Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 20 5 2019
Manuscript Category

OXFORD

Genome analysis

Statistics for Proteomics: the possible pitfalls when analyzing spliced peptides

Bianca De Saedeleer [1,*], Lotte Pollaris [1*],Sam Dierickx [1,*], Lieven Clement [2,5,a], Lennart Martens [3,4,5,a]

[1]Faculty of Bioscience Engineering, Master of Science in Bioinformatics, Ghent University, Ghent, Belgium.
[2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium.
[3]VIB-UGent Center for Medical Biotechnology, Ghent, Belgium.
[4]Department of Biochemistry, Ghent University, Ghent, Belgium.
[5]Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium.

*,These authors contributed equally to this work
a, whom correspondence should be addressed

Associate Editor:

Received on 16 may 2019

Abstract

Motivation: The composition of the immunopeptidome would be valuable information in the research for the development of new vaccines and immunotherapies against autoimmunity, infectious diseases and cancers. A recent study [18] reports the identification of 19.3 to 19.6% of the immunopeptidome to be spliced peptides, which account for 23.6% of its diversity. This remarkable claim could mean a breakthrough in medicine, but is contraintuitive according to current insights in biology. We hypothesize that the analysis is not performed in the right statistical manner, violating crucial assumptions that need to hold in order to rely on the adapted theoretical principles, such as the Target-Decoy Approach (TDA) and the False Discovery Rate (FDR) procedure.
Results: In this project, the assumptions considering the representativeness of the decoys are checked as well as their abundance relative to the targets. A more valid construction of the decoys and an alternative and more reliable calculation for the FDR are proposed. In addition, the publicly available code of the authors is revised and improved. Starting from identical data from the human fibroblast HLA-I immunopeptidome, the analysis is repeated and the obtained results do not confirm the findings of Liepe et al.[18], only one spliced peptide could be reported in this paper compared to the 1154 by Liepe et al. [17].
Availability: All scripts and links to the data are available on
https://github.ugent.be/samdieri/Design_project_2019_Statomics
Contact: lieven.clement@ugent.be

1 Introduction

The 26S proteasome is a large protease complex that degrades all unneeded and damaged proteins into peptides by pulling the protein sequence through its 20S core and cleaving it through canonical peptide bond hydrolysis. This results in non-spliced peptides. In addition, the proteasome is also able to paste the non-spliced peptide sequences by proteasome catalyzed peptide splicing (PCPS), resulting in so called spliced peptides that are not found in the original protein sequence. If the fragments originate from the same molecule, cis-PCPS is performed, if they are derived from different molecules, the phenomenon is called trans-PCPS. It is acknowledged that cis-PCPS occurs the most, however, how this ligation is executed exactly by the proteasome is until now unknown [3, 9, 17]. Spliced and non-spliced peptides can function as epitopes that are part of an antigen (protein, peptide, or polysaccharide), to which an antibody, B-cell or T-cell

of the immune system binds, eventually inducing an immune response [5]. These peptides migrate from the cytosol to the lumen of the endoplasmic reticulum by a transporter associated with antigen processing. There, the major histocompatibility complex (MHC) molecules, which are called Human Leukocyte Antigen (HLA) class I molecules in humans, will bind to the peptides of 9 to 12 amino acids (AA) long, and transport the epitopes from the cytosol to the cell surface. If these displayed peptides are foreign, for example from a pathogen, virus or tumor, the T-cells (CD8+ white blood cells containing a T-receptor) will recognize them and will destroy these cells [17, 26]. Such epitopes are interesting to discover, because they could help in the development of new vaccines and immunotherapies against autoimmunity, infectious diseases and cancers. The collection of epitopes associated to the HLA's is called the immunopeptidome. Finding out its composition is a challenging problem and a hot topic as it could mean a breakthrough in the medical field [3, 6, 17, 20, 26]. The proteasome is able to generate spliced peptides, which would increase the variety within the immunopeptidome [19]. Nonetheless, the relevance and frequency of these peptides is unknown. Considering that the immune system needs to distinguish foreign and cell specific epitopes, it seems that spliced peptides would be too random for this recognition. Intuitively, spliced peptides are expected to be low abundant in the immunopeptidome and are generally seen as byproducts of the main proteasome activity to manage cellular protein homeostasis and AA recycling [26]. It is proven that PCPS does not happen completely random. Berkers et al. [3] indicate the existence of certain splicing rules, but further investigation is needed. In order to perform PCPS, the fragments need to have a minimum length of 3 AA and contain a C-terminal nucleophile. The latter influences, along with many other factors such as the distance between the two fragments that will be pasted into a spliced peptide, the PCPS efficiency [9]. This efficiency is low [9, 25] and has been measured to be maximum 0.01%, for spliced peptides detected on the cell surface of HLA class I molecules. These initiate an immune response, indicating a possible role of spliced peptides as immunogenic epitopes [3]. Nevertheless, it is not confirmed that spliced peptides would be more immunogenic or more preferred than non-spliced peptides. Furthermore, the low peptide splicing efficiency contradicts a large presence in the immunopeptidome [26].

Detection of peptides is mostly performed by use of cell surface elution and mass spectrometry (MS). Nonetheless, analyzing the immunopeptidome using MS is complex due to technical limitations in the isolation and purification steps as well as the lack of uniformly defined computational standards. Despite that MS is not directly applicable for quantitative comparison of single peptides, it is approved for large datasets [6, 20, 22]. The obtained spectra are matched against theoretical ones from peptide databases [3] using search engines such as MaxQuant and MS-GF+ [15, 24]. However, these databases only contain linear fragments of the proteome, making it hard to observe spliced peptides and to determine their diversity and whether spliced peptides are abundant in the immunopeptidome or play a minor role [26]. Due to the unknown mechanism of the proteasome and its ability to create spliced peptides, the required databases will be large. Filtering steps are needed to make the searches feasible [26].

In contrast to previous studies [9, 19, 25] which manage to identify maximum five spliced peptides, Liepe et al. [17] claim that the spliced peptide pool represents for 25% of the HLA class I immunopeptidome in terms of abundance and 21 to 32% in terms of variety. They identify 6592 non-spliced and 3417 spliced 9-12 mer peptides in the HLA-I immunopeptidome of the GR lymphoblastoid cell line, and 1154 spliced and 3039 non-spliced 9-12 mers in the human fibroblast HLA-I immunopeptidome. They state that this discovery increases the amount of known antigens in general by 34% and the number of identified HLA-I ligands by 50%. Mylonas et al. [21] invalidate these claims by de-novo sequencing. They also suspect violations of the required decoy assumptions, but this is not assessed with subsequent analysis. In 2018, Liepe et al. [18] published a follow-up paper based on their conclusions of their study in 2016. Here, they perform some additional filtering steps, leading to the conclusion that the immunopeptidome contains 19.3 to 19.6% spliced peptides which account for 23.6% of the diversity. Despite the fact that this is still a high abundance considering the biological concerns, the revised conclusions are remarkably lower than previously stated. Moreover, Liepe et al. [18] report a False Discovery Rate (FDR) of 2-4% for the spliced and 1% for the non-spliced peptides.

Due to the biological contradiction of this claim, i.e. the appearance of spliced peptides is expected to be a rare event, the aim of this project is to repeat the searches of Liepe et al. [17] and examine if the analyses are statistically correct and to verify the reliability of the reported conclusions by Liepe et al. [17]. Presumably, the authors do not check the assumptions, the distribution of the decoy hits might not be representative for the distribution of the bad target hits, because the decoys do have a too random AA distribution and the amount of decoys is dissimilar to the amount of targets (Figure 1, 1a and 1b). This would have given rise to incorrect FDR controls, which can be checked using the Target-Decoy Approach (TDA) (Figure 1, 2). Because only the scripts for making the initial search space are publicly available and scripts or a clear explanation for reconstructing the decoy database are lacking, the study of Liepe et al. [17] becomes irreproducible. We will first build the decoy database in the right statistical and biological manner, taking into account the above stated assumptions, making sure the decoys are a good simulation of the bad target hits. Second, we will perform a FDR control on the output of a search engine, as this calculation will be influenced by the properties of the decoys, taking into account that we are mainly interested in a subset of the data. Third, we will rewrite and improve the code in order to get better results and shorten the computational time (the publicly available scripts have a running time of weeks to months, depending on the capacity of the used computer) (Figure 1, 3a and 3b).

2 Materials and Methods

2.1 Pipeline of Liepe et al. [17]

For the detailed pipeline, Liepe et al. [18] refer to their paper of 2016 [17] as they do not change their strategy in the most recent study. Therefore, the concerns expressed in this paper will mainly be based on the paper of 2016. From the entire human proteome database (Swissprot), only the 9-12 mer spliced and non-spliced peptides with a maximum intervening length of 25 AA are used to generate the databases. Additionally, peptides are filtered on the molecular weights (MW) of interest, i.e. from peptides present in the immunopeptidome. An error of ± 6 ppm is allowed. The target database is built by merging the spliced and non-spliced peptide databases and transforming them into a FASTA format with a structure that mimics the human proteome. To construct the decoy database, the target sequences are randomized. Using Mascot, the peptides are matched against MS/MS spectra and then filtered based on their Mascot ion score at a 1% FDR cutoff, resulting in a ranked list of potential peptide sequences. In the latest publication of Liepe et al. [18] an additional filtering step is performed where if the first ranked peptide is non-spliced, that peptide is assigned to the MS/MS match. If a spliced peptide is ranked first, the authors check if a non-spliced peptide with a lower rank occurs among the matches and whether the ion score difference between the spliced and non-spliced peptide is less than 30%. When this is the case, the non-spliced peptide is considered as a true hit. If the ion score difference is larger than 30%, the non-spliced peptide is not assigned and the first ranked spliced

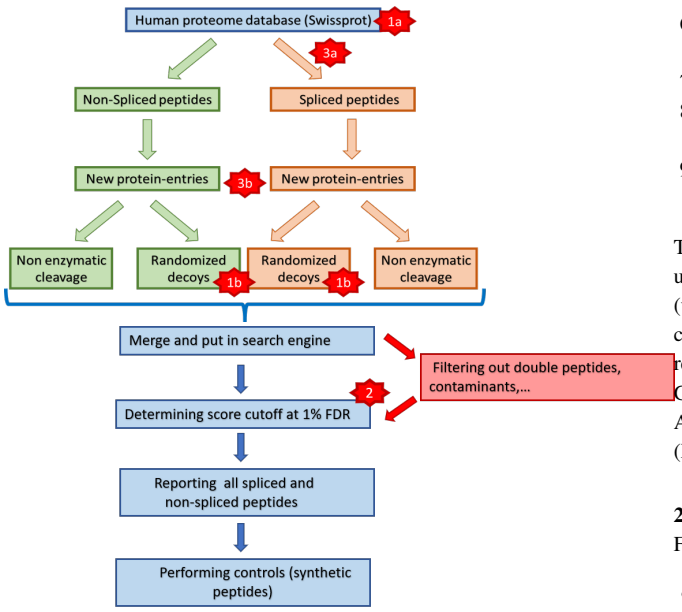

Fig. 1. Schematic overview of the pipeline of Liepe et al. [17]. A more detailed explanation on the pipeline is given in section 2.1. The red stars indicate where possible mistakes are made. (1a) The use of Swissprot instead of Uniprot leaves out possible good targets, making it likely that some decoys are in fact targets or that some non-spliced peptides are wrongly considered to be spliced peptides. (1b) The decoys are randomly created, which could lead to decoys not being a good representation of the targets; (2) The wrong formula is used for the FDR calculation and some of the assumptions are perhaps violated; (3a) The spliced peptides are inefficiently made leading to computational burden. (3b) The recreation of new protein entries is unnecessary, causing longer running time and inducing possible mistakes.

peptide is compared to another spliced peptide in the list with the lowest possible rank and an ion score difference of at least 30%. If such a spliced peptide occurs in the list, the first ranked spliced peptide is assigned to the MS/MS match. The final obtained list contains identified non-spliced peptides and assigned spliced peptide candidates. To confirm the found spliced peptides, the authors create synthetic peptides from the obtained spliced spectrum matches and put these through the mass spectrometer in order to construct real spectra of these synthetic peptides and match them against the theoretical spectra of interest. Another control is performed by partially flipping the sequences of the decoys. Further elaboration on the introduced controls can be found in section 4.

2.2 Reconstruction of the Search Space

Since an unambiguous strategy and code for reconstructing the decoy database is lacking and the provided scripts (https://data.mendeley.com/datasets/y2cvb5nvgn/1) by the authors for building the target database are computationally inefficient, we propose a new algorithm to create the target and decoy database and to merge these together into the proper search space containing the non-spliced and spliced peptides. The following strategy is proposed:

1. Make the non-spliced peptides (non-spliced targets) based on the input settings chosen by the user.
2. Remove duplicate non-spliced peptides.
3. Form a spliced peptide (spliced target) by pasting two non-spliced peptides, created in the previous step, together. Repeat this for all combinations considering the input settings.
4. Remove duplicate spliced peptides.
5. Filter the peptides based on their MW.
6. Remove peptides that occur in the non-spliced as well as in the spliced database.
7. Create the decoys by reversing all remaining peptide sequences.
8. Remove decoys that occur in the target (spliced and non-spliced) database.
9. Merge target and decoy databases together into a complete search space and store in FASTA file.

The algorithm is written in R (v3.5.1) and the following packages are used: Seqinr (v3.4-5), Tidyverse (v1.2.1), Stringi(v1.2.4), Arrangements (v1.1.5), Parallel (v3.5.1) and MSnbase (v2.8.3). The full annotated code (prep_sp.R, compute_sp_function.R, terminal_make_sp.R and removeduplictes.R) with accompanying information can be found on GitHub (https://github.ugent.be/samdieri/Design_project_2019_Statomics). All searches are run on an Ubuntu server (v18.04.2 LTS, 18 cores (Intel (R) Xeon (R) CPU ES-2420 v2 @2.20GHz), 22GB RAM).

2.2.1 Input Settings

Following inputs need to be provided by the user:

- k-mer lengths of interest
- database containing the protein sequences of which the fragments originate
- mgf file with the mass spectrometry data
- allowed number of AA between two fragments that generate a spliced peptide

To build the desired search space, the same inputs are chosen as in the discussed paper: kmers of length 9, 10, 11 and 12 AA, the Swissprot human fasta file (HUMAN.fasta) and the MS data from the human fibroblast HLA-I immunopeptidome (20130504_EXQ3_MiBa_SA_Fib-2.mgf file) from Bassani-Sternberg et al. [1], provided in the PRIDE Archive (https://www.ebi.ac.uk/pride/archive/projects/PXD000394/files;jsessionid=3C4BED0B6A64BF2AE929529A16E72E3E) and a maximal permitted distance between the fragments of 25 AA.

2.2.2 Computation of the Index Matrix

The search space consists of non-spliced and spliced peptides. First, the non-spliced peptides are made, which are then pasted together to form the spliced peptides. This is done by computing index matrices that contain the required indices in order to make such peptides. The non-spliced peptides index matrix contains all start and stop positions of the different fragments that can be made with the given input settings and its size is determined by the largest protein of the given FASTA file. This allows the reuse of the matrix for smaller proteins by selecting the desired start-stop combinations. In order to compute the matrix, all fragments of length 1 to 11 AA are made. The start and stop indices of these fragments are generated by drawing two natural numbers from the interval that ranges from 1 to the length of the longest protein, such that the desired total length of a spliced peptide can be obtained by pasting two non-spliced peptides together. Because RAM memory is limited, the calculation is done in bins of 1100 AA. The first bin starts with one and the last bin ends with the length of the protein. Every bin contains a new part of the protein sequence with a small overlap with the previous bin. This overlap is the length of the longest k-mer fragment minus one, which allows the use of bins without losing possible fragments. The new minimal starting index of the current bin will be:

$$Start_{current} = End_{previous} - k + 1 \tag{1}$$

with $Start_{current}$ the start position of the current bin, $End_{previous}$ the end position of the previous bin,k the considered k-mer.

This will create duplicate fragments, which are filtered out after all fragments are generated. The remaining couples are separated based on the length of the peptide they represent. This results in a list with for every k-mer (1-12) a database with corresponding fragments.
After obtaining all start-stop couples of the non-spliced peptides, spliced peptides with length 9 to 12 are generated by a permutation. They are represented as two start-stop couples, which refer to the two non-spliced peptides that need to be pasted together to form the spliced peptide. These combinations of start-stop couples are generated by considering all combinations within two databases of the non-spliced list. The first database is one that contains all fragments of a certain length n. The second database contains all fragments of length m in order to get the total length of the k-mer of interest. All position couples of the two non-spliced peptides that can form the k-mer are allocated in the spliced peptide index matrix. The permutation is run in smaller bins of 1100 AA to save RAM space. The separation into bins is based on the start and stop value of the fragments, both have to be between certain border values. The new border values are based on the previous values obtaining again a small overlap.

$$Start_{current} = End_{previous} - gap - k + 1 \tag{2}$$

with $Start_{current}$ the start position of the current bin, $End_{previous}$ the end position of the previous bin, gap the maximal allowed distance (in AA) between two fragments and k the considered k-mer.

In each bin, all fragment combinations that do not meet the constraints are removed. To obtain all spliced peptides for a certain k-mer this procedure is repeated so that every database of the non-spliced peptides has been selected as first database. Finally, duplicate peptides are removed from the database. This procedure creates all spliced peptides for one specific k-mer and is repeated for every k-mer of interest, resulting in a list of databases, with for every k-mer a database of two start-stop couples representing the spliced peptides.

2.2.3 Calculation of the Mass Matrix

To be able to filter the peptides based on their molecular weight, a mass matrix is computed. The mass matrix contains the MW of the precursor fragments obtained from the MS data. The MW is computed using following formula:

$$MW = mz \cdot z \tag{3}$$

with MW the molecular weight, z the charge of the peptide fragment and mz the mass-to-charge ratio. A precursor tolerance of 10 ppm on the m/z-ratio is usually used in proteomics [27].
The error on the MW depends on the charge of the precursor and is calculated with this formula:

$$Error_{MW} = Error_{mz} \cdot z \tag{4}$$

with $Error_{MW}$ the error on the MW, $Error_{mz}$ the error on the m/z-ratio and z the charge of the peptide fragment. This will result in a matrix containing different MW intervals for every data point $\pm Error_{MW}$.

2.2.4 Construction of the Search Space

After computing the index and mass matrices, the search space can be made. First, the start-stop couples from the index matrix are translated into the corresponding peptide fragments. Then, the MW of every fragment is computed. The peptides are filtered on their MW, only if the MW is within one of the intervals in the mass matrix, the peptide fragment is kept. For the spliced peptides an additional filter is used to ensure all peptide fragments that also occur in the non-spliced database are removed.
To create the decoys, all remaining peptide fragments are reversed (standard procedure to create decoys [10]). This procedure results in four databases: non-spliced targets, spliced targets, non-spliced decoys and spliced decoys. The obtained sequences are then written to separate FASTA files, according to the database they originate from. Finally, the decoys need an additional filtering step: decoy sequences that occur in the non-spliced as well as in the spliced database are removed. This is done with "gt -sequniq" function of GenomeTools (v1.5.9).

2.3 Performing the Searches

In order to match the observed MS spectra against theoretical spectra from a certain peptide database and to identify and quantify the peptides, search engines such as MaxQuant and MS-GF+ can be used. These engines assign scores to the matching peptides for which a higher score means a better or more precise match [15, 24]. Liepe et al. [17] use the engine Mascot, which is not open source, and therefore not an option for this project.

2.3.1 MS-GF+ Search

A first explorative analysis is done using MS-GF+ Beta (v10072) on the non-spliced peptides from the human fibroblast HLA-I immunopeptidome generated by Bassani-Sternberg et al. [1]. This mgf file contains the non-centered spectra, in order to perform the search, the file is first centered using MSConvert (v3.0.18288-56376f01b). The same parameters as mentioned by Bassani-Sternberg et al. [1] are used, except for the minimal and maximal allowed length, which are set at 9 and 12 AA, respectively. Two searches are performed: one where the search engine itself builds the decoys (by reversing the targets) and one using the decoys generated by applying the proposed strategy of section 2.2. The output of an MS-GF+ search is an mzid file, which needs to be converted into an tsv extension in order to be read in in R for further evaluation of the targets and decoys. Scripts for the search are available on GitHub (cmdMSGFPlus.R and FDR_MSGF.Rmd). If the search engine generates the decoys itself, the size of the decoy database is not known, and this information is needed in order the estimate the FDR with a correction factor for the imbalance in size between the target and decoy databases (see section 2.4). Therefore, a search should be executed on the merged search space containing the spliced and non-spliced targets and decoys built with the procedure described in section 2.2.

2.3.2 MaxQuant Search

To validate the findings from the MS-GF+ search, a MaxQuant (v1.5.8.3) search is performed on the non-centered spectra with the same settings as reported by Bassani-Sternberg et al. [1], but with the FDR set to 1 in order to retrieve all the decoys. Again, the search engine itself generates the decoys by reversing the targets. One of the generated output files of MaxQuant is the evidence.txt file, which is read in in R for further evaluation on the targets and decoys (FDR_MaxQuant.Rmd on GitHub). Subsequently, a search is performed on the merged search space made by using the proposed strategy (section 2.2) in order to compare our results to those of Liepe et al. [17]. The FDR_MQ_FULL.Rmd provided on GitHub contains the code for the executed analysis including an all-all, all-sub and sub-sub approach which will be explained further in section 2.4.5.

2.4 Theoretical Principles

2.4.1 Estimating the False Positives using the FDR

The main goal of using mass spectrometry in proteomics is to report as much correct and as less incorrect peptide identifications as possible. When matching peptide spectra to theoretical spectra (peptide-to-spectrum matches (PSMs)), correct and incorrect identifications will occur [10]. It is crucial to measure the uncertainty on the reported PSMs. When reporting results, a balance between the amount of False Positives (FP) and True

Positives (TP) needs to be maintained. Ideally, all TPs will be reported, but this implies a high number of FPs among the reported positives (all hits), while it is desired to minimize the amount of FPs. Hence, a certain amount of FPs will be allowed in order to report enough TPs. Since the real proportion between FPs and TPs is unknown, statistics are needed to control this amount of FPs, and thus to measure the uncertainty on the results. An approach to control this uncertainty is the False Discovery Rate (FDR)[2]. The FDR is defined as the proportion of expected FPs on all hits kept:

$$FDR = E\left[\frac{FP}{FP+TP}\right] \tag{5}$$

This method is validated for proteomics [14], and used by Liepe et al. [17] as well. In proteomics, this method is used to determine a certain score threshold. All PSMs with a score above this threshold will be accepted as a hit. Among these accepted hits, the expected fraction of FPs is on average equal to a certain value, defined in the FDR. In this research, the FDR is controlled at 1%, which allows 1% FPs among the reported positives. In order to make use of the FDR, the distributions of the FPs and all hits, are required. Based on the output of a search engine, the total distribution can be estimated by linking the PSMs to their scores. The distribution of the FPs is more difficult, as it is not known which matches are correct. To simulate this FP distribution, the target-decoy approach (TDA) is often used.

2.4.2 The Competitive Target-Decoy Approach

To estimate the distribution of the bad hits (FPs), the competitive target-decoy approach can be used. In this approach, a search is performed against a search space containing targets (real peptides of interest) and decoys (non-existent sequences that are not targets). Peptides that match against the nonsense decoy database are assumed to estimate the distribution of the bad target hits. In the conventional competitive TDA, the crucial assumption holds that bad hits are equally likely to match a decoy than a target [10]. If this is not the case, the distribution of the decoys will not be a good simulation of the distribution of the bad target hits, leading to an invalid FDR control. At least, the following constraints need to be satisfied, in order to obtain similar distributions of the decoys and bad target hits:

1. Similar amino acid distributions as the target protein sequences.
2. Similar amount of predicted peptides as the target protein list.
3. No predicted peptides in common between target and decoy sequence lists [11].

The first constraint makes sure that the decoy database is a good simulation of the bad target hits. If the decoy database would be too random, bad hits wouldn't be equally likely to match the decoy or target database. Not all AA compositions of the proteins from where the peptides present in the immunopeptidome originate, are known, but these will most likely be similar to the known proteins. Also, some orders of AA are more likely than others [12, 16]. If the decoy database contains a lot of AA that are rare in the organism, PSMs would not score high on these peptides, resulting in a shift of the decoy distribution towards the left. This way the distribution wouldn't be a good representation of the FP distribution. As it is impossible to know the magnitude of the shift, it is impossible to estimate a good bad hit distribution.
The second constraint considers balanced database sizes, so that a bad hit will have an equal probability to match a target as a decoy. If database A is twice as big as database B, the chance to match randomly a peptide in database A is twice as high as for database B (assuming that the first assumption is met). The shape of the distribution of the decoy database would still be similar to the one of the bad target hit database, but the area would be smaller. A correction is needed if this constraint is not fulfilled, which will be elucidated in subsection 2.4.4.

The third constraint reasons that a certain sequence that is present as a target, is not a nonsense peptide, because it has a meaning in the context of the experiment. Matching this peptide could be a good hit, and good hits shouldn't be present in the decoy database. Peptides occurring in both the target and decoy database complicate the case. If, for example, a target peptide present in the decoy database provides a good hit, it can get quite a high score, and thus influences the distribution. The distribution on the decoys would shift more to the right than the distribution of the bad target hits. Removing all targets from the decoy space is therefore a solution.

2.4.3 FDR and TDA in formulas

Based on the Supplementary Materials of Sticker et al. [23] and the course notes of Statistical Genomics by L. Clement [7], the FDR in the TDA is estimated as follows if the assumptions on the decoy database are fulfilled:

$$\widehat{FDR}(x) = \frac{\#decoys|x \geq t}{\#targets|x \geq t} = \frac{\hat{\pi}_0 \cdot \bar{F}_0(x)}{\bar{F}(x)} \tag{6}$$

with x the PSM scores, t a certain threshold to control the permitted amount of FPs, $\hat{\pi}_0$ $(= \frac{\#decoys}{\#targets})$ the estimated probability of a hit being a bad hit (a FP), $\bar{F}_0(x)$ $(= \frac{\#decoys|x \geq t}{\#decoys})$, and $\bar{F}(x)$ $(= \frac{\#targets|x \geq t}{\#targets})$ the empirical complementary cumulative distribution functions (ECCDF) of the decoys and targets, respectively. By using this formula it is clear that when the distributions are known, an estimation of the FDR can be made.

2.4.4 Corrected π_0 and FDR estimate

If the sizes of the decoy and target databases differ, the FDR cannot be estimated using equation (6), but a correction needs to be introduced. Let $\#B$ be the total amount of bad hits (against the targets and decoys), $\#D$ the number of decoys, $\#T$ the number of targets, and $\#T_B$ the number of bad hits against the targets of which the probability is equal to π_0. The chance of having a bad hit against a decoy is π_D.

$$\#B = \#D + \#T_B = \#D + \pi_0 \cdot \#T \tag{7}$$

$$\#D = \pi_D \cdot \#B \Rightarrow \#B = \frac{\#D}{\pi_D} \tag{8}$$

Combining (7) and (8):

$$\#T_B = (1-\pi_D) \cdot \#B \Rightarrow \pi_0 \cdot \#T = (1-\pi_D) \cdot \frac{\#D}{\pi_D} \tag{9}$$

Which leads to

$$\pi_0 = \frac{(1-\pi_D)}{\pi_D} \cdot \frac{\#D}{\#T} \tag{10}$$

π_D can be estimated as

$$\hat{\pi}_D = \frac{Size_{decoys}}{Size_{decoys} + Size_{targets}} \tag{11}$$

Then

$$\hat{\pi}_0 = \frac{Size_{targets}}{Size_{decoys}} \cdot \frac{\#D}{\#T} \tag{12}$$

Applying this on (6):

$$\widehat{FDR}(x) = \frac{Size_{targets}}{Size_{decoys}} \cdot \frac{\#decoys|x \geq t}{\#targets|x \geq t} \tag{13}$$

2.4.5 All-Sub Approach by Sticker et al. [23]

The goal of the paper by Liepe et al. [17] is to evince the abundance of spliced peptides. In this research, these spliced peptides are the main set of interest and the non-spliced peptides are of less importance. When calculating the FDR on both spliced and non-spliced peptides together, and

reporting them separately afterwards, the uncertainty on the results can differ between the different subgroups, giving rise to possible mistakes. This can lead to a lower FDR in one subgroup than in the other.
In most peptide identification cases a search-all-assess-all method (all-all) is used to calculate the FDR for the obtained PSMs. However, this strategy often leads to an overestimation of the FDR and thus an unreliable FDR control. Some researchers advocate for a search-subset-assess-subset approach (sub-sub) where the parts that are not of interest are removed from the dataset. However, depending on the present sequences in the dataset, there is a possibility that some sequences are forced to match against the wrong target or decoy sequence, harming the assumptions of the TDA. In this case many PSMs will score low, which is visualized by a steep increase of the score distribution (Figure 1 of [23]). These switched low scoring PSMs will have higher scores in the complete database, implying that they have a higher chance to be FPs. The advantage of this approach is that the search is kept quite small, and thus more identifications are possible. A search-all-assess-subset (all-sub) approach, searches against all expected peptides and only keeps the PSMs of interest for further FDR calculation (based on TDA). This would force less spectra towards the wrong distribution without influencing the further data analysis and providing a more valid FDR control [23]. All three strategies will be performed in this analysis and the obtained results will be compared to those of Liepe et al. [17] who used the all-all approach.

Table 1. The amount of targets and decoys generated by using the proposed code and after filtering out the targets in the decoys.

Peptide	Amount after filtering
non-spliced targets	315 711
non-spliced decoys	312 480
spliced targets	52 832 999
spliced decoys	52 501 066

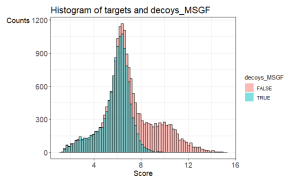

Fig. 2. The score distributions of target (red) and decoy (green) hits after a MS-GF+ search considering the non-spliced search space (using FDR_MSGF.Rmd). A clear overlap between the targets with a low score (the bad target hits) and the decoy hits is observed.

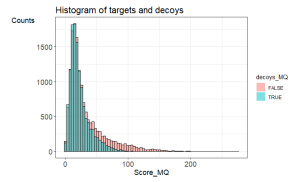

Fig. 3. The score distributions of target (red) and decoy (green) hits after a MaxQuant search considering the non-spliced search space (using FDR_MaxQuant.Rmd). A clear overlap between the targets with a low score (the bad target hits) and the decoy hits is observed.

3 Results

In order to measure the uncertainty on the obtained results and get insight into the proportion of expected FPs among them, the distribution of the bad target hits needs to be simulated by use of the decoy distribution. If these distributions coincide, the decoy distribution simulates the bad hits correctly, and π_0 can be estimated, which is required to control the FDR. For this, at least the three assumptions described in section 2.4.2 need to hold. In addition, if the amount of decoys is not equal to the amount of targets, an correction factor need to be introduced in order to estimate π_0 correctly. In this analysis, a search space is build containing the non-spliced and spliced targets and decoys by applying the proposed procedure described in section 2.2. First, a sub-sub approach on the non-spliced peptides only is performed to get used to the strategy and set everything into place for the complete search, because the complete search is computationally expensive. Then, an all-all and all-sub approach on the search space containing the spliced and non-spliced peptides is executed, the assumptions are checked and the obtained results are compared to those of Liepe et al. [17].

3.1 Representative Decoys

Using the suggested approach in section 2.2, the spliced and non-spliced targets are created. In order to make sure that the decoys contain the right amino acid composition, decoys are generated by reversing the target sequences. Additionally, decoys that are identical to the present targets, are removed, which results in the quantities shown in Table 1. The amount of non-spliced decoys is approximately equal to the amount of non-spliced targets (ratio of non-spliced targets to non-spliced decoys is 1.010), as well as for the spliced peptides (ratio of 1.006). This means that $\hat{\pi}_0$, the estimated chance of a hit being a bad hit, can be estimated by the amount of decoy hits on the amount of target hits. In this case no correction for differences in database sizes is required in the TDA.

To evaluate the search space, the search engines and the output, some searches are performed on the non-spliced peptides database only (no recombination into spliced peptides has to be executed, resulting in a smaller search space). First, a search is performed using MaxQuant and MS-GF+, where the target and decoys sequences are created by the search engines themselves without filtering based on the molecular weights. Then, the searches are repeated, but with the search space generated using the proposed strategy (section 2.2). Figures 2 and 3 show that the distribution of the non-spliced decoys coincides well with the distribution of the bad target hits, meaning that the decoys are a good simulation of the bad target hits and that the assumptions of the TDA are not violated. The FDR can be calculated in the correct statistical manner. However, a big difference can be noticed between the output of the different search engines. In Figure 2 (MS-GF+) a second peak of the target distribution is observed, showing that MS-GF+ is able to handle the big search space well, as it can make a good distinction between the good and bad hits, which results in more peptides that can be reported. This indicates that MS-GF+ is more suitable for this kind of analysis. Comparing to Figure 3, it is clear that MaxQuant suffers more from the size of the search space, leading to less peptides that can be reported.
Subsequently, a search is performed on the merged search space containing the generated spliced and non-spliced targets and decoys together. A search using MS-GF+ is not executed, because of an out-of-memory error (the search space is 6GB). Figure 4 displays the score distribution obtained from a MaxQuant search. This plot indicates that the decoy hits are almost completely coinciding with the target hits, meaning that the decoys are a good simulation of the bad target hits, and implying a lot of bad target hits.

The representativeness of the decoys can also be checked using PP-plots (probability–probability plots) where the empirical cumulative distribution function (eCDF) of the decoys is plotted against the eCDF of the subset target PSMs. Ideally, the decoy dataset from the complete search is representative for the distribution of the incorrect PSMs in the subset and can be used to estimate $\bar{F}_0$ in the FDR calculation [23]. Figures 5 and 6 visualize desired PP-plots. The first part of the curve is linear with a slope equal to π_0, illustrating that the decoy distribution and the incorrect PSMs of the target distribution coincide. Then, the line deviates towards higher percentiles due to the upper tail of the incorrect and the lower tail of the correct subset PSMs overlap. Finally, the line becomes vertically, because all the decoys have been observed (decoy percentile=1) before detecting all the targets. This profile shows that the subset of decoys as well as the

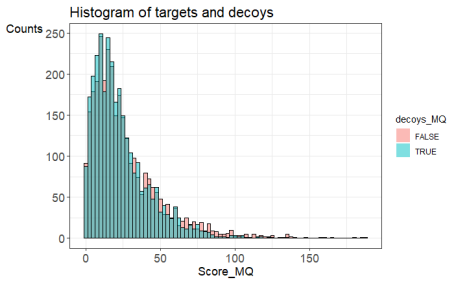

Fig. 4. The score distributions of spliced and non-spliced target (red) and decoy (green) hits performed with MaxQuant (using FDR_MQ_FULL.Rmd). A clear overlap between the targets with a low score (the bad target hits) and the decoy hits is observed.

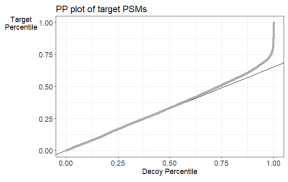

Fig. 5. PP-plot of the outcome of the MS-GF+ search on the non-spliced peptides (using FDR_MSGF.Rmd), the x-axis shows the percentage of decoys passed, the y-axis the percentile of targets passed at the same score. The straight line has a slope of π_0.

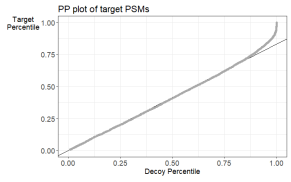

Fig. 6. PP-plot of the outcome of the MaxQuant search on the non-spliced peptides (using FDR_MaxQuant.Rmd), the x-axis shows the percentage of decoys passed, the y-axis the percentile of targets passed at the same score. The straight line has a slope of π_0.

decoys from the complete search are representative for the incorrect PSMs of the target distribution. It can be observed that the vertical tail starts earlier by the MS-GF+ search, indicating that this engine can report more peptides, as already seen in Figure 2. When the first part of the curve does not follow the π_0 line, but still is a straight line, it means that π_0 is wrongly estimated. This can occur when there is an imbalance between the amount of decoys and targets in the search space for which is not corrected with the proposed correction factor (section 2.4.4).

The PP-plot of the output of the MaxQuant search where the merged search space is considered (Figure 7), shows that the line of the target/decoys is slightly under the π_0 line. This indicates that the decoys have slightly higher scores than the targets, and that they are overly conservative. The vertically tail on the right is very small, and thus very few peptides can be reported.

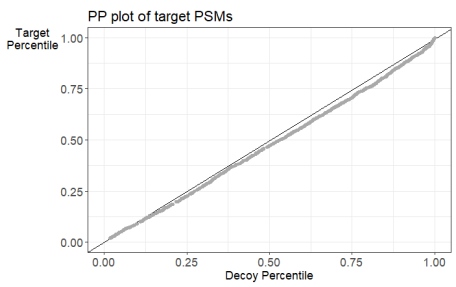

Fig. 7. PP-plot of the MaxQuant search outcome (using FDR_MQ_FULL.Rmd) on the merged search space (spliced and non-spliced peptides), the x-axis shows the percentage of decoys passed, the y-axis the percentile of targets passed at the same score. The straight line has a slope equal to π_0.

3.2 Reporting the Peptides of interest

Figure 8 displays an unequal amount of non-spliced targets that is reported at a 1% FDR using different methods. As mentioned in subsection 3.1, MS-GF+ handles the search space better and is able to report more non-spliced peptides. This shows the instability of the amount of reported peptides, and that the choice of the search engine and small adaptations in filtering, and cutoff influence the results a lot.

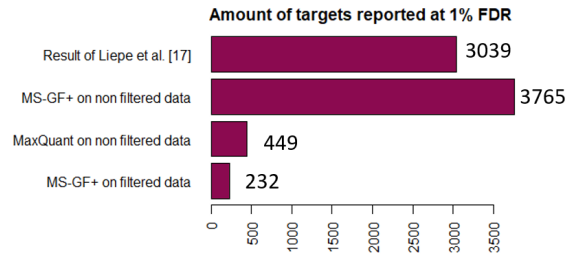

Fig. 8. The amount of non-spliced targets reported at a 1% FDR. The upper bar represents the amount of targets that Liepe et al. [17] report using Mascot. The second bar is the result of a MS-GF+ search (peptides created by the search engine, without mass filtering). The third bar is the outcome of a MaxQuant search (peptides created by the search engine, without mass filtering). The lower bar is the output of a MS-GF+ search on the non-spliced targets and decoys created using the proposed strategy of section 2.2 (with the suggested mass filtering of subsection 2.2.3).

The output of the MaxQuant search on the complete search space containing the spliced and non-spliced peptides is further examined using first an all-all approach and subsequently an all-sub approach (suggested in section 2.4.5). When controlling the FDR at 1% in the all-all approach, six peptides can be reported, of which five are non-spliced peptides, and one is spliced. Putting this in a larger perspective, from the 6154 peptide spectra that have been matched in total, 125 are non-spliced peptides, and thus 6029 are spliced. In the all-sub approach, where the FDR is calculated on the spliced peptides only, the same singular spliced peptide can be reported at a 1% FDR, which is logical as it has a higher score than all the decoys. This spliced peptide is modified with a deamination on N on the third position, which makes it suspicious, especially because no non-spliced peptide with this modification can be found back, when looking for non-spliced peptides who contain that piece of the spliced peptide in which the modification is present. Performing an all-sub approach on the non-spliced peptides, and thus calculating the FDR based on the non-spliced peptides only, leads to 73 non-spliced peptides that can be reported. Although this is way less than what can be reported when only considering the non-spliced peptides as search space, the amount is still higher than the outcome of the all-all approach, showing the power of the all-sub approach.

3.3 Improvement of the Code

Liepe et al. [17] mention that their provided code takes weeks to months to run, depending on the amount of clusters and cores available. By rewriting their code, the computational time speed up tremendously: using 18 cores with 22GB of RAM memory the running time is shortened to three days. Several bottlenecks are found and solved. The first bottleneck is the computation of the index matrix, Liepe et al. [17] recreate this matrix for each protein separately. This is unnecessary, and solved by creating one big index matrix, based on the largest protein, where for every smaller protein the corresponding indices are selected. Additionally, the functions to translate the indices to real peptide fragments and to calculate the MW

are modified to run more efficiently. This makes the algorithm already significantly faster, but the RAM memory fills up quickly. At most it takes two days to build the databases. The server has 18 cores but they are not all used because the 22GB RAM memory is the limiting factor. Therefore, the non-spliced and spliced peptides are generated separately, using fifteen and eight cores, respectively. The original code runs on a cluster, however, for this project only one node is disposable, thus the code needs to be adapted for this as well. The use of for-loops in R by Liepe et al. [17] is very inefficient. To be able to parallelize the code, these for-loops are replaced by the more efficient lapply and mclapply functions.

4 Discussion

In this study, the outcome of Liepe et al. [17] is questioned. Suspicion rises that the authors do not check the minimal assumptions that need to hold in order to obtain coinciding distributions of the decoys and bad target hits, which is required to perform the competitive TDA (discussed in 2.4.2). This can lead to inaccurate numbers that are reported. Because the publicly available pipeline of Liepe et al. [17] is incomplete, their exact analysis cannot be repeated. Therefore, starting from the data of the human fibroblast HLA-I immunopeptidome, a search space containing spliced and non-spliced peptides is contructed using the described strategy in section 2.2, which is then fed to the MaxQuant engine to obtain PSMs. By observing the obtained scoring distributions and evaluating the amount of spliced peptides that can be reported, we try to examine whether the research of Liepe et al. [17] is trustworthy.

Striking differences are observed when comparing our results (section 3) to those of Liepe et al. [17, 18]. The amount of non-spliced peptides that are found back at a 1% FDR by performing a default search with MS-GF+ is higher than obtained by Liepe et al. [17], i.e. 3765 versus 3039 (Figure 8, section 3.2). This is expected, because it is observed that MS-GF+ is able to distinguish the good and bad hits better, resulting in more peptides that can be reported (Figure 2, section 3.1). The use of another search engine (MS-GF+ versus Mascot) can also contribute to the amount of peptides that can be reported. Repeating the same search (with comparable parameters) on MaxQuant, a remarkably lower amount of non-spliced peptides can be reported, i.e. 449, which is not expected. A reason for this can be that MaxQuant is unable to handle this particular kind or search, as it is made to perform protein searches rather than peptide searches and therefore MaxQuant cannot make a valid distinction between the distribution of the good and bad hits. In addition, Liepe et al. [17] apply a too stringent FDR, which will be further explained in this discussion. When considering a search space containing the non-spliced targets and decoys made by the suggested strategy (section 2.2), a completely different result (MS-GF+ search) is obtained, i.e. 232. The mass filtering included in this procedure might be too conservative and can be the cause. This shows that the obtained results depend tremendously on the used search engine and the chosen parameters, and thus substantiates the uncertainty on the credibility of the claims of Liepe et al. [17].
Applying an all-all approach on the output of a MaxQuant search on the merged search space, results in one spliced peptide and five non-spliced peptides that can be reported. This is in conflict with the amount published by Liepe et al. [17], i.e. 1154 and 3039 respectively, and might be due to the imbalance in size between the target and decoy data sets, which is not checked and therefore for which is not corrected by the authors. They report 72.48% ($\frac{3039}{3039+1154}$) of the peptides as non-spliced, which keeps up the amount of TPs, and state that the scores of non-spliced peptides are generally lower than the scores of the spliced ones. The proportion of FPs among the non-spliced peptides can be considerably lower than the fraction among the spliced peptides. Only the overall FDR is controlled at 1%, the FDR in the subset can be notably higher, which weakens the conclusions of Liepe et al. [17]. Additionally, their decoy composition can be too random, leading to a shift to the left of the estimated distribution of the bad target hits. As a consequence, the authors are too optimistic in the peptides they report, containing more FPs than estimated. Furthermore, getting a high score for a good hit becomes more difficult in a larger search space (Table 1, section 3.1), because the score metric is based on the difference in scores between the best hit for a certain spectrum and its second best hit. The more spectra are joining, the smaller the score difference will be between the first and the second best hit, as they have a bigger chance on being similar. This will lead to a lower amount of peptides that can be reported. As the spliced peptide database is 167.68 times bigger than the non-spliced one (ratio of the total amount of spliced peptides to the amount of non-spliced peptides: $\frac{52832999+52501066}{315711+312480}$ displayed in Table 1 (section 3.1)), it is expected to get more matches against spliced peptides relative to the non-spliced peptides as the spliced peptides are overrepresented in the search space. Given that 6029 peptides match against spliced peptides (section 3.2), which is 3.48 ($\frac{167.68*125}{6029}$) times less than expected by random chance, matches against non-spliced peptides occur more often than against spliced peptides. This shows that blowing up the search space is not a good idea and should be handled more carefully, as it will influence the results.
The all-sub approach (with non-spliced peptides as subset) reports 73 non-spliced peptides, which is more than the five that can be reported in the all-all approach, showing the strength of the all-sub methodology. However, this result is remarkably lower than the amount reported in the sub-sub approach (i.e. 3765), where only the non-spliced peptides are taken into account. Again, a possible reason for this can be the blown up search space, because more spectra will get a better match by chance. Since theoretical spectra are considered, a lot of similar spectra will be present among the spliced peptide database, leading to decoys that will be too similar to the targets, which will induce many hits that are in fact FPs (i.e. bad target hits) and decrease the obtained matching scores. This will cause less matches that can be reported on a certain FDR cutoff, as observed in this comparison. This finding favors the use of a sub-sub approach when only non-spliced peptides are of interest. Since the spliced peptide search space is very big, the identification of spliced peptides becomes difficult as they are expected to be not redundant.

As described in section 2.4.1, the FDR is used as a measure of the uncertainty on the results and is calculated by equation (5). However, Liepe et al. [17] make use of the following formula, which is wrongly introduced by Elias & Gygi [10], taking into account all peptide targets and decoys above a certain threshold:

$$FDR = \frac{2 \cdot decoys}{decoys + targets} \tag{14}$$

Here, the decoys are taken into account twice, once to represent the bad hits due to decoy matches, and once to represent the bad hits in the target space. Apart from being a simulation of the bad target hits, the decoys do not play a role in the calculation of the FDR. Taking them into account lowers the amount of peptides that can be reported which is a conservative method of calculating the FDR. This is observed by comparing our results of the sub-sub approach considering the non-spliced peptides using MS-GF+ and the outcome of Liepe et al. [17] (Figure 8, section 3.2).

The distribution of the decoys, as a representation of the bad target hits, needs to be simulated correctly in order to be able to calculate the expected amount of FPs among the reported spliced peptides. Moreover, π_0 (the probability of a hit being a bad hit), needs to be estimated as well. The competitive TDA is used to estimate the required distributions,

which should coincide. At least the three constraints described in section 2.4.2 need to be satisfied, in order to get overlapping distributions. It is presumable that the first constraint, i.e. representative decoys, doesn't hold in the study of Liepe et al. [17]. Because the AA distribution and peptide order within humans are not taken into account. If completely random decoys are used, higher chances exist that matches against a decoy will have lower scores than bad matches against a target. This results in a too optimistic approximation of bad hits, and thus in reporting too many FPs.
Another aspect that also affects the results and the obtained scores is the database used for the human proteins. Liepe et al. [17] utilize the Swissprot database, which only contains the proteins with proven existence. There are proteins that are likely to exist, but that are not present in this search space. These can be found in Uniprot [4]. As they are not included in the original search space, they can be accidentally considered as a spliced peptide, which influences the results. Moreover, the proteasome also cuts bacterial, viral and tumor proteins besides human, which spectra aren't included either, and this can bias the results as well. However, solving this is quite impossible, as the amount of bacteria and viruses potentially present is enormous. In addition, many of their genomes are unknown, not to mention their proteins. Taking all these sequences into account would blow up the search space even more.
However, if only the second constraint, i.e. equal amount of targets as of decoys, wouldn't hold, π_0 and the FDR, can be estimated using the alternative formulas (section 2.4.4, equation 12 and 13). Otherwise, if the amount of decoys is less than the amount of targets, the peptide to decoy matches are an underestimation of the amount of bad target hits.
In the article of Liepe et al. [17] the size of the decoy database is unknown, which hinders a correct FDR control. The spliced decoy database is created by randomizing the spliced peptides, but the exact pipeline has not been described. Moreover, the authors paste the spliced peptides into 'proteins' with lengths that mimic human proteins as can be seen in Figure 1 (1). These proteins are then fed to the search engine, which cut them unspecifically, i.e. after every amino acid. By doing so, peptides are created that are not spliced, nor non-spliced, nor randomized decoy, but consist of three different protein pieces. Namely, a piece of the first part of a spliced peptide ('Spl peptide 1 part 1' on Figure 9), the whole second part of that same spliced peptide ('Spl peptide 1 part 2' on Figure 9) and a piece of the first part of another spliced peptide ('Spl peptide 2 part 1' on Figure 9), the total length of these three pieces together will be 9-12 AA. These undefined peptides are only assigned as decoys when they match against another spectra, leading to the inability to estimate the size of the decoy database, because it is unknown how many of such undefined peptides are created. As a consequence, π_0 can never be estimated correctly, creating complications in controlling the FDR and thus assessing the uncertainty on the results. The assumption of having an equally large target database as decoy database cannot be checked either.

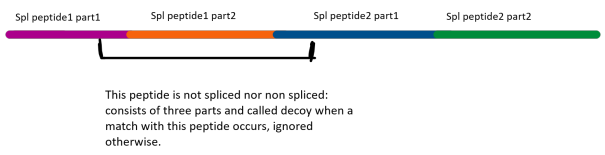

Fig. 9. Visualization of a possible peptide created by the search engine when putting all spliced peptides together, as carried out by Liepe et al. [17]. This peptide is not a non-spliced peptide, nor a spliced peptide nor a predefined decoy.

The third constraint that a bad hit is equally likely to match a target than to match a decoy is not checked as this is nowhere in the article mentioned, nor is the data to perform this, provided. This results in coinciding distributions of the bad hit targets and the decoys. This assumption is easily verified using PP-plots, if the PP-plot follows a straight line on the π_0 estimate, the distribution is valid, and the TDA can be used. The score distribution (Figure 3) and PP-plot (Figure 6) shown in section 3 confirm that the proposed strategy (section 2.2) for building the decoy database produces reliable decoys that are a good simulation of the bad target hits and that an equal amount of targets and decoys is obtained (Table 1, section 3.1).

Besides the doubt that all required assumptions hold, some other remarks according Liepe et al.'s work can be made. Liepe et al. [17] try to confirm the obtained spliced peptides by matching the obtained spectra of synthetic peptides against the theoretical spectra of interest, but these theoretical spectra are not very accurate (all peaks have the same height). The correlation between the synthetic and theoretical spectra is measured, which is expected to be close to one, as these should be spectra from the same peptide [13, 27]. However, the obtained correlations are not always very high, and the scale even goes up to negative correlations, indicating that this control provides a weak conformation and is rather invalid [27]. The scaling choice in the way the correlations are visualized is misleading (correlations going from -1 to 1, Figure 5S in Supplementary Materials [17]). As a second control to verify the decoys, which are constructed by randomizing the target sequences, Liepe et al. [17] flip the second half of the decoy sequences. This partial flip is expected to be a bad idea as the first part of the sequence stays the same. When the sequence is cut after a certain AA, some parts will attract more ions depending on the present AA, and the remaining piece will be considered as less important by the mass spectrometer resulting in less spectra for that particular part, which is here, the flipped piece. This will lead to similar spectra and thus similar scores as the original decoy sequences. Hence, this is an invalid control. Additionally, the difference in peaks between the target and the half reversed decoy can become so small that the score for the target will drop (as it is calculated based on the matching difference between the best and the second best PSM) so that the target won't get reported anymore. As the theoretical spectra aren't completely the same as the (experimental) peptide spectra, a good spectrum might accidentally match against a decoy. This will result in decoys scoring too high, and not being representative anymore for the bad target hits, leading to a wrong estimation of these bad hits. When the bad hits are modelled wrong, the good hits will be modelled wrong as well. If the decoys are too random or if the amount of decoys is too small relative to the number of targets, the distribution of the decoys is a too optimistic estimation of the distribution of the bad hits. These bad hits will actually have higher scores than estimated, leading to a FDR of the PSMs above a certain threshold that is higher than the stated 1%. To quantify the elevated FDR, the analysis should be redone, which has not been possible in the provided time of this project.
In the more recent article of Liepe et al. [18], the pipeline is expanded by some extra filtering steps to reduce the amount of FPs (shortly described in section 2.1). The fact that these filters are used indicates that the researchers are not sure about their own FDR control. Furthermore, the percentage of spliced peptides found back in all peptides often drops tremendously (almost never more than one copy). Having only one replicate of a peptide is odd in the context of the research, as the peptides in the immunopeptidome need to be recognized [8] by the immune system, repeats are plausible. A lack of repeats will hinder the recognition. Compared to the the non-spliced peptides, the difference is very clear (Figure S3 in Supplemental Figures [18]).

Hence the publicly available scripts are incomplete to reproduce the entire pipeline of Liepe et al. [17] exactly, comparing the performance and required running times of the code is difficult and quite impossible, but there can be concluded that some more effective manners are used in the

improved code.

5 Conclusion

The obtained results by performing the proposed strategy in this paper do not confirm the claims of Liepe et al. [18], i.e. the immunopeptidome contains 19.3 to 19.6% spliced peptides which account for 23.6% of the diversity. Since the exact pipeline and required scripts are not publicly available, the analysis of Liepe et al. [17] is irreproducible, making it impossible to explicitly point out the flaws in the analysis. Based on the theoretical principles in statistics and considering the assumptions that need to hold in order to perform an accurate competitive TDA, i.e. decoys need to be a good simulation of the bad target hits and the amount of decoys has to be equal to the amount of targets without having sequences in common, we suspect that the decoys generated by the authors are irrepresentative and/or that few decoys are present relative to the targets. Moreover, the wrong, too stringent formula for the FDR is used, leading to an incorrect FDR control and inaccurate reported results. Further research is necessary on whether the claims of Liepe et al. [17] are truthful, preferably with the original code of the authors and using the same technical supplies (Mascot, same server). If this is impossible, the executed analysis using the proposed strategy should be repeated on another search engine such as X!Tandem as an additional verification of the obtained results. Its output can be analyzed in PeptideShaker in order to gain more insight in the ions and to track down the precise faults and violations. Ideally, the searches should be performed using Mascot, if a license can be attained. It would be interesting to examine to which spectra the non-spliced peptides in the sub-sub approach switch compared to the all-sub approach. Additionally, new data should be generated and investigated to gain more insight in the context, as well as qualitative controls should be performed. We also strongly emphasize the need of a golden standard to handle such peptide data in order to create more accurate and valid outcomes that can serve as a reliable source for its further possible application in medicine.

Acknowledgements

We thank Lieven Clement, Lennart Martens and Adriaan Sticker for the expertise on the theoretical aspects required for this project. We thank Niels Hulstaert, Jarne Pauwels and An Staes for the technical support, Caroline De Tender for guiding us in writing a qualitative paper and overall, StatOmics group of the University of Ghent University and the CompOmics group of VIB-UGent for the accommodation and supplies.
All scripts are publicly available on GitHub: https://github.ugent.be/samdieri/Design_project_2019_Statomics

Funding

This work has been supported by Ghent University.

References

[1]M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, and M. Mann. Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Molecular & Cellular Proteomics*, 14(3):658–673, 2015.

[2]Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[3]C. R. Berkers, A. De Jong, K. G. Schuurman, C. Linnemann, H. D. Meiring, L. Janssen, J. J. Neefjes, T. N. Schumacher, B. Rodenko, and H. Ovaa. Definition of proteasomal peptide splicing rules for high-efficiency spliced peptide presentation by mhc class i molecules. *The Journal of Immunology*, 195(9):4085–4095, 2015.

[4]E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch. Uniprotkb/swiss-prot. In *Plant bioinformatics*, pages 89–112. Springer, 2007.

[5]L. Candela. Boundless anatomy and physiology: Antigens. https://courses.lumenlearning.com/boundless-ap/chapter/antigens/, 2019. [Online; accessed 27-April-2019].

[6]E. Caron, R. Aebersold, A. Banaei-Esfahani, C. Chong, and M. Bassani-Sternberg. A case for a human immuno-peptidome project consortium. *Immunity*, 47(2):203–208, 2017.

[7]L. Clement. Statistical genomics. https://github.com/statOmics/statisticalGenomicsCourse, 2018.

[8]M. Colonna and J. Samaridis. Cloning of immunoglobulin-superfamily members associated with hla-c and hla-b recognition by human natural killer cells. *Science*, 268(5209):405–408, 1995.

[9]A. Dalet, N. Vigneron, V. Stroobant, K.-i. Hanada, and B. J. Van den Eynde. Splicing of distant peptide fragments occurs in the proteasome by transpeptidation and produces the spliced antigenic peptide derived from fibroblast growth factor-5. *The Journal of Immunology*, 184(6):3016–3024, 2010.

[10]J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207, 2007.

[11]J. E. Elias and S. P. Gygi. Target-decoy search strategy for mass spectrometry-based proteomics. In *Proteome bioinformatics*, pages 55–71. Springer, 2010.

[12]S. Istrail, L. Florea, B. V. Halldórsson, O. Kohlbacher, R. S. Schwartz, V. B. Yap, J. W. Yewdell, and S. L. Hoffman. Comparative immunopeptidomics of humans and their pathogens. *Proceedings of the National Academy of Sciences*, 101(36):13268–13272, 2004.

[13]L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923, 2007.

[14]L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, 7(01):29–34, 2007.

[15]S. Kim and P. A. Pevzner. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications*, 5:5277, 2014.

[16]N. Le Floc'h, D. Melchior, and C. Obled. Modifications of protein and amino acid metabolism during inflammation and immune system activation. *Livestock Production Science*, 87(1):37–45, 2004.

[17]J. Liepe, F. Marino, J. Sidney, A. Jeko, D. E. Bunting, A. Sette, P. M. Kloetzel, M. P. Stumpf, A. J. Heck, and M. Mishto. A large fraction of hla class i ligands are proteasome-generated spliced peptides. *Science*, 354(6310):354–358, 2016.

[18]J. Liepe, J. Sidney, F. K. Lorenz, A. Sette, and M. Mishto. Mapping the mhc class i–spliced immunopeptidome of cancer cells. *Cancer immunology research*, 7(1):62–76, 2019.

[19]A. Michaux, P. Larrieu, V. Stroobant, J.-F. Fonteneau, F. Jotereau, B. J. Van den Eynde, A. Moreau-Aubry, and N. Vigneron. A spliced antigenic peptide comprising a single spliced amino acid is produced in the proteasome by reverse splicing of a longer peptide fragment followed by trimming. *The Journal of Immunology*, 192(4):1962–1971, 2014.

[20]M. Mishto and J. Liepe. Post-translational peptide splicing and t cell responses. *Trends in immunology*, 38(12):904–915, 2017.

[21]R. Mylonas, I. Beer, C. Iseli, C. Chong, H.-S. Pak, D. Gfeller, G. Coukos, I. Xenarios, M. Müller, and M. Bassani-Sternberg. Estimating the contribution of proteasomal spliced peptides to the hla-i ligandome. *Molecular & Cellular Proteomics*, 17(12):2347–2357, 2018.

[22]A. W. Purcell, N. P. Croft, and D. C. Tscharke. Immunology by numbers: quantitation of antigen presentation completes the quantitative milieu of systems immunology! *Current opinion in immunology*, 40:88–95, 2016.

[23]A. Sticker, L. Martens, and L. Clement. Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nature methods*, 14(7):643, 2017.

[24]S. Tyanova, T. Temu, and J. Cox. The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, 11(12):2301, 2016.

[25]N. Vigneron, V. Stroobant, J. Chapiro, A. Ooms, G. Degiovanni, S. Morel, P. van der Bruggen, T. Boon, and B. J. Van den Eynde. An antigenic peptide produced by peptide splicing in the proteasome. *Science*, 304(5670):587–590, 2004.

[26]N. Vigneron, V. Stroobant, V. Ferrari, J. A. Habib, and B. J. Van den Eynde. Production of spliced peptides by the proteasome. *Molecular immunology*, 2018.

[27]S. Yilmaz-Rumpf, E. Vandermarliere, and L. Martens. Methods to calculate spectrum similarity. In S. Keerthikumar and S. Mathivanan, editors, *Proteome bioinformatics*, volume 1549 of *Methods in Molecular Biology*, pages 75–100. Springer, 2017.