distance at most two. (c) Peptide-spectrum match (PSM) of the neoantigen “RLFKGYEGSLLK”.

(d) PSM score distributions of *de novo*, database, and decoy peptides. (e) Length distributions of *de novo* and database peptides. (f) Venn diagram of *de novo*, database, and IEDB peptides. (g) Binding affinity distributions of *de novo*, database, and IEDB peptides. (h) Binding affinity rank distributions of *de novo*, database, and IEDB peptides. (i) Binding motif of *de novo* peptides for allele HLA-A68:01.


LC-MS/MS
Step 1
Build the immunopeptidome of the patient
• Standard database search engines
• UniProt/Swiss-Prot human database
• No enzyme specificity
Peptide-spectrum matches
Unlabeled spectra
Step 2
Train personalized machine learning model
• Neural networks to learn peptidome
• Counter overfitting to detect mutations
Step 3
Personalized de novo sequencing
• labeled spectra for quality control
• unlabeled spectra for mutated peptides
Identified mutated peptides
Step 4
Quality control
• De novo confidence score
• Second-round search with FDR control
Step 5
Neoantigen selection
• Binding affinity for patient's HLA alleles
• Missense mutations against wild-types
• Neoantigen expression level

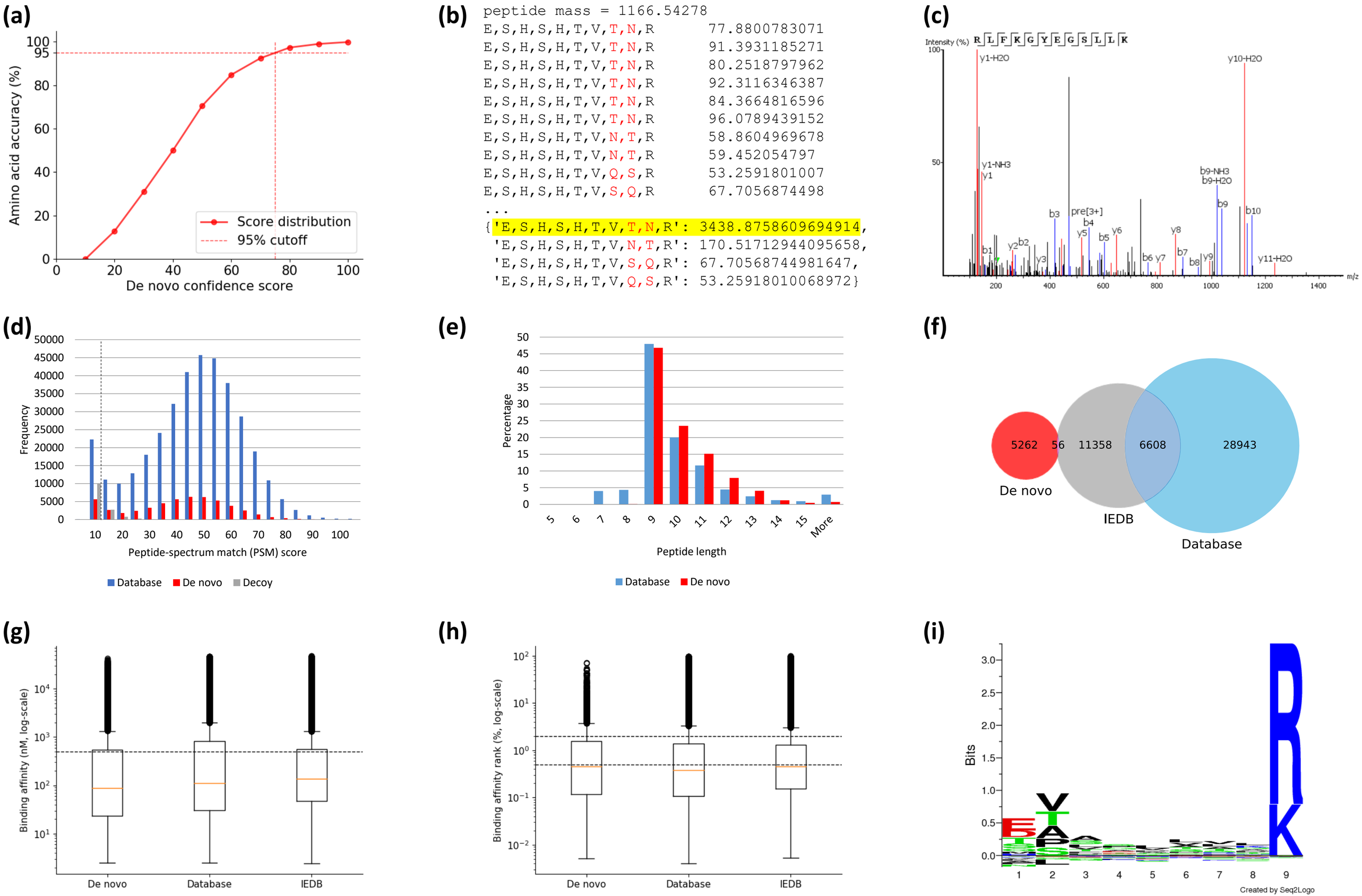

(a)
Amino acid accuracy (%)
De novo confidence score
Score distribution
95% cutoff
(b)
peptide mass = 1166.54278
E,S,H,S,H,T,V,T,N,R 77.8800783071
E,S,H,S,H,T,V,T,N,R 91.3931185271
E,S,H,S,H,T,V,T,N,R 80.2518797962
E,S,H,S,H,T,V,T,N,R 92.3116346387
E,S,H,S,H,T,V,T,N,R 84.3664816596
E,S,H,S,H,T,V,T,N,R 96.0789439152
E,S,H,S,H,T,V,N,T,R 58.8604969678
E,S,H,S,H,T,V,N,T,R 59.452054797
E,S,H,S,H,T,V,Q,S,R 53.2591801007
E,S,H,S,H,T,V,S,Q,R 67.7056874498
...
{'E,S,H,S,H,T,V,T,N,R': 3438.8758609694914,
'E,S,H,S,H,T,V,N,T,R': 170.51712944095658,
'E,S,H,S,H,T,V,S,Q,R': 67.70568744981647,
'E,S,H,S,H,T,V,Q,S,R': 53.25918010068972}
(c)
Intensity (%)
R L F K G Y E G S L L K
y1-H2O
y10-H2O
y1-NH3
y1
b9-NH3
b9-H2O
b9
b10
b3
pre[3+]
b4
y5
b5
y6
y8
b1
y2
b2
y3
b6
y7
b7
b8
y9
y11-H2O
m/z
(d)
Frequency
Peptide-spectrum match (PSM) score
Database
De novo
Decoy
(e)
Percentage
Peptide length
More
Database
De novo
(f)
5262
56
11358
6608
28943
De novo
IEDB
Database
(g)
Binding affinity (nM, log-scale)
De novo
Database
IEDB
(h)
Binding affinity rank (%, log-scale)
De novo
Database
IEDB
(i)
Bits
Created by Seq2Logo

Table 1. Identified neoantigens for patient Mel-15.Green rows: MHC class 1; yellow row: MHC class 2; red letters: mutated amino acids.

Chr	Position	Ref Allele	Alt Allele	Genename	Transcript ID	Effect	Aa change	wildtype peptide	de novo peptide
X	100687163	G	A	SYTL4	ENST00000263033	missense_variant	Ser363Phe	GRIAFSLKY	GRIAFFLKY
14	32822259	G	A	AKAP6	ENST00000280979	missense_variant	Met1482Ile	KLKLPMIMK	KLKLPIIMK
10	17229543	G	A	VIM	ENST00000224237	missense_variant	Gly41Ser	SLGSALRPSTSRSLY	SLSSALRPSTSRSLY
7	158680743	G	A	NCAPG2	ENST00000441982	missense_variant	Pro134Leu	KPILWRGLK	KLILWRGLK
8	30474849	C	T	RBPMS	ENST00000517860	missense_variant	Pro46Leu	RPFKGYEGSLIK	RLFKGYEGSLIK

Table 2. Alignment of candidate mutated peptides against the reference sequence from the MHC class 2 dataset of patient Mel-15. The mutated site is highlighted in green and yellow colors, for reference and mutated amino acids respectively. The columns provide supporting evidence of binding affinity rank (lower is better), number of peptide-spectrum matches (PSMs), the total confidence score of PSMs, and the total abundance of PSMs. Two candidate neoantigens “SLSSALRPSTSRSLY” and “TSTRTYSLSSALRPS” are highlighted in red color. “SLSSALRPSTSRSLY.1” shows the identification of this peptide from the MHC class 1 dataset.

…SYVTTSTRTYSLGSALRPSTSRSLY…	binding %	num_psm	total_score	total_abundance
YVTTSTRTYSLSSALRPSTS		8	573.18	4.26E+06
VTTSTRTYSLSSALRPSTS		5	290.37	2.25E+06
SLSSALRPSTSRSLY	0.08	8	351.79	1.70E+07
SLSSALRPSTSRSLY.1	0.08	11	622.22	3.44E+07
SYVTTSTRTYSLSSALRPSTS		8	666.41	9.78E+06
VTTSTRTYSLSSALRPS		3	115.28	3.03E+06
TTSTRTYSLSSALRPS		6	340.06	1.10E+07
YVTTSTRTYSLSSALRPST		2	104.34	3.68E+05
TTSTRTYSLSSALRPSTS		3	120.71	7.46E+05
YVTTSTRTYSLSSALRPS		2	121.59	8.45E+05
TSTRTYSLSSALRPS	0.43	12	498.53	1.26E+07

Table S1. Personalized workflow of neoantigen discovery for patient Mel-15

Neoantigen discovery workflow	HLA-I *de novo* peptides	HLA-II *de novo* peptides
Step 1: Build the immunopeptidome of the patient Standard database search engines UniProt/Swiss-Prot human database No enzyme specificity	(not yet available)	(not yet available)
Step 2: Train personalized machine learning model Neural networks to learn peptidome Counter overfitting to detect mutations	(not yet available)	(not yet available)
Step 3: Personalized de novo sequencing Labeled spectra for quality control Unlabeled spectra for mutated peptides	366309	93983
Step 4: Quality control De novo confidence score Second-round search with FDR control	25665 5320	3867 863
Step 5: Neoantigen selection Binding affinity for patient's HLA alleles Missense mutations against wild-types Neoepitope expression level	158	37

Figure S1. Distributions of supporting peptide-spectrum matches (PSMs) of candidate neoantigens and the T-cell-stimulating neoantigen (red dot).

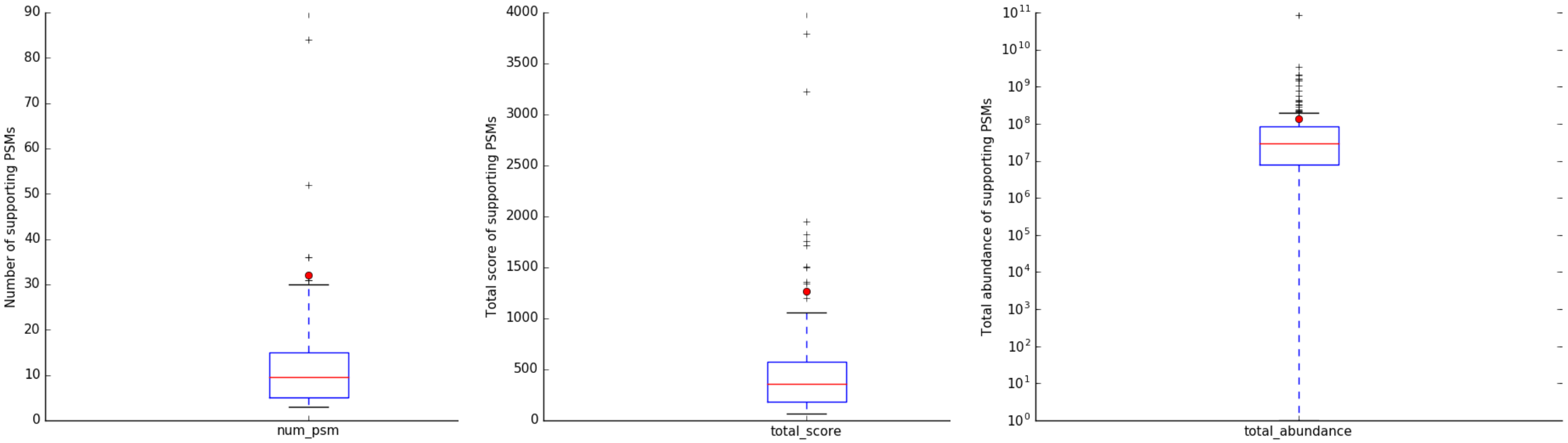
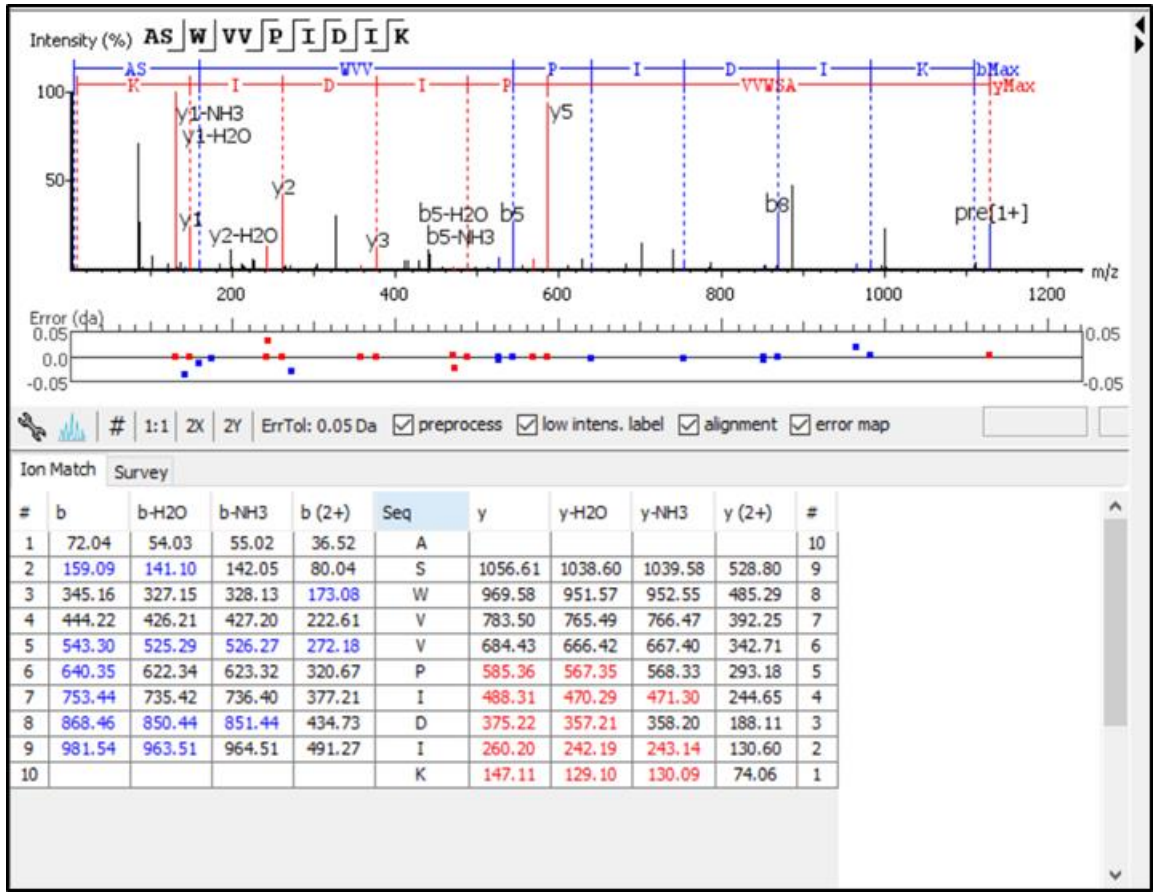


Figure S2. Peptide-spectrum matches of MaxQuant and DeepNovo for 3 candidate neoantigens that are likely to be false positives.

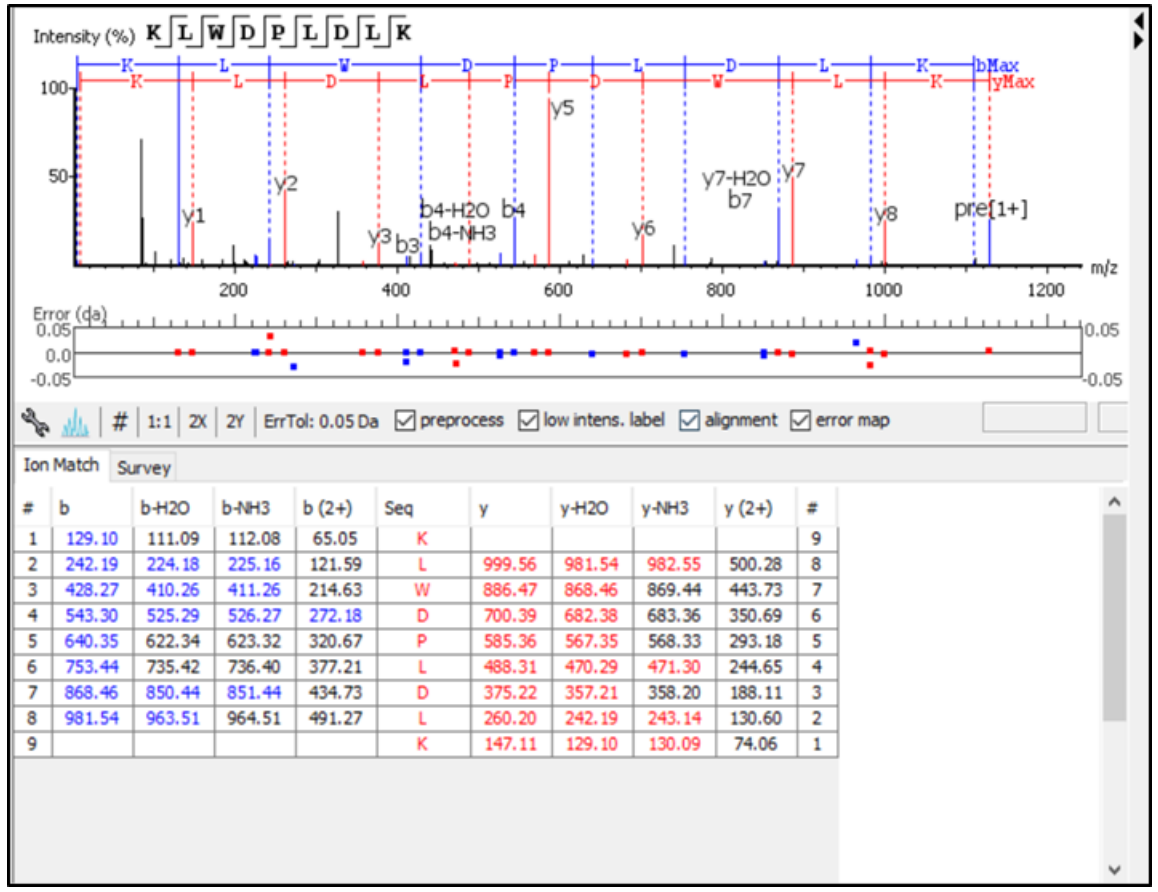
Fraction: 20141208_QEp7_MiBa_SA_HLA-I-p_MM15_4_B.raw
Scan ID: 49534
Retention time: 82.974
M/z: 564.327
Charge: 2

MaxQuant



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	72.04	54.03	55.02	36.52	A					10
2	159.09	141.10	142.05	80.04	S	1056.61	1038.60	1039.58	528.80	9
3	345.16	327.15	328.13	173.08	W	969.58	951.57	952.55	485.29	8
4	444.22	426.21	427.20	222.61	V	783.50	765.49	766.47	392.25	7
5	543.30	525.29	526.27	272.18	V	684.43	666.42	667.40	342.71	6
6	640.35	622.34	623.32	320.67	P	585.36	567.35	568.33	293.18	5
7	753.44	735.42	736.40	377.21	I	488.31	470.29	471.30	244.65	4
8	868.46	850.44	851.44	434.73	D	375.22	357.21	358.20	188.11	3
9	981.54	963.51	964.51	491.27	I	260.20	242.19	243.14	130.60	2
10					K	147.11	129.10	130.09	74.06	1

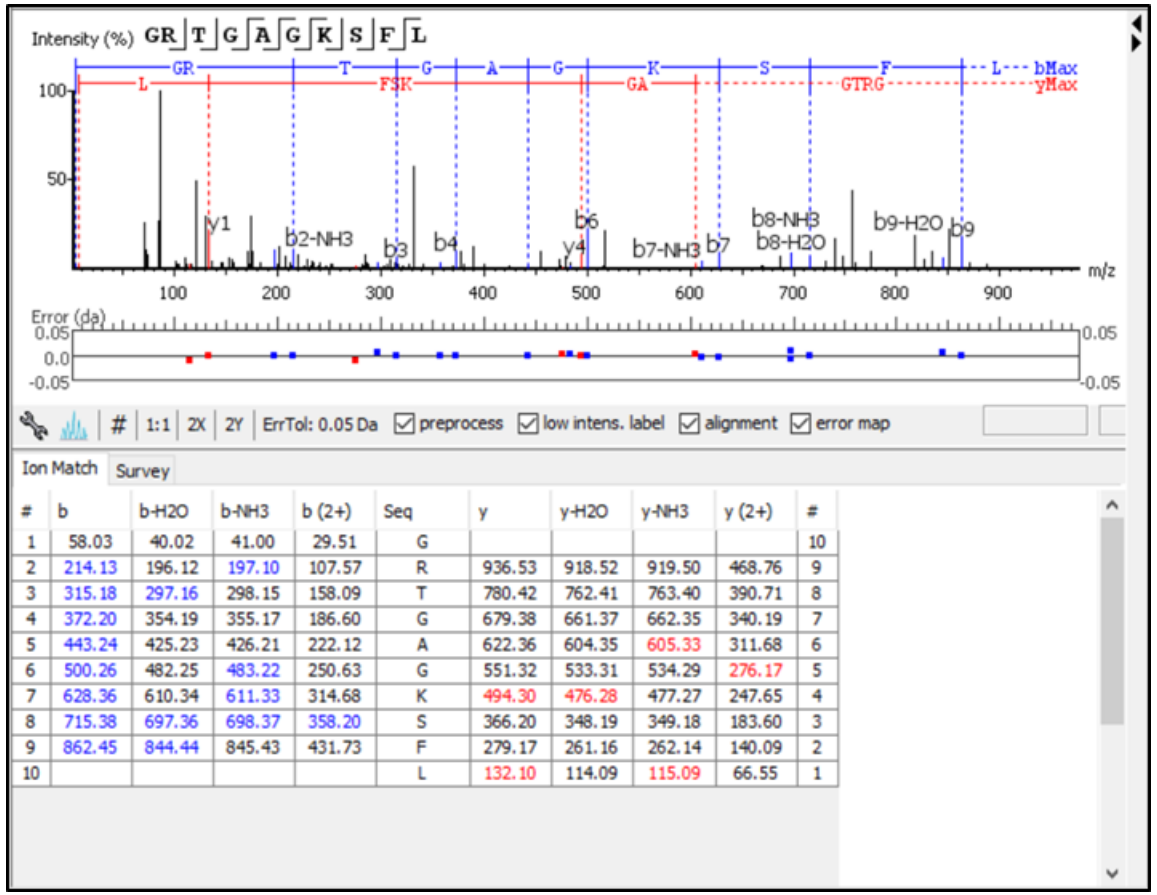
DeepNovo



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	129.10	111.09	112.08	65.05	K					9
2	242.19	224.18	225.16	121.59	L	999.56	981.54	982.55	500.28	8
3	428.27	410.26	411.26	214.63	W	886.47	868.46	869.44	443.73	7
4	543.30	525.29	526.27	272.18	D	700.39	682.38	683.36	350.69	6
5	640.35	622.34	623.32	320.67	P	585.36	567.35	568.33	293.18	5
6	753.44	735.42	736.40	377.21	L	488.31	470.29	471.30	244.65	4
7	868.46	850.44	851.44	434.73	D	375.22	357.21	358.20	188.11	3
8	981.54	963.51	964.51	491.27	L	260.20	242.19	243.14	130.60	2
9					K	147.11	129.10	130.09	74.06	1

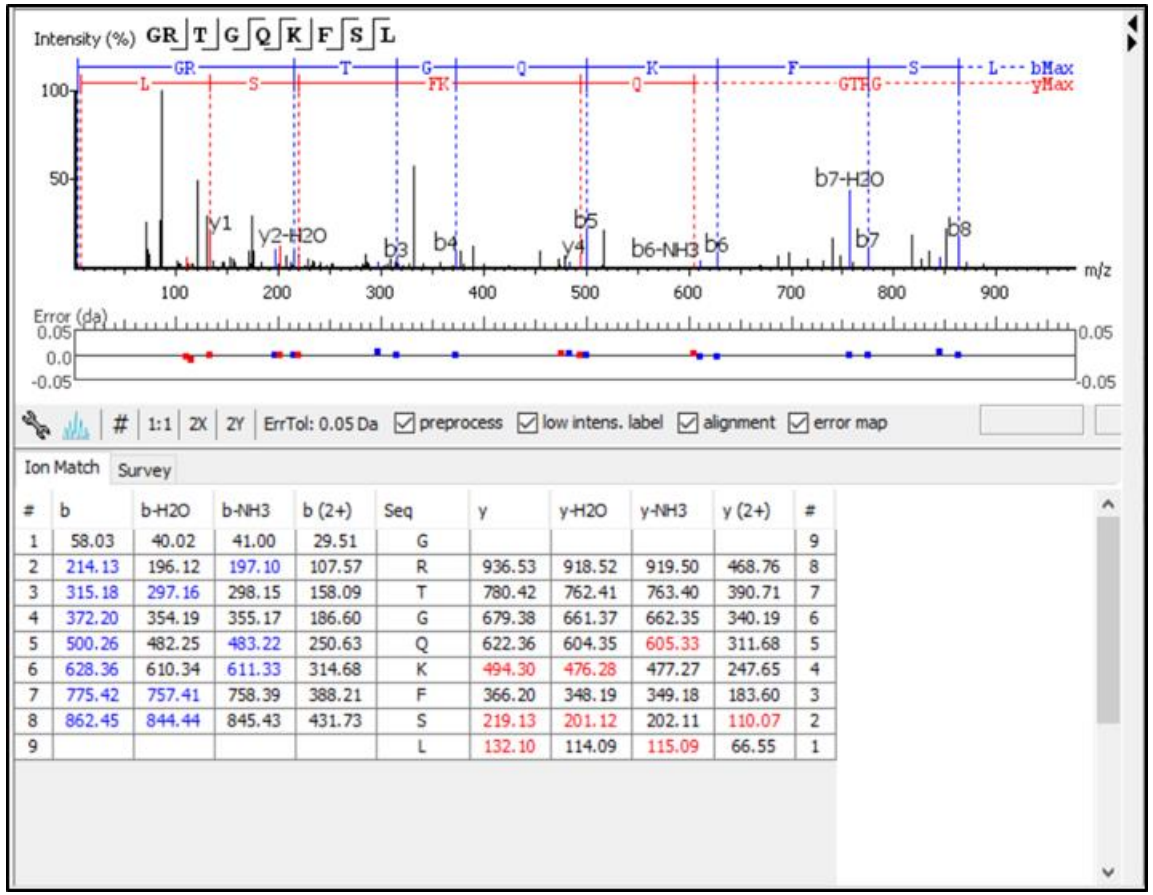
Fraction: 20141210_QEp7_MiBa_SA_HLA-I-p_MM15_2_B_1.raw
Scan ID: 21931
Retention time: 37.371
M/z: 331.854
Charge: 3

MaxQuant



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.03	40.02	41.00	29.51	G					10
2	214.13	196.12	197.10	107.57	R	936.53	918.52	919.50	468.76	9
3	315.18	297.16	298.15	158.09	T	780.42	762.41	763.40	390.71	8
4	372.20	354.19	355.17	186.60	G	679.38	661.37	662.35	340.19	7
5	443.24	425.23	426.21	222.12	A	622.36	604.35	605.33	311.68	6
6	500.26	482.25	483.22	250.63	G	551.32	533.31	534.29	276.17	5
7	628.36	610.34	611.33	314.68	K	494.30	476.28	477.27	247.65	4
8	715.38	697.36	698.37	358.20	S	366.20	348.19	349.18	183.60	3
9	862.45	844.44	845.43	431.73	F	279.17	261.16	262.14	140.09	2
10					L	132.10	114.09	115.09	66.55	1

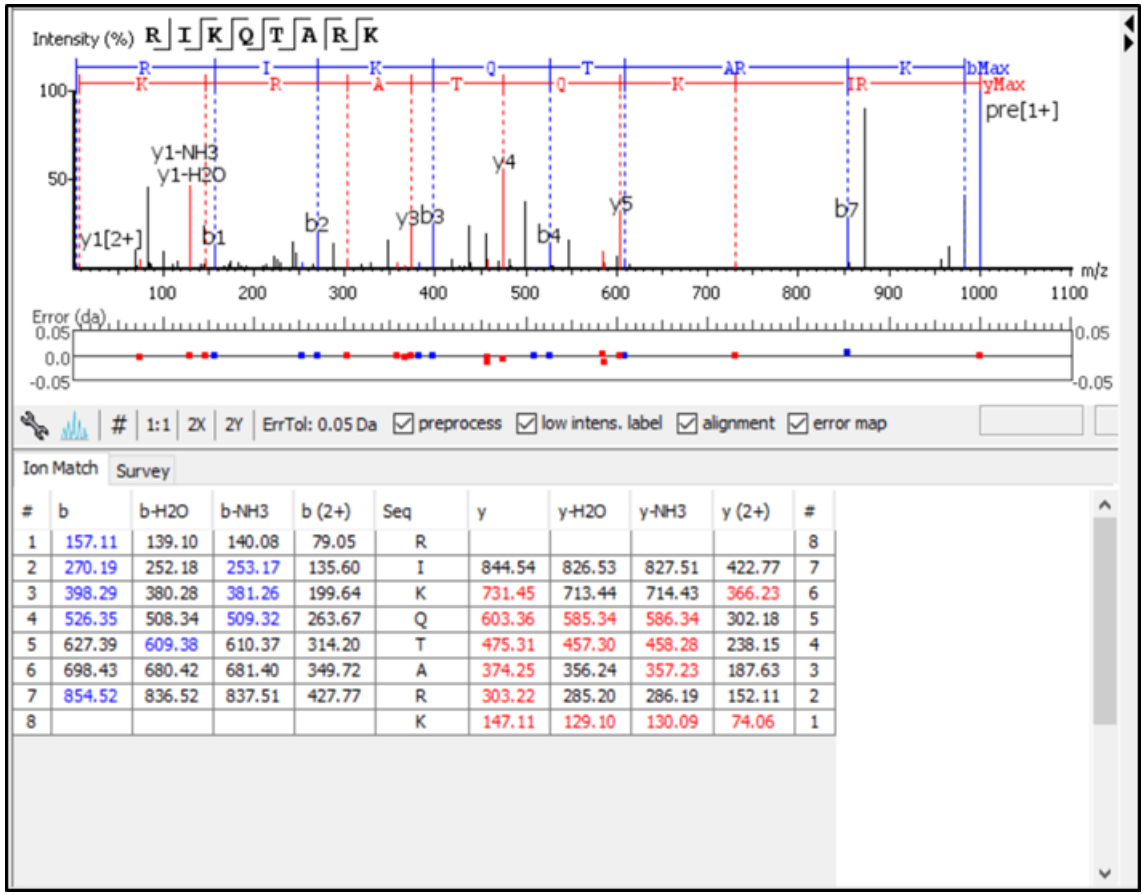
DeepNovo & PEAKS DB



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.03	40.02	41.00	29.51	G					9
2	214.13	196.12	197.10	107.57	R	936.53	918.52	919.50	468.76	8
3	315.18	297.16	298.15	158.09	T	780.42	762.41	763.40	390.71	7
4	372.20	354.19	355.17	186.60	G	679.38	661.37	662.35	340.19	6
5	500.26	482.25	483.22	250.63	Q	622.36	604.35	605.33	311.68	5
6	628.36	610.34	611.33	314.68	K	494.30	476.28	477.27	247.65	4
7	775.42	757.41	758.39	388.21	F	366.20	348.19	349.18	183.60	3
8	862.45	844.44	845.43	431.73	S	219.13	201.12	202.11	110.07	2
9					L	132.10	114.09	115.09	66.55	1

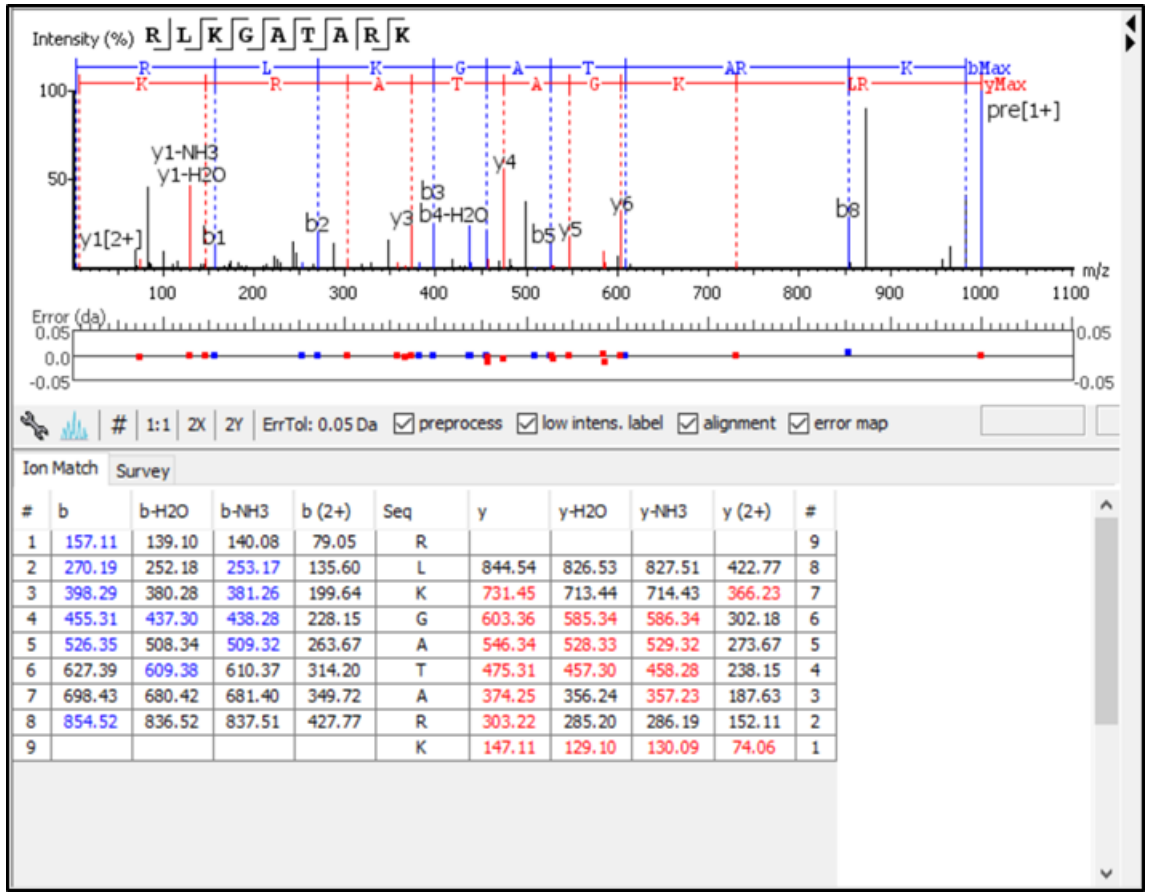
Fraction: 20141208_QEp7_MiBa_SA_HLA-I-p_MM15_3_B.raw
Scan ID: 2606
Retention time: 6.317
M/z: 334.217
Charge: 3

MaxQuant



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	157.11	139.10	140.08	79.05	R					8
2	270.19	252.18	253.17	135.60	I	844.54	826.53	827.51	422.77	7
3	398.29	380.28	381.26	199.64	K	731.45	713.44	714.43	366.23	6
4	526.35	508.34	509.32	263.67	Q	603.36	585.34	586.34	302.18	5
5	627.39	609.38	610.37	314.20	T	475.31	457.30	458.28	238.15	4
6	698.43	680.42	681.40	349.72	A	374.25	356.24	357.23	187.63	3
7	854.52	836.52	837.51	427.77	R	303.22	285.20	286.19	152.11	2
8					K	147.11	129.10	130.09	74.06	1

DeepNovo



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	157.11	139.10	140.08	79.05	R					9
2	270.19	252.18	253.17	135.60	L	844.54	826.53	827.51	422.77	8
3	398.29	380.28	381.26	199.64	K	731.45	713.44	714.43	366.23	7
4	455.31	437.30	438.28	228.15	G	603.36	585.34	586.34	302.18	6
5	526.35	508.34	509.32	263.67	A	546.34	528.33	529.32	273.67	5
6	627.39	609.38	610.37	314.20	T	475.31	457.30	458.28	238.15	4
7	698.43	680.42	681.40	349.72	A	374.25	356.24	357.23	187.63	3
8	854.52	836.52	837.51	427.77	R	303.22	285.20	286.19	152.11	2
9					K	147.11	129.10	130.09	74.06	1