UGC Care Group I Journal
Journal of Management & Entrepreneurship Vol-10Issue-02 2021
ISSN 2229-5348

Identification of alternative splice variants in Aspergillus flavus through comparison of multiple tandem MS search algorithms

Mrs.K.Swapna Kumari , Abdul Gaffor

Assistant Professor, Professor
1,2 Department of H&S
1,2Global Institute of Engineering and Technology,Moinabad,RR District, Telangana State

Abstract

Background:

Database searching is the most frequently used approach for automated peptide assignment and protein inference of tandem mass spectra. The results, however, depend on the sequences in target databases and on search algorithms. Recently by using an alternative splicing database, we identified more proteins than with the annotated proteins in Aspergillus flavus. In this study, we aimed at finding a greater number of eligible splice variants based on newly available transcript sequences and the latest genome annotation. The improved database was then used to compare four search algorithms: Mascot, OMSSA, X! Tandem, and InsPecT.

Results: The updated alternative splicing database predicted 15833 putative protein variants, 61% more than the previous results. There was transcript evidence for 50% of the updated genes compared to the previous 35% coverage. Database searches were conducted using the same set of spectral data, search parameters, and protein database but with different algorithms. The false discovery rates of the peptide-spectrum matches were estimated < 2%. The numbers of the total identified proteins varied from 765 to 867 between algorithms. Whereas 42% (1651/3891) of peptide assignments were unanimous, the comparison showed that 51% (568/1114) of the RefSeq proteins and 15% (11/72) of the putative splice variants were inferred by all algorithms. 12 plausible isoforms were discovered by focusing on the consensus peptides which were detected by at least three different algorithms. The analysis found different conserved domains in two putative isoforms of UDP-galactose 4-epimerase.

Conclusions: We were able to detect dozens of new peptides using the improved alternative splicing database with the recently updated annotation of the A. flavus genome. Unlike the identifications of the peptides and the RefSeq proteins, large variations existed between the putative splice variants identified by different algorithms. 12 candidates of putative isoforms were reported based on the consensus peptide-spectrum matches. This suggests that applications of multiple search engines effectively reduced the possible false positive results and validated the protein identifications from tandem mass spectra using an alternative splicing database.

Background

Tandem mass spectrometry (MS/MS) has been one of the most effective high-throughput approaches for protein identification and quantification. In a typical “bottom-up” approach, also known as the shotgun proteomics strategy, the enzyme-digested protein mixture is analyzed using single- or multi-dimensional chromatography coupled with tandem mass spectrometry [1,2]. A variety of computational approaches have been developed to assign peptide sequences to the acquired MS/MS data. Database searching algorithms are the most frequently used methods for large-scale proteomics studies [3]. The most popular commercial MS/MS search engines are SEQUEST [4] (Thermo Fisher Scientific Inc.) and Mascot [5] (Matrix Science Ltd.). Open source tools are also available, such as OMSSA [6], X! Tandem [7], and Andromeda [8]. Although each implementation is different, the general approach of MS/MS search algorithms is similar [9]. Given a protein sequence database, the search algorithm first generates all in silico-digested peptides upon the specified parameters, such as digestive enzymes, missed cleavages, and

modifications. For each MS/MS spectrum, the search engine only evaluates the candidate peptide sequences within a user-defined precursor mass tolerance window. A scoring function is used to calculate a score which represents how well the theoretical spectrum of each candidate peptide matches the observed spectrum. The top scoring peptide hit is reported and then the peptide sequence is assigned to the experimental MS/MS spectrum. Protein identifications are inferred by grouping the peptide-spectrum matches [10].

Another approach for identifying peptides from fragment ion spectra combines partial de novo sequencing and database searching. Short peptide sequence tags are inferred from MS/MS spectra using de novo algorithms. The list of candidate peptides in the database search can be reduced to only those containing the tag [11]. The algorithms will then try to extend the sequence tag by finding masses of the flanking residues in the database peptide which match masses of the prefix and suffix regions of the tag [12]. Although the hybrid approach is still reliant on protein sequence databases, it is an alternative strategy while analyzing peptides with novel modifications or sequence variations [13].

Alternative pre-mRNA splicing (AS) enables eukaryotes to generate distinct mRNAs and therefore multiple protein variants from a single gene. The common approach to developing an alternative splicing database is based on automated large-scale mapping of transcripts and genomic sequences. The massively parallel picolitrescale sequencing system developed by the 454 Life Sciences Corporation was capable of sequencing 25 million bases in a four-hour run [14]. The 454 sequence reads are short, averaging 80-120 bases per read. The massively parallel sequencing-by-synthesis technology has been used to generate EST data of a human prostate cancer cell line, and 25 novel alternative exon splicing events were identified [15].

Recently, we expanded the target database to include putative alternatively spliced isoforms with the aim that the MS/MS spectra can be better interpreted [16]. The results showed that our approach was able to identify more proteins from the experimental spectra and to provide evidence for improving the genome annotation. Subsequently, the Aspergillus flavus NRRL3357 whole genome shotgun project had a major update in 2009. Among 41 peptides discovered in our previous study, 6 of them were included in the second version of genome annotation. Meanwhile, 454 sequencing data of A. flavus became available locally. The first goal of this study was to rebuild the alternative splicing database using the latest genome annotation and newly acquired 454 sequencing data as transcript evidence. The second part of the study aimed at comparing four MS/MS search algorithms for isoform identifications using the resulting alternative splicing database. We tested three probability-based algorithms, Mascot [5], OMSSA [6], and X! Tandem [7], and one sequence tag-based algorithm, InsPecT [12]. The design of the study is illustrated in Figure 1.

Results

Rebuilding A. flavus alternative splicing database

Genome annotation is the result of continuous efforts. An updated version of A. flavus genome annotation was released in 2009. Compared to the prior genome project, the second version dropped 360 previously documented genes and added 1000 novel ones (Figure 2A). A newly acquired collection of 454 sequence reads and ESTs provided the transcription information of half of the genes for predicting splice variants (Figure 2B). An updated alternative splicing database was then built using the second version of the genome and all available transcripts. The RefSeq database (release 40) contained 13487 A. flavus genes and corresponding proteins, with no splice isoform. The updated alternative splicing database predicted another 15833 putative protein variants (Figure 2C). It was estimated that 15.4% (2077/13487) of the total genes encoded more than one protein, 7.62 (15833/2077) putative isoforms per gene on average. The predicted variant sequences were appended to the collection of the RefSeq proteins to form a combined database for the following database searches.

Comparison of MS/MS search algorithms on identifying putative isoforms

In order to compare the performance of identifying putative splice variants, the same set of MS/MS spectra were searched against the resulting combined database by Mascot, OMSSA, X! Tandem, and InsPecT. Although each algorithm already reported internal statistical measures like p-value or E-value, the cut-off thresholds were selected to ensure the search results had an estimated false discovery rate (FDR) < 2% for peptide identification (see Additional file 1). While several isoforms were encoded from the same gene, sometimes the different protein products could not be distinguished by the identified peptides.

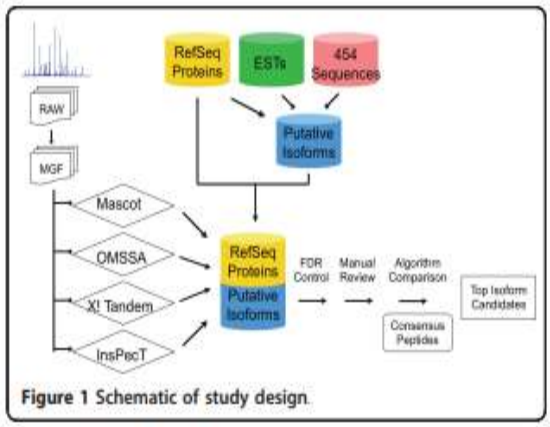

Figure 1 Schematic of study design.

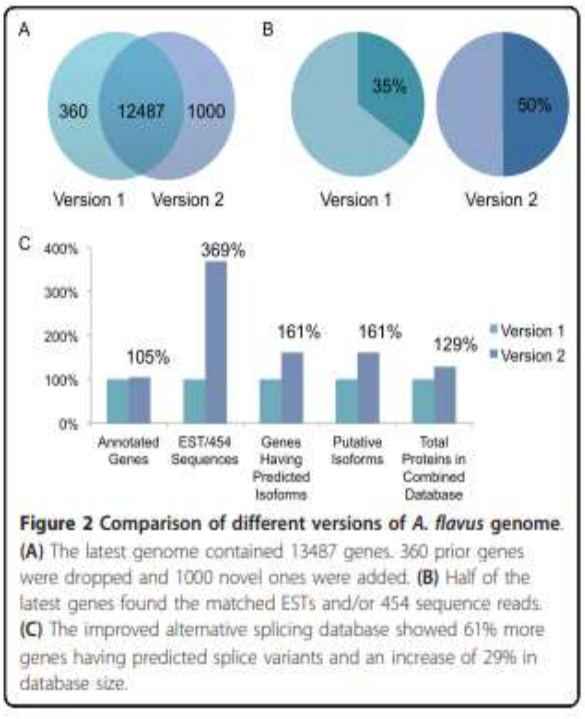

Figure 2 Comparison of different versions of *A. flavus* genome. **(A)** The latest genome contained 13487 genes. 360 prior genes were dropped and 1000 novel ones were added. **(B)** Half of the latest genes found the matched ESTs and/or 454 sequence reads. **(C)** The improved alternative splicing database showed 61% more genes having predicted splice variants and an increase of 29% in database size.

In such a scenario, it was observed that Mascot would pick the protein with the longest sequence from all possible candidates. InsPecT would also report one protein from the list of candidate sequences, but not necessarily the longest one. In contrast, OMSSA and X! Tandem would report all matched proteins and let users interpret the findings. In order to present the results concisely, we accepted the longest protein sequence to represent the group of all possible matches. If a group of peptides could be mapped to either the RefSeq protein or the splice variant of the same gene, we conservatively assigned the identification to the RefSeq protein since no clear conclusion was possible. The number of identified peptides, RefSeq proteins, and splice variants by algorithms are listed in Table 1.

To study the consistency between different algorithms on search results, the identified hits were categorized by the algorithms having the same finding (Table 2). The overlaps were illustrated in four-way Venn diagrams as well (Figure 3). For the peptide-spectrum matches, 42% (1651/3891) of peptide assignments were concurred by all four algorithms. Since we introduced predicted isoform sequences into the database, the protein identification was divided into two subgroups: RefSeq proteins and putative splice variants. 51% (568/1114) of the identified RefSeq proteins were consistent across all algorithms. In contrast, only 15% (11/72) of the putative splice variants were identified unanimously.

To investigate whether different algorithms assigned the same spectrum to different peptide sequences, the peptide-spectrum matches were examined within and between algorithms (Table 3). It was observed for all algorithms that 1% or fewer spectra were assigned to different peptides by the same tool. The inconsistency expanded but never exceeded 2% while comparing the assignment of the same spectrum between different algorithms. It also appeared that InsPecT assigned more spectra differently in comparison with other three probability-based algorithms. The multiple peptides assigned from the same spectra between algorithms might account for a part of the identification variations.

It was not surprising to see that the number of peptidespectrum matches and protein hits dropped while reducing the false discovery rate. However, most of the removed hits belonged to the identifications reported by only one algorithm (see

Additional file 2). The consensus hits of multiple algorithms seemed more likely to be the correct identification. In the comparison of the overlaps between search results, the identified splice variants between different algorithms showed greater variations than the RefSeq proteins. It is noted that the prediction of all possible splice variants from ESTs tends to be overestimated. To reduce the false positive results, we compiled a list of top splice isoform candidates by taking advantage of the consensus peptides. By focusing on those variant-specific peptides identified by at least three different algorithms, 12 putative isoforms were reported (Table 4). 11 splice variants were inferred by all four algorithms. The scores, p-values, and E-values of the assignments looked satisfying. None of these specific peptide sequences appeared in any RefSeq proteins. In addition, no two consensus peptides came from the same spectra. As an example, one putative isoform discovered through the strategy was further analysed below.

Table 1 Number of identified peptides and proteins by algorithms with a FDR < 2%

Algorithm	Threshold	Number of Identified Peptides	MS/MS FDR (%)	Number of Identified RefSeq Proteins	Number of Identified Splice Variants
Mascot	E-value < 0.001	2731	1.88	747	21
OMSSA	E-value < 0.09	2457	1.89	745	20
X! Tandem	E-value < 0.04	2692	1.76	857	36
InsPecT	p-value < 0.002	2569	1.91	738	33

Table 2 Overlap of identified peptides and proteins between algorithms with a FDR < 2%

Algorithm	Peptides		RefSeq Proteins		Putative Isoforms	
	Count	%	Count	%	Count	%
Mascot only	275	7.1	58	5.2	6	8.3
OMSSA only	152	3.9	72	6.5	8	11.1
X! Tandem only	471	12.1	156	14.0	23	31.9
InsPecT only	334	8.6	98	8.8	20	27.8
Mascot, OMSSA	115	3.0	24	2.2	0	0.0
Mascot, X! Tandem	100	2.6	13	1.2	2	2.8
Mascot, InsPecT	80	2.1	10	0.9	1	1.4
OMSSA, X! Tandem	83	0.8	17	1.5	0	0.0
OMSSA, InsPecT	44	1.1	5	0.4	0	0.0
X! Tandem, InsPecT	108	2.8	12	1.1	0	0.0
Mascot, OMSSA, X! Tandem	185	4.8	36	3.2	0	0.0
Mascot, OMSSA, InsPecT	209	5.4	16	1.4	1	1.4
Mascot, X! Tandem, InsPecT	116	3.0	22	2.0	0	0.0
OMSSA, X! Tandem, InsPecT	21	0.5	7	0.6	0	0.0
Mascot, OMSSA, X! Tandem, InsPecT	1651	42.4	568	51.0	11	15.3
Grand Total	3891	100.0	1114	100.0	72	100.0

Conserved domain analysis of putative isoforms of UDPgalactose 4-epimerase

UDP-glucose 4-epimerase (UGE) [KEGG: EC 5.1.3.2] plays a pivotal role in normal galactose metabolism, converting UDP-galactose back to UDP-glucose in the final step of the Leloir pathway [17]. NAD+ is required

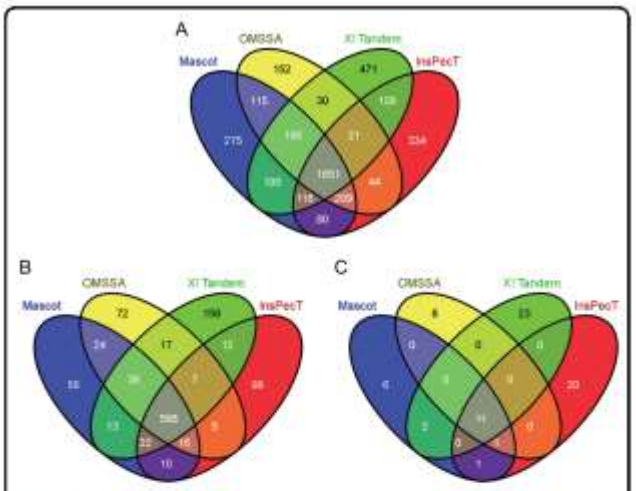

Figure 3 Overlap of peptide and protein identifications using different search algorithms. The 4-way Venn diagrams generated by the VENNY program [33] illustrate the intersections of **(A)** all peptides, **(B)** RefSeq proteins, and **(C)** putative splice variants identified by Mascot, OMSSA, X! Tandem, and InsPecT. In addition to 42% (1651/3891) of identified peptides overlapping, all four algorithms agreed on 51% (568/1114) of RefSeq protein identifications but only 15% (11/72) of the putative splice variants. All search results had an estimated FDR < 2% for peptide identification.

to be a cofactor in the catalytic mechanism. Five UGE isoforms encoded in the Arabidopsis thaliana genome differed in enzymatic properties, transcript regulation, and subcellular localization [18]. The MS/MS spectrum which was used to assign the consensus peptide FAVETAITDVINAQR in the putative UGE isoform was examined (Figure 4). The abundant matched b- and yions, accurate precursor ion mass, and expected mass difference from the SILAC pair observed in the spectrum correlated well with the low expectation value or p-value reported by algorithms.

According to the annotation of RefSeq release 40, A. flavus UDP-glucose 4-epimerase [Entrez Gene: 7919639] contained four coding exons (Figure 5A). The corresponding splice variant generated from our prediction had three exons instead: the first two were constitutive and the third was alternative (Figure 5B). Since different sets of peptide-spectrum matches were used to conclude the protein identification between search algorithms,

the peptides shown in Figure 5 are based on Mascot's result. The alternative exon in the protein variant was supported by the distinctive peptide FAVETAITDVINAQR which was located in an intron of the corresponding RefSeq protein. The encoding variant sequence ended approximately in the middle of the third coding exon of the RefSeq counterpart. A group of 9 peptides which were mapped to the remaining coding sequence supported the identification of the RefSeq protein.

While multiple protein products are encoded from the same gene, different isoforms are usually destined for performing various biological functions. Thus, we were interested in learning whether two identified UGE

Table 3 Number of MS/MS spectra assigned to different pe

Algorithm	Number of assigned spectra	Assigned to different peptides by Mascot
Mascot	9410	87 (0.92%)
OMSSA	8531	128 (1.50%)
X! Tandem	8820	76 (0.86%)
InsPecT	8832	132 (1.49%)

isoforms had different functional motifs among their sequences. The Conserved Domain Database (CDD), part of NCBI's Entrez database system, is a protein annotation resource that consists of a collection of wellannotated multiple sequence alignment models as position-specific score matrices (PSSMs) [19]. Two motifs were found by searching the RefSeq sequence against CDD (version 2.23, containing 37407 PSSMs) (Figure 5A). One was a member of the Rossmann-fold NAD(P) (+)-binding proteins superfamily, 3-ketoacyl-(acyl-carrier-protein) reductase [CDD: PRK12825], and the other was UDP-glucose 4-epimerase [CDD: PLN02240].

Conclusions

The prediction of the alternatively spliced variants based on EST sequences by a computational pipeline inclines to be over-estimated and may contain errors. The introduction of putative isoforms into the protein database can further lower the p-value of peptide identifications because of the increasing size of the database. Consensus decision making exploits the goodness of multiple search algorithms to validate the assignment results of spectral data at a relatively low cost. The approach is particularly valuable while making inferences in isoform identifications from an alternative splicing database.

References

1. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR III: Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol 1999, 17:676-682.

2. Washburn MP, Wolters D, Yates JR III: Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 2001, 19:242-247.

3. Sadygov RG, Cociorva D, Yates JR III: Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods 2004, 1:195-202.

4. Eng JK, McCormack AL, Yates JR III: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 1994,

5:976-989. 5. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999, 20:3551-3567.

6. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: Open mass spectrometry search algorithm. J Proteome Res 2004, 3:958-964.

7. Craig R, Beavis RC: TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004, 20:1466-1467.

8. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M: Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 2011, 10:1794-1805.

9. Balgley BM, Laudeman T, Yang L, Song T, Lee CS: Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. Mol Cell Proteomics 2007, 6:1599-1608.

10. Nesvizhskii AI, Vitek O, Aebersold R: Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007, 4:787-797.

11. Hughes C, Ma B, Lajoie GA: De novo sequencing methods in proteomics. Methods Mol Biol 2010, 604:105-121.

12. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V: InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 2005, 77:4626-4639.

13. Tabb DL, Saraf A, Yates JR III: GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 2003, 75:6415-6421.

14. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437:376-380.

15. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, et al: Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 2006, 7:246.

16. Chang KY, Georgianna DR, Heber S, Payne GA, Muddiman DC: Detection of alternative splice variants at the proteome level in Aspergillus flavus. J Proteome Res 2010, 9:1209-1217.

17. Holden HM, Rayment I, Thoden JB: Structure and function of enzymes of the Leloir pathway for galactose metabolism. J Biol Chem 2003, 278:43885-43888.

18. Barber C, Rosti J, Rawat A, Findlay K, Roberts K, Seifert GJ: Distinct properties of the five UDP-D-glucose/UDP-D-galactose 4-epimerase isoforms of Arabidopsis thaliana. J Biol Chem 2006, 281:17276-17285.

19. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, et al: CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 2009, 37:D205-D210.

20. Thoden JB, Wohlers TM, Fridovich-Keil JL, Holden HM: Human UDPgalactose 4-epimerase. Accommodation of UDP-N-acetylglucosamine within the active site. J Biol Chem 2001, 276:15131-15136.

21. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ: An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 2005, 5:3475-3490.

22. Searle BC, Turner M, Nesvizhskii AI: Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. J Proteome Res 2008, 7:245-253.

23. Edwards N, Wu X, Tseng CW: An unsupervised, model-free, machinelearning combiner for peptide identifications from tandem mass spectra. Clin Proteomics 2009, 5:23-36.

24. Yu W, Taylor JA, Davis MT, Bonilla LE, Lee KA, Auger PL, Farnsworth CC, Welcher AA, Patterson SD: Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. Proteomics 2010, 10:1172-1189.

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403-410.

26. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 1998, 8:967-974.

27. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA: Splicing graphs and EST assembly problem. Bioinformatics 2002, 18(Suppl 1):S181-S188.

28. Georgianna DR, Hawkridge AM, Muddiman DC, Payne GA: Temperaturedependent regulation of proteins in Aspergillus flavus: whole organism stable isotope labeling by amino acids. J Proteome Res 2008, 7:2973-2979.

29. Proteome Commons Database [http://proteomecommons.org]. .

30. Choi H, Nesvizhskii AI: False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res 2008, 7:47-50.

31. Käll L, Storey JD, MacCoss MJ, Noble WS: Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 2008, 7:29-34.

32. Elias JE, Gygi SP: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007, 4:207-214.

33. Oliveros JC: VENNY. An interactive tool for comparing lists with Venn Diagrams [http://bioinfogp.cnb.csic.es/tools/venny/index.html]. 2007