Identification of rare alternative splicing events in MS/MS data reveals a significant fraction of alternative translation initiation sites

Integration of transcriptome data is a crucial step for the identification of rare protein variants in mass-spectrometry (MS) data with important consequences for all branches of biotechnology research. Here, we used Splooce, a database of splicing variants recently developed by us, to search MS data derived from a variety of human tumor cell lines. More than 800 new protein variants were identified whose corresponding MS spectra were specific to protein entries from Splooce. Although the types of splicing variants (exon skipping, alternative splice sites and intron retention) were found at the same frequency as in the transcriptome, we observed a large variety of modifications at the protein level induced by alternative splicing events. Surprisingly, we found that 40% of all protein modifications induced by alternative splicing led to the use of alternative translation initiation sites. Other modifications include frameshifts in the open reading frame and inclusion or deletion of peptide sequences. To make the dataset generated here available to the community in a more effective form, the Splooce portal (http://www.bioinformatics-brazil.org/splooce) was modified to report the alternative splicing events supported by MS data.

Identification of rare alternative splicing events in MS/MS data reveals a significant fraction of alternative translation initiation sites

José E. Kroll[1,2], Sandro J. de Souza[2] and Gustavo A. de Souza[3*]

1. Institute of Bioinformatics and Biotechnology, R. da Palestina 99 sala 109, CEP 59092-460 Natal, Brazil
2. Brain Institute, UFRN, Av. Nascimento de Castro 2155, CEP 59056-450, Natal, Brazil
3. Dept. of Immunology and Centre for Immune Regulation, Oslo University Hospital HF Rikshospitalet, University of Oslo, Oslo, Norway.

* Correspondence should be addressed to:

Gustavo A. de Souza,
PO Box 4950 Nydalen, 0424 Oslo, Norway
g.a.d.souza@medisin.uio.no (+47 23074223)

Running Title: Alternative splicing identification by mass spectrometry

ABBREVIATIONS

ASE – Alternative splicing events

TIS – Translational initiation site

FDR – False discovery rate

GTI-Seq – Global translational initiation sequencing

INTRODUCTION

The development of large-scale technologies, including genomics, has revolutionized life sciences. For example, the sequencing of the human genome in 2001 was a milestone in the characterization of our genetic framework (Lander et al., 2001; Venter et al., 2001). The advancement of sequencing technologies in the last few years has allowed the genome sequencing of more than a thousand human individuals (1000 Genomes Project) (Consortium 2012). Likewise, the characterization of the transcriptome was also facilitated by these new sequencing technologies. RNA-Seq techniques have allowed the identification of transcripts with low copy numbers. Thus, the complete characterization of the transcriptome of different cell types is already a reality today (Au et al., 2013; Peng et al., 2012; Xue et al., 2014). We know for example about the large variability found in the transcriptomes of eukaryotes due to alternative splicing and alternative polyadenylation. As a consequence of the emergence of these technologies, an explosion of this type of data in public databanks and data repositories is already occurring and exponential growth is expected for the next years. Improving bioinformatics capabilities is crucial for the processing, storage and interpretation of results from large-scale technologies.

While the technologies for sequencing of nucleic acids developed at an impressive speed, the same did not happen with technologies for sequencing amino acids and proteins. Recently, mass spectrometry-based proteomics achieved enough comprehensiveness and throughput to allow in-depth characterization of “complete proteomes” (Beck et al., 2011; Nagaraj et al., 2011). However, proteomic data acquisition is still restricted to few groups, even though public availability of high depth proteomic data is increasing (Desiere et al., 2006; Vizcaino et al., 2013; Vizcaino et al., 2014).

Alternative splicing is defined, basically, as a process in which identical pre-mRNA molecules are processed in different ways in terms of usage of splice sites. It is a fundamental process in all multi-cellular organisms being responsible for the creation of a large diversity of proteins from a relatively small number of genes (Cork et al., 2012). Alternative splicing events

(ASE) have been extensively characterized using transcriptome data. On the other hand, only recently proteome data have been used for global discovery of ASEs (Brosch et al., 2011; Severing et al., 2011; Tress et al., 2008). The reason lies on the following: protein identification by mass spectrometry is still routinely performed through the use of protein databases cataloged and curated by public repositories such as nrNCBI and Uniprot. Most of these databanks contain only a limited number of protein sequence isoforms, and single nucleotide polymorphisms and ASEs are normally under-represented. This is generally so because peptide identification approaches in proteomics mostly use probabilistic-based algorithms, and excessively large databases would result in spurious spectral matches and, therefore, reduced number of positive identifications (Wang et al., 2012). Thus, new approaches should be developed where ASEs can be investigated without compromising database size and protein identification rates. Several researchers have created strategies that use MS data repositories such as Peptide Atlas and in silico protein database design using nucleotide sequence repositories or merging protein sequence databases (Blakeley et al., 2010; Brosch et al., 2011). However, very few had applied RNA-Seq data to offer isoform information at the transcriptome level, which then could be validated at the protein level. For example, Sheynkman and colleagues (Sheynkman et al., 2013) developed a strategy where RNA-Seq and MS data collected from the same samples had been applied for the identification of splice junction peptides. However, applying such different expertise in any project might not be a reality for a majority of laboratories. Therefore, creating strategies that rely on heavy bioinformatics analysis of nucleotide de novo sequence and validation through MS is relevant.

Here, we investigated whether ASEs could be satisfactorily identified using size-limited FASTA database, built from repositories of expressed sequences, which was then challenged by MS data. Our group had recently developed Splooce, a database that integrates information from transcriptome analysis, including RNA-Seq, to identify splicing variants (Kroll et al., 2012).

Protein entries created from Splooce were evaluated using MS/MS analysis, and a large number of novel proteins isoforms were identified. Surprisingly we found that around 40% of all modifications at the protein level were related to the use of alternative translation initiation sites (TIS).

MATERIALS & METHODS

Protein variants identification using mass spectrometry and MaxQuant

Predicted proteins (in FASTA format) were collected from the full Splooce database and filtered for entries showing alternative splicing events supported only by ESTs and/or RNASeq expressed sequences. Those events were tagged as rare since they were not found in the set of full-insert cDNA sequences (RefSeq, mRNA), which usually have well characterized coding sequences. Any pattern of combined alternative splicing event was allowed. As default parameter, Splooce only reports events that are supported by at least two expressed sequences. For the prediction of protein sequences, Splooce uses a simple ab-initio strategy. Briefly, human entries from the Reference Sequence database (Pruitt et al., 2014) were modified by introducing alternative splicing patterns observed from the transcriptome data. Thus, full-length alternative cDNA sequences were created from expressed sequence fragments that often cover only a small fraction of coding sequences. As a final step, prior to the translation process, new open reading frames are predicted based on their length. Our final set of predicted proteins, containing 120,299 entries, can be downloaded from http://www.bioinformatics-brazil.org/~jkroll/sploocemm. Human entries from Uniprot (from December 2013) (Magrane & Consortium 2011) were added to the Splooce dataset to facilitate the visualization of identified peptides that are not unique to the Splooce set. The final dataset contained 209,927 entries.

We submitted the collection of entries from Splooce plus Uniprot to a dataset of MS/MS peptide information collected from 11 tumor cell lines that were publicly available at the Tranche

Network (currently discontinued). The whole collection of MS data was derived from the laboratory of Dr. Mathias Mann (Geiger et al., 2012). Four RAW files from this dataset were not used because they were apparently corrupted in the depository. We submitted the remaining files to a MaxQuant (version 1.4.1.2) (Cox & Mann 2008) search using the following parameters: initial search with a precursor mass tolerance of 20 ppm that were used for mass recalibration; main search precursor mass and fragment mass were searched with mass tolerance of 6 ppm. The search included variable modifications such as Met oxidation, N-terminal acetylation (protein), and Pyro-Glu (Q)(E). Carbamidomethyl cysteine was added as a fixed modification. Minimal peptide length was set to 7 amino acids and a maximum of two miscleavages were allowed. The false discovery rate (FDR) was set to 0.01 for peptide and protein identifications. In the case of identified peptides that are shared between two proteins, these are combined and reported as one protein group. Protein table output was filtered to eliminate the identifications from the reverse database, and common contaminants.

Protein variants identification using a *de novo* strategy

We also decided to test the ability to identify peptides characterizing ASEs using a *de novo* approach rather than a probabilistic one using a database. MS raw files were submitted to *de novo* sequence identification using the PEAKS software (Ma et al., 2003). Parameters were set as: i) trypsin with no proline restriction as enzyme, ii) two miscleavages allowed and iii) precursor ion and fragment ion error of 10 ppm. Furthermore, carbamydomethyl (Cys) as fixed modification, while protein N-term acetylation, Met oxidation and pyro-Glu (Q / E) were also allowed as variable modifications. Only peptide sequences with more than 80% average coverage certainty were selected for further analysis. Coverage certainty is calculated on an amino acid per amino acid basis, i.e., only in cases where the software was able to precisely detect mass of the amino acid removed from two neighboring daughter ions.

Identification of peptides supporting alternative splicing events

The output file of identified peptides obtained from MaxQuant and PEAKS were filtered for peptides observed specifically on Splooce entries. As described above, all MaxQuant peptides showing reversed and contaminant tags were removed from the data set. The resulting peptides were then compared against an unmodified set of RefSeq sequences, which Splooce uses as template for predicting new proteins. Any peptide observed for a Splooce entry, but not observed for its respective unmodified RefSeq, was classified as an ASE supporting peptide since it aligns uniquely to the alternative protein sequence. Additionally, any ASE supporting peptides matching the beginning of proteins were classified as alternative translation start sites.

A clear limitation in a “database-based” approach is a reduction in peptide/protein identification due to an increase in the search space by creating an excessively large database. Therefore we restricted our database to a size approximately twice as big as Uniprot. Protein identification using our database obtained approximately 500 proteins less than the original publication, a variation of less than 5%. Since the original publication used a version of the discontinued International Protein Index database, we also submitted the dataset to Uniprot database without our in house Splooce sequences (data not shown), since Uniprot and IPI would have closer number of entries and therefore, similar search space. The Uniprot result identified approximately 200 proteins less than the original publication. Such differences are probably due to: i) different identified unique entries in Uniprot or IPI, ii) small differences in the parameters between our MaxQuant search and the original publication, and/or iii) differences in MaxQuant performance since we used an updated version compared to the one used the original publication. Regardless, we concluded that even doubling the database size with Splooce entries, protein identification penalty was irrelevant for the approach efficiency.

RESULTS AND DISCUSSION

Identification of splicing variants in the MS/MS data

Splooce was used as a source to create a database of predicted protein isoforms in FASTA format, which was then searched against MS/MS spectra. A data set of 120,299 non-redundant protein sequences was created based on rare ASEs that were not observed for full-insert cDNA sequences (see Experimental Procedures for more details). That data set was merged to 89,602 Uniprot entries from the December 2013 release. A public collection of MS RAW files was then selected for protein identification. Only files from a publication that reported good level of instrument sensitivity and proteomic depth (Geiger et al., 2012) were used and such MS dataset was challenged against the Splooce-derived protein sequences using two peptide identification approaches, one based in probabilistic method and another one based on *de novo* sequencing (Figure 1). Both methods offer unique advantages and limitations. *De novo* sequencing provides unbiased peptide identification, not limited to its theoretical existence in a database. On the other hand, sequence information can only be obtained from good to high quality MS/MS data, and partial sequence information is generally discarded. Algorithms using a protein database overall offer a higher identification rate, since partial sequence information, together with accurate mass measurement of the precursor peptide ion, can still provide positive identification. *De novo* data also offer additional possibilities since once a given sequence information is obtained it can be aligned against sequence repositories to provide protein identification.

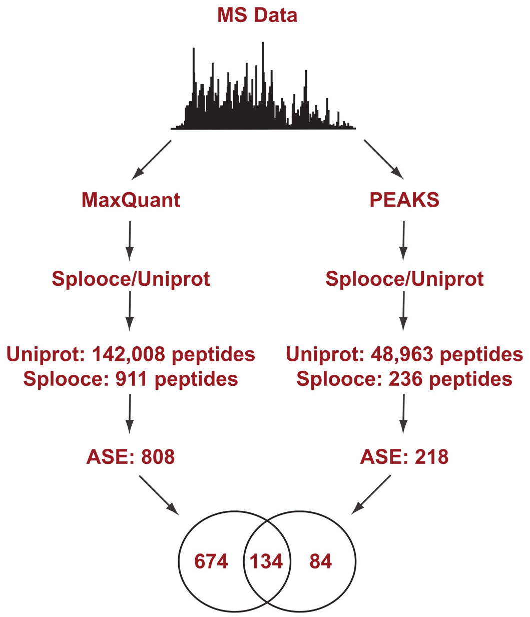


Figure 1. Experimental design flowchart. Briefly, public MS data from 11 cell lines (Geiger et al., 2012) were submitted to peptide identification using a Splooce database either by a probabilistic approach (MaxQuant) or a *de novo* approach (PEAKS). Identified peptides were sorted and those characterizing alternative splicing events not present in Uniprot were compared.

Initial analysis using the probabilistic approach (MaxQuant) allowed us to identify a total of 142,926 unique peptides representing 11,237 protein groups. Supplementary files S1 reports the MaxQuant peptide output containing the identification features for both the total peptides identified and the ones identified only in the Splooce database. As expected, the vast majority (142,008) of these peptides are already present in Uniprot. However, 911 peptides, representing 808 ASE, were only observed for Splooce entries.

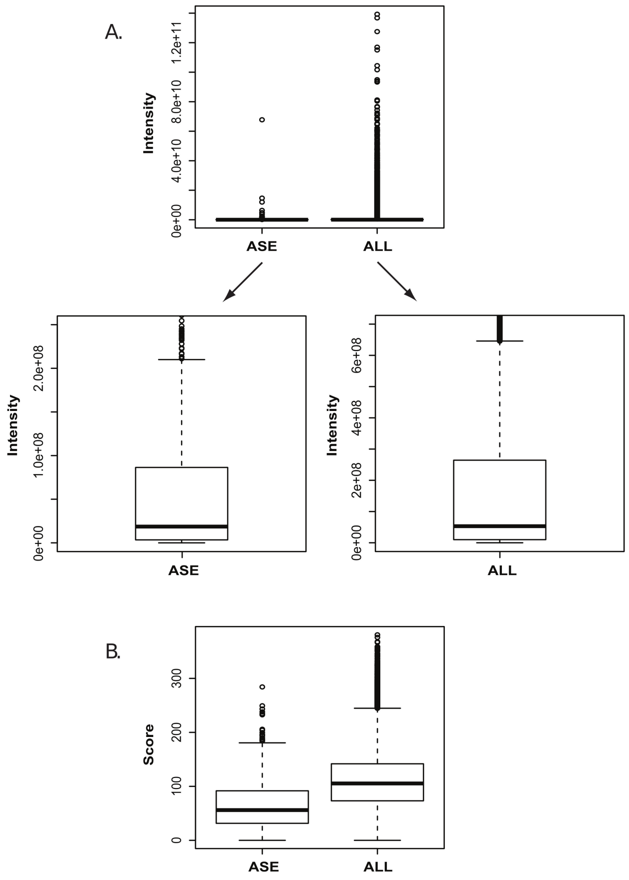


Figure 2. Peptide signal intensity (A) and scoring (B) distribution for all peptides (ALL) and sorted alternative splicing events (ASE) in the probabilistic approach. ASE peptides were on average an order of magnitude less abundant than the whole peptide population, consequently with lower average scoring.

We next plotted individual peptide intensities and scores from both the complete peptide dataset and peptides uniquely identified in Splooce. Data overview of the complete dataset showed, as previously reported, an intensity span of 7 orders of magnitude. The peptides characterizing the rare ASEs were observed mostly at the bottom half of the intensity distributions, with an average distribution approximately one order of magnitude lower than the complete Uniprot set (Figure 2A). While the score distribution seemed similar, ASE-derived peptides, on average, had a lower distribution (Figure 2B), which could be a consequence of poorer MS/MS from lower intensity ions.

In addition, the same RAW files collection was submitted to PEAKS, a software capable of determining a MS/MS sequence without the support of a database. Since no FDR can be estimated without the support of reversed sequences artificially created from a database, this analysis was restricted to spectra where fragment ion mass sequences could be measure with an average confidence of at least 80%. Using this approach, approximately 50,000 peptides were identified in Uniprot and Splooce (data not shown), and from those only 236 peptides, confirming 218 splicing events, could be identified in the same Splooce-derived database as used in the probabilistic approach. From those, 134 ASE were already observed in the probabilistic approach. By merging the results of the two strategies, we characterized a total of 892 ASE (Supplementary File S2 and S3).

As expected, the *de novo* method identified a smaller proportion of proteins and peptides than the probabilistic method when submitted to a BLAST-like alignment versus the same Splooce database. In fact, a smaller number of splicing events were detected in the *de novo* method when compared to the probabilistic one. An explanation for this could be that since most ASE events characterized by the probabilistic method are seen in the bottom part of signal intensity, they most probably generated partial MS/MS information that did not fulfilled the criteria required by us for reporting good quality *de novo* sequences. With this observation we therefore conclude that performing a probabilistic method using an in house database generates more information than *de novo* sequencing.

The frequency of each type of alternative splicing was next calculated for all events identified in our strategy. Simple events like exon skipping, alternative splice borders and intron retention showed proportional frequencies when compared to general Splooce statistics (Table 1). Moreover, no ASEs resulting from dual-specificity splice sites were identified, since these events are very uncommon and usually found within UTR sequences (Zhang et al., 2007). Splooce is also a database that focus on the analysis of combined ASEs (CASEs), and it was previously shown that approximately half of all alternative expressed sequences may have more than one ASE along their sequences (Kroll et al., 2012). The analysis presented here confirms the same finding at the proteome level. The most frequent combined event was the skipping of several adjacent exons (up to 11 exons), followed by adjacent alternative splice sites.

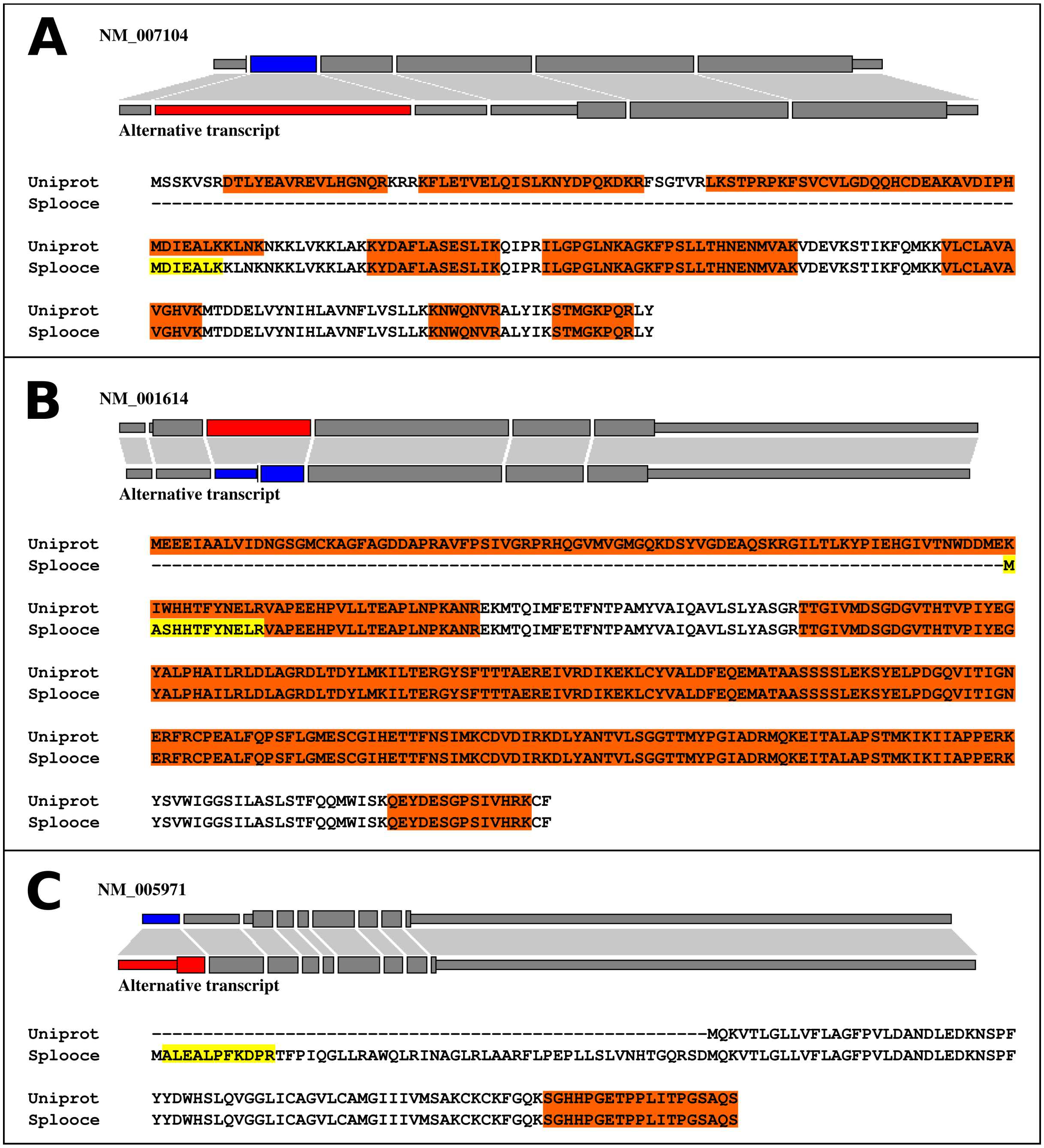
Alternative Splicing Event	**Total Events from Splooce**	**Events identified by the MS/MS analysis**
Exon skipping	38060 (35%)	182 (39%)
Alternative 3' splice site	30172 (29%)	130 (28%)
Alternative 5' splice site	27585 (25%)	90 (20%)
Intron retention	12632 (11%)	61 (13%)
Dual-specific splice site	112 (0%)	0 (0%)

Table 1. Amount of alternative splicing events identified by the MS/MS analysis compared to the total number of events available from the Splooce database.

Alternative TIS represents the majority of events at the proteome level

We further explored what types of events were observed in the identified peptides. Interestingly, 355 ASEs, out of the 892 (40%), showed a pattern consistent with the use of an alternative TIS due to an ASE (Figure3, Supplementary File S2). The remaining 537 proteins showed different types of variations along their protein sequences (Supplementary file S3). Files S2 and S3 not only contain a resumed version of the results described in this section, but also report protein sequence alignments for Uniprot and Splooce sequences of all proteins identified with a rare ASE. Peptides shared between both databases, in addition to the Splooce-specific peptide(s), are highlighted in the alignment. Most importantly, each alignment contains a link to the Splooce website where information and statistics for that rare ASE can be collected.

The high proportion of alternative TIS was further explored. All new protein isoforms showing an alternative TIS were searched against the TISdb database (Wan & Qian 2014), a collection of TIS obtained from a genome-wide method (GTI-Seq) developed by the same authors (Lee et al., 2012). We found that only one TIS present in our list was present in the TISdb providing therefore a proteome validation for that respective TISdb entry. Several reasons could explain the small overlap between the two datasets such as: i) the different nature of the samples used in both studies, ii) the fact that most of the TIS present in TISdb are non-canonical and start with others codons than ATG (we restricted our analysis to ATG-associated TIS) and iii) the lack of proteome validation in most of the studies that populated TISdb.



```
Uniprot     MSSKVSRDTLYEAVREVLHGNQRKRRKFLETVELQISLKNYDPQKDKRFSGTVRLKSTPRPKFSVCVLGDQQHCDEAKAVDIPH
Splooce     ------------------------------------------------------------------------------------

Uniprot     MDIEALKKLNKNKKLVKKLAKKYDAFLASESLIKQIPRILGPGLNKAGKFPSLLTHNENMVAKVDEVKSTIKFQMKKVLCLAVA
Splooce     MDIEALKKLNKNKKLVKKLAKKYDAFLASESLIKQIPRILGPGLNKAGKFPSLLTHNENMVAKVDEVKSTIKFQMKKVLCLAVA

Uniprot     VGHVKMTDDELVYNIHLAVNFLVSLLKKNWQNVRALYIKSTMGKPQRLY
Splooce     VGHVKMTDDELVYNIHLAVNFLVSLLKKNWQNVRALYIKSTMGKPQRLY
```

```
Uniprot     MEEEIAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEK
Splooce     -----------------------------------------------------------------------------------M

Uniprot     IWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEG
Splooce     ASHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEG

Uniprot     YALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSYELPDGQVITIGN
Splooce     YALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSYELPDGQVITIGN

Uniprot     ERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERK
Splooce     ERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERK

Uniprot     YSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF
Splooce     YSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF
```

```
Uniprot     ------------------------------------------------------MQKVTLGLLVFLAGFPVLDANDLEDKNSPF
Splooce     MALEALPFKDPRTFPIQGLLRAWQLRINAGLRLAARFLPEPLLSLVNHTGQRSDMQKVTLGLLVFLAGFPVLDANDLEDKNSPF

Uniprot     YYDWHSLQVGGLICAGVLCAMGIIIVMSAKCKCKFGQKSGHHPGETPPLITPGSAQS
Splooce     YYDWHSLQVGGLICAGVLCAMGIIIVMSAKCKCKFGQKSGHHPGETPPLITPGSAQS
```

Figure 3. Alignments between normal (Uniprot/RefSeq) and alternative (Splooce) proteins, showing different categories of alternative TIS observed for our data. Sequences highlighted in orange represent MS peptides found for the Uniprot/RefSeq proteins, and sequences highlighted in yellow represent peptides found exclusively in the alternative sequences from Splooce. Peptides that align specifically to a sequence from Splooce are supposed to characterize ASEs. A: Alternative TIS is downstream the original one; B: Same as A, although the beginning of the

protein sequence is directly affected by the ASE. C: Alternative TIS is upstream the original one.

Wilson and colleagues have suggested that the association between ASE and TIS are restricted to the amino-terminus of proteins where both events are used to produce isoforms that differ at their amino end. Almost 2,000 events like that were identified at the transcriptome level but few (17 instances) were confirmed in a limited search against MS/MS data (Wilson et al., 2014). We wondered whether this type of event would be frequent in our dataset of 355 TIS. Visual inspection of all 355 cases identified only 29 instances (8%) that would fit the model from Wilson et al., (2013) (for more details, see Supplementary file S2). The low level of validation of such cases at the proteome level, also seen by the authors in their original report, raises doubts about their widespread occurrence. All remaining 326 cases of TIS in our dataset were analyzed to identify the effect of the ASE in the protein sequence originally present in the reference sequence. In only three cases, the alternative TIS was upstream of the original ATG codon. In all remaining cases, the ASE occurred upstream of the alternative TIS and disrupted the respective ORF. An alternative ATG codon, always located downstream of the ASE, is then used as a new TIS. Interestingly, only in 15% of these cases (48 out of 323) the ATG codon used in the TIS is the first one downstream of the ASE.

CONCLUSIONS

A limitation one is facing in this type of analysis the definition of a proper false discovery rate when adding entries in a database *ad infinitum*. Any observed MS/MS information in such approaches will be tagged to the “best-fit” theoretical peptide present in the database, regardless if that is the correct one. Even though identification engines such as Mascot and MaxQuant have proof-check algorithms to quantify FDR rate, incorrect MS/MS information might still be reported as true. Therefore there will be always the risk that peptides that are present in the sample but not represented in the database are incorrectly assigned. In addition, there will be a size limit where adding more protein entries created by RNAseq information will be detrimental to the analysis, rather than beneficial. For a good isoform discovery phase study to reliably work, a compromise between database size and validation rounds using complementary databases must be created. A desirable strategy would be to create a collection of public, high quality datasets such as the one used in this work and use them for database-based splicing discovery using different versions of the Splooce database. Recently, similar approaches have been successfully implemented for mapping expressed genes, pseudogenes and characterization of new open reading frames (Kim et al., 2014; Wilhelm et al., 2014), but little was shown regarding splicing isoforms. Therefore, such approach using Splooce databases with public MS data for ASE discovery is feasible and promising for further characterization of the human proteome draft.

In summary, a new strategy for the identification of splicing variants in MS/MS data is provided here allowing us to confirm at the proteome level more than 800 new variants. We extended previous observations linking ASE and TIS and provided validation for hundreds of new TIS events. We have upgraded the Splooce portal to take into account the integration of MS/MS data in the validation of splicing variants.

References

Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, van Bakel H, Schadt EE, Reijo-Pera RA, Underwood JG, and Wong WH. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci U S A* 110:E4821-E4830.

Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J, and Aebersold R. 2011. The quantitative proteome of a human cell line. *Molecular systems biology* 7:549.

Blakeley P, Siepen JA, Lawless C, and Hubbard SJ. 2010. Investigating protein isoforms via proteomics: a feasibility study. *Proteomics* 10:1127-1140.

Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, Verstraten R, Adams DJ, Harrow J, Choudhary JS, and Hubbard T. 2011. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res* 21:756-767.

Consortium TGP. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.

Cork DMW, Lennard TWJ, and Tyson-Capper AJ. 2012. Progesterone receptor (PR) variants exist in breast cancer cells characterised as PR negative. *Tumour Biol* 33:2329-2340.

Cox J, and Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* 26:1367-1372.

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, and Aebersold R. 2006. The PeptideAtlas project. *Nucleic Acids Res* 34:D655-658.

Geiger T, Wehner A, Schaab C, Cox J, and Mann M. 2012. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* 11:M111 014050.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, and Pandey A. 2014. A draft map of the human proteome. *Nature* 509:575-581.

Kroll JE, Galante PA, Ohara DT, Navarro FC, Ohno-Machado L, and de Souza SJ. 2012. SPLOOCE: a new portal for the analysis of human splicing variants. *RNA Biol* 9:1339-1343.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, and International Human Genome Sequencing C. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

Lee S, Liu B, Lee S, Huang SX, Shen B, and Qian SB. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* 109:E2424-2432.

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, and Lajoie G. 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 17:2337-2342.

Magrane M, and Consortium U. 2011. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011:bar009.

Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, and Mann M. 2011. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* 7:548.

Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, and et al., 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology* 30:253–260.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen Fc, ,oise, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, and Ostell JM. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 42:D756-D763.

Severing EI, van Dijk AD, and van Ham RC. 2011. Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. *BMC Plant Biol* 11:82.

Sheynkman GM, Shortreed MR, Frey BL, and Smith LM. 2013. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol Cell Proteomics* 12:2341-2353.

Tress ML, Bodenmiller B, Aebersold R, and Valencia A. 2008. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol* 9:R162.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, and Zhu X. 2001. The sequence of the human genome. *Science* 291:1304-1351.

Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, O'Kelly G, Schoenegger A, Ovelleiro D, Perez-Riverol Y, Reisinger F, Rios D, Wang R, and Hermjakob H. 2013. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 41:D1063-1069.

Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolome S, Apweiler R, Omenn GS, Martens L, Jones AR, and Hermjakob H. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* 32:223-226.

Wan J, and Qian SB. 2014. TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Res* 42:D845-850.

Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, and Zhang B. 2012. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J Proteome Res* 11:1009-1017.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, and Kuster B. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509:582-587.

Wilson LO, Spriggs A, Taylor JM, and Fahrer AM. 2014. A novel splicing outcome reveals more than 2000 new mammalian protein isoforms. *Bioinformatics* 30:151-156.

Xue J, Schmidt SV, Sander J, Draffehn A, Krebs W, Quester I, De Nardo D, Gohel TD, Emde M, Schmidleithner L, Ganesan H, Nino-Castro A, Mallmann MR, Labzin L, Theis H, Kraut M, Beyer M, Latz E, Freeman TC, Ulas T, and Schultze JL. 2014. Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity* 40:274-288.

Zhang C, Hastings ML, Krainer AR, and Zhang MQ. 2007. Dual-specificity splice sites function alternatively as 5' and 3' splice sites. *Proc Natl Acad Sci U S A* 104:15028-15033.