SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data

Yafeng Zhu, Lina Hultin-Rosenberg, Jenny Forshed, Rui M M Branca, Lukas M Orre & Janne Lehtiö

Cancer Proteomics Mass Spectrometry, Science for Life Laboratory, Karolinska Institutet, Stockholm, Sweden

Mailing address:
Yafeng Zhu: yafeng.zhu@ki.se
Lina Hultin-Rosenberg: lina.hultin-rosenberg@ki.se
Rui M M Branca: rui.mamede-branca@ki.se
Lukas M Orre: lukas.orre@ki.se
Jenny Forshed: jenny.forshed@ki.se
Janne Lehtiö: janne.lehtio@ki.se

To whom correspondence should be addressed. Email: janne.lehtio@ki.se
RUNNING TITLE
Splice variant analysis in MS based proteomics

LIST OF ABBREVIATIONS:

MS  Mass spectrometry
LC  Liquid chromatography
EST Expressed sequence tag
PSM Peptide spectrum match
iTRAQ Isobaric tags for relative and absolute quantification
LTQ Linear quadropole ion trap
PQPQ Peptide quantification by peptide quality control
snRNP Small nuclear ribonucleoprotein particle
EVDB Evidence Viewer Database
RT-PCR Reversed transcription polymerase chain reaction
SVSP Splice variant specific peptide
MSF Magellan storage file
FDR False discovery rate
iPS cells Induced pluripotent stem cells
hESCs Human embryotic stem cells
Summary

Alternative splicing is a pervasive process in eukaryotic organisms. More than 90 percent of human genes have alternatively spliced products and aberrant splicing has been shown to be associated with many diseases. Current methods employed in the detection of splice variants include prediction by clustering of expressed sequence tags (ESTs), exon microarray and mRNA sequencing, all methods focusing on RNA level information. There is a lack of tools to analyze splice variants at the protein level. Here, we present SpliceVista, a tool for splice variant identification and visualization based on mass spectrometry (MS) proteomics data. SpliceVista retrieves gene structure and translated sequences from alternative splicing databases and maps MS identified peptides to splice variants. The visualization module plots the exon composition of each splice variant and aligns identified peptides to its transcript positions. If quantitative MS data is used, SpliceVista plots the quantitative patterns for each peptide and provides users the option to cluster peptides based on their quantitative patterns. SpliceVista can identify splice variant specific peptides (SVSPs) hence providing the possibility for variant specific analysis. The tool was tested on two experimental datasets (PXD000065, PXD000134). In A431 cells treated with gefitinib, 2983 SVSPs corresponding to 939 splice variants were identified. By comparing splice variant centric, protein centric and gene centric quantification, several genes (e.g. EIF4H) were found to have differentially regulated splice variants after gefitinib treatment. The same discrepancy between protein centric and splice centric quantification was detected also in the other dataset, in which induced pluripotent stem cells were compared with parental fibroblast and human embryotic stem cells. In addition, SpliceVista can be used to visualize novel splice variants inferred from peptide level evidence. In summary SpliceVista enables visualization, detection and differential quantification of protein splice variants that are often missed by current proteomics pipelines.
Introduction

Eukaryotic genes are composed by exonic (protein coding) and intronic (non-coding) regions. Alternative splicing is a process in which pre-mRNA is cut at junction sites and the resulting exonic sequences are reconnected together in different ways to form different versions of mature messenger RNA. It has been shown that 92-94% of human genes can undergo alternative splicing (1, 2). This process plays an essential role in increasing the proteome diversity in eukaryotic organisms. For multiexon mRNAs, different splicing patterns can occur, e.g. exon skipping (an exon is either included or excluded from the mature mRNAs), alternative 5' or 3' splicing (exons are spliced in different lengths) or mutually exclusive splicing (exons are selectively spliced to be exclusively present in different splice forms). Alternative splicing is carried out by the spliceosome, which consists of five small nuclear ribonucleoprotein particles (snRNPs), U1, U2, U4, U5, U6 and more than 150 other proteins (3). Mutations in splicing sites or in the main components of the splicing machinery will affect the genes’ splicing patterns, and potentially give rise to alternative protein products which may have different conformation, function or subcellular location. Disruption of the splicing machinery has been shown to be associated with many human diseases such as cystic fibrosis, Alzheimer disease and cancer (4-6).

Large efforts have been put into identification of gene products generated by alternative splicing. This is a challenging task since alternative splice forms are often temporal, tissue specific and low abundant (7). So far most work has been done starting at the mRNA level, taking use of the vast amount of public domain expressed sequence tag (EST) data as well as RNA sequencing data (8-11). ESTs or RNA-seq reads that belong to one gene are clustered together and then aligned to the genomic sequence in order to identify alternative splicing events. These efforts have resulted in many publically available alternative splicing databases. Most of them are generated by mining data from GenBank, UniGene and
SwissProt. (i) The Evidence Viewer Database (EVDB) is one of the relational databases which support searching splice variants by sequence and gene symbol querying (12). This database uses high quality transcripts from NCBI Genbank and RefSeq, which are then aligned to the chromosomal sequence in order to determine their exon structures. EVDB contains 81142 non-redundant human splice variants in the most recent build (completed in June 2010). (ii) The ECgene database is an alternative splicing database which is constructed by genome based EST clustering. Splice variants in the database are assigned with different evidence levels based on the minimum number of clones used to cover the transcripts (13).

Mass spectrometry (MS) based proteomics enables large-scale identification and quantification of proteins. The most commonly used workflow for MS based proteomics is the so-called bottom up approach or shotgun proteomics where proteins are digested into peptides to facilitate efficient MS analysis. Bioinformatic methods are then used to infer the protein level events (14). A challenge in shotgun proteomics is the protein inference problem, which refers to the task of determining which proteins the identified peptides are derived from. The difficulty is due to that some peptides are shared by several proteins. Once the protein mixture is digested by a protease, peptides from all proteins are mixed together and the protein context of each and every peptide is lost. This will lead to ambiguities in the identification of proteins present in the sample. The existence of several protein isoforms (e.g. alternative splicing forms) further complicates the identification process, since protein isoforms usually have very similar sequence. After tryptic digestion, it is impossible to distinguish between different protein isoforms with absolute certainty if no splice variant specific peptides (SVSPs) are identified. Nevertheless, MS based proteomics has been used to identify known and novel splice variants by including the sequence of known and predicted protein variants in the search database (15-17).
In quantitative proteomics the protein inference problem also affects the accuracy of protein quantification since protein quantity measurements can be compromised by peptides that are wrongly assigned to a particular protein or protein variant. To address this problem we recently developed a tool, PQPQ (protein quantification by peptide quality control) (18), to detect protein variants in MS based shotgun proteomics data. This method is based on the assumption that peptides derived from a given protein variant will have a correlated quantitative pattern over samples. PQPQ takes all high confidence peptide spectrum matches (PSMs) and clusters them based on their quantitative pattern over samples. PSMs derived from different protein isoforms that are differentially expressed or regulated, will have different quantitative patterns and will consequently be grouped in different clusters by PQPQ. PQPQ may thus detect protein variants based on quantitative patterns, even in cases where the database search of the MS/MS data has failed to detect those protein variants.

Here we present a novel tool, SpliceVista, which enables and facilitates splice variant centered interrogation of shotgun proteomics data. SpliceVista retrieves gene structure and translated sequences from two alternative splicing databases, EVDB and ECgene, and maps identified peptides to splice variants. The visualization module plots the exon composition of each splice variant and aligns identified peptides to its transcript positions. If quantitative MS data is used, SpliceVista plots the quantitative patterns for each peptide. In addition, a simplified version of the PQPQ-algorithm is included in the package to provide users the option to cluster peptides based on their quantitative patterns. Since the splice variants effect gene function, and aberrant splicing forms have been shown to be related with many human diseases such as cancer(4-6); we envision that SpliceVista will be an important tool in splice variant associated biomarker discovery and biological research on variant specific proteome changes.
Materials and Methods

Algorithm Availability and Requirements

SpliceVista was written in Python 2.7.2. It consists of five modules: converter.py, mergepsm.py, download.py, mapping.py and visualization.py. It also includes a simplified version of PQPQ (named clusterpeptide.py) which mainly does the peptide clustering. A detailed manual for the program can be found in Supplementary file 1. The program is free to use and can be downloaded from https://github.com/yafeng/SpliceVista.

Preprocessing of MS data

The following information needs to be extracted from the MS output from database searching: protein accession id, peptide sequence and quantitative data (if available). The gene symbol is then assigned to each protein by the python script converter.py and is used to retrieve known protein splice variants from the EVDB database.

SpliceVista workflow

SpliceVista is designed to identify and visualize splice variants based on MS-identified peptides. There are four main parts of the program (Fig.1):

1. Data preprocess. In this step, all PSMs will be assigned a gene symbol from its protein ID and grouped into peptides.

2. Download. SpliceVista uses the gene symbol to retrieve all known splice variants of a gene from the EVDB database. The identifiers of splice variants used in EVDB are consistent with those in Genbank. Nucleotide sequences are extracted from Genbank by their IDs and then translated into amino acid sequences.

3. Mapping. In this step, all identified peptides are grouped by gene, and genes are
analyzed one by one. For each gene, all the identified peptides are mapped to the gene’s splice variants from the EVDB or ECgene databases. The genomic and transcript positions of each peptide are reported in the output file.

4. **Visualization.** The data from previous steps is used for visualizing the exon structures of each splice variant lined up with the identified peptides. The splice variants in EVDB are visualized with only exons scaled to size. The predicted splice variants derived by ECgene database (including both known, defined as present in Ensembl 72, and unknown ones, only in ECgene database) are visualized with intron and exon scaled to corresponding size. In addition, if PQPQ is used, the peptide clusters based on quantitative patterns are visualized allowing connection between specific peptides and detected quantitative peptide clusters.

*Suggested position for Fig.1*

**Output files from SpliceVista**

There are two important output files from SpliceVista: mappingout.txt and genestatistic.txt. See Supplementary file 2, Table S1 and Table S2. The genomic and transcript positions, quantitative data and PQPQ clustering results of each peptide can be found in the mappingout.txt file. Other files, subexons.txt, splicingvar.txt, and varseq.fa, which are retrieved from the databases (EVDB and GenBank) are necessary for mapping peptides in the visualization module. See the User Manual (Supplementary file 1) for detailed information.

**Visualization module of SpliceVista**

Given a gene symbol for the protein, SpliceVista (visualization.py) can generate an image that contains three panels (Fig.2). The top panel displays the exon structure of all known
splice variants. The mid panel displays the transcript positions of identified peptides. If PQPQ is applied, each peptide is assigned to a cluster in which all peptides show correlated quantitative pattern. In the bottom panel, the quantitative patterns of the different clusters are drawn in the same order as in the mid panel. The bars represent the mean intensity ratio of all peptide spectrum matches (PSMs) for each unique peptide, the standard deviation being indicated by vertical lines (error bars).

*Suggested position for Fig.2*

**In silico analysis of all human protein isoforms**

A list of 21,494 protein-coding genes was downloaded from Ensembl 63. Of those, 18,372 genes have splice variants in the EVDB database, *i.e.* 76,827 splice variants in total. 3072 of them belong to genes with only one known splice variant in the database. *In silico* trypsin digestion of the human proteome (Ensembl 63) was performed, resulting in 832,421 unique peptides (6aa=<peptide length<= 40aa) that were later mapped to all splice variants. To compare with trypsin, *in silico* lysC digestion of human proteome, which generates on average longer peptides, was also performed yielding 447,880 unique peptides (6aa=<peptide length<= 40aa). In the simulated protease digestion, one missed cleavage was allowed only in cases of consecutive cleavage sites (KK, KR, RK, or RR for trypsin, and KK for lysC), and no cutting was done if the cleavage site was followed by Proline.

**A431 human cell line proteomics data**

*Sample preparation*

In order to exemplify the key functions of SpliceVista, it was used to analyze A431 human cell line (epidermoid carcinoma cell line, cell line number ACC 91) proteomics data. The sample preparation and mass spectrometry experiment was done in (19). Briefly, twenty-four
hours after seeding, A431 cell cultures (in duplicates) were treated with gefitinib and harvested at 2h, 6h and 24h after treatment. Controls were left untreated (duplicates at 0h). Protein samples from A431 whole cell extraction and three subcellular fractions were then digested by trypsin (see experimental setup and subcellular fractionation procedures in Supplementary file 2, Fig.S1). The resulting peptide mixtures were arranged into four sets (whole, light, medium and heavy) and peptides at different time points in each set were labeled with 8plex-iTRAQ (ABsciex). Peptide mixtures (200 µg) in each set were separated by isoelectric focusing (20) using five different IPG (immobilized pH gradient) gel strips (provided by GE Healthcare Bio-Sciences AB, Uppsala, Sweden; pH ranges of the five strips were 3.7-4.9, 3.70-4.05, 4.00-4.25, 4.20-4.45 and 4.39-4.99, all of them 24 cm long). After isoelectric focusing completion, each IPG strip was divided into 72 fractions and peptides from each fraction were transferred into a 96-well micro-titer plate by liquid handling robotics (GE Healthcare prototype) and dried in a speedvac. Five MS-based experiments (corresponding peptides collected from five IPG gel strips) were performed using a hybrid LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific) for each of the four peptide mixtures sets (whole cell, light, medium and heavy fraction). Detailed mass spectrometry analysis can be found in (19).

**Mass spectrometry data processing**

1. Searching Ensembl 63 human protein database

All MS/MS spectra were searched by Sequest/Percolator under the software platform Proteome Discoverer (PD, v1.3.0.339, Thermo Scientific) using a target-decoy strategy. The reference database used was the human protein subset of Ensembl 63 (76,501 protein entries). Precursor mass tolerance of 10 ppm and product mass tolerances of 0.02 Da for HCD-FTMS and 0.8 Da for CID-ITMS were used. Additional settings were: trypsin with 1 missed cleavage; carbamidomethylation on cysteine and iTRAQ-8plex on lysine and N-terminal as
fixed modifications; and oxidation of methionine and phosphorylation on serine, threonine or tyrosine as variable modifications. Quantification of iTRAQ-8plex reporter ions was done using an integration window tolerance of 20ppm. PSMs found at 1% FDR (false discovery rate) were exported.

Reporter ion based quantification of proteins was done following Proteome Discoverer's default settings: only PSMs from unique peptides and with precursor interference < 50% were used for quantification; the quantitative ratios of each PSM were normalized to have the same protein median ratios between iTRAQ channels. The protein tables of all identified proteins (at 1% FDR) and their quantitative data can be found in Supplementary file 3. The raw data, pep.XML and Proteome Discoverer MSF files associated with this manuscript are available in the repository ProteomeXchange (dataset ID: PXD000065).


ECgene splice variant database (at high and low evidence level) was downloaded from http://genome.ewha.ac.kr/ECgene/ (see peptide overlap of ECgene databases and Ensembl 72 in Supplementary file 2, Fig.S2). MS/MS spectra from A431 whole cell (five MS experiments, peptides separated on IPG gel strips of pH ranges 3.7-4.9, 3.70-4.05, 4.00-4.25, 4.20-4.45 and 4.39-4.99) were searched against two different databases: ECgene database (high evidence level) concatenated with the Ensembl 72 database, and ECgene database (low evidence level) concatenated with the Ensembl 72 database. The same software and parameters were used as described above except that the only variable modification used was oxidation of methionine. Peptide and protein quantification was not included in this workflow.

Results

Human protein isoforms with unique sequences and isoform specific peptides
To evaluate the potential and limitations of bottom up mass spectrometry based proteomics to do splice variant specific analysis, we performed a theoretical analysis. In the simulated trypsin digestion, 146,818 (18%) tryptic peptides uniquely mapped to a specific splice variant, i.e. they are splice variant specific peptides (SVSPs) (Fig.3). 65,916 of them mapped to a gene with only one known splice variant. Conversely, 16,935 (22%) splice variants were shown to have SVSPs. Because lysC generates longer peptides than trypsin (see peptide length distribution by trypsin and lysC digestion in Supplementary file 2 Fig.S3), one could expect a better coverage of splice junction sites by lysC digestion. However, the proportion of SVSPs produced by lysC digestion (17% SVSPs corresponding to 13,593 splice variants) was not higher than that of those produced by trypsin. This result can in part be explained by the fact that many lysC peptides will be too long (>40 aa) to be detected by LC-MS analysis.

Combining peptides generated by both trypsin and lysC resulted in a modest increase in the number of isoforms with SVSPs (23%) compared to trypsin alone. These results indicate that three out of four protein isoforms cannot be uniquely identified (Fig.3) using these two enzymes.

Identification of splice variants in A431 cell line data

As described previously, a splice variant is identified if there is one or more peptides uniquely mapped to its sequence. To test the applicability of the method on proteomics data generated by MS-based shotgun proteomics, we used SpliceVista to analyze human cancer cell line data (A431). In the whole cell lysate, 607 unambiguous splice variants and 1680 SVSPs were identified (Table 1). All SVSPs reported in A431 dataset are derived from genes with multiple splice variants (see these SVSPs and their mapping output from SpliceVista in Supplementary file 4). SVSPs from single isoform genes were not counted. By subcellular
fractionation, the number of unique peptides identified increased, as did the number of SVSPs and corresponding splice variants (Fig.4). Expectedly, this data demonstrates that by using subcellular fractionation, we can increase SVSPs and splice variant identifications due to increased peptide coverage (see sequence coverage with and without subcellular fractionation in Supplementary file 2, Fig.S4).

**Suggested position for Table 1**

**Suggested position for Fig.4**

**Detection of differentially regulated splice variants in A431 data**

The A431 dataset contains quantitative information generated by iTRAQ 8-plex labelling of samples from a gefitinib treatment time course study. Duplicate samples were taken at 2, 6 and 24h after gefitinib treatment and compared with duplicate untreated controls, and reported as ratios using the average of the controls as denominator. In the current study, we performed three different quantitative analyses on the genes with splice variants identified from this dataset (Table 1); gene centric analysis, protein centric analysis and splice variant centric analysis (Fig.5). In the gene centric analysis, the relative expression level of a gene is calculated as the mean ratio of all PSMs identified for this gene. In protein centric analysis, the conventional approach in proteomics, the relative expression level is calculated as the mean ratio of all PSMs identified by the search engine for this protein. In splice variant centric analysis, the relative expression level of a certain splice variant of a gene is calculated by taking the mean ratio of PSMs specific to that splice variant. The difference to the conventional protein centric analysis is that only PSMs that uniquely map to a splice variant are used to quantify it. Since more than 90% of genes can undergo alternative splicing (1, 2), there is a potential risk of averaging out the differences of differentially regulated splice
variants when doing gene or protein centric quantitative analysis if the gene or protein contains peptides shared among splice variants.

Three genes identified in the A431 dataset were selected to exemplify three typical situations where different results are observed depending on whether gene centric, protein centric or splice variant centric analysis are used (Fig.5): i) down regulation of gene EIF4H is only seen in splice centric analysis while no obvious regulation is observed from neither gene centric nor protein centric analysis (Fig.5A). The protein reported (ENSP00000265753) in the conventional protein centric analysis contains the SVSPs used for splice variant centric analysis. The distinct quantitative pattern detected by the splice variant centric analysis implies that at least one additional unreported splice variant is present. Moreover the unreported variant is highly abundant compared to the variant identified by SVSPs. When doing gene centric or protein centric analysis, the unreported but dominant splice variant averages out the down-regulation signal of the identified splice variant. ii) gene centric, protein centric and splice centric analysis all showed different results for the ITGB4 gene (Fig.5B). Here, three different PSM populations were found by protein centric analysis, corresponding to two different variants (ENSP00000200181, ENSP00000344079) and a protein group containing shared PSMs between these two variants. In addition, the quantitative pattern of that protein group indicates that at least one additional variant could be present and dominant. iii) for the SHROOM3 gene, two variants with different quantitative patterns are found by protein centric analysis. One of these variants contains an SVSP (2 PSMs) mapping to the variant reported in the splice centric analysis (Fig.5C). However, the other variant (37 PSMs) is the highly abundant one, and thus has the major contribution to the gene centric signal. In all three cases, the gene centric quantification result is an averaged outcome of all identified PSMs mapped to the gene and is dominated by the protein variant which contains the most PSMs. With SpliceVista, we are able to quantify splice variants
specifically and in some cases infer hidden variants by comparing splice centric analysis to
gene centric and protein centric analysis (see the comparison for all genes from heavy
fraction with SVSP identified in Supplementary file 2, Fig. S6).

A PQPQ based quantitative clustering combined with peptides’ transcript position
information was investigated to see if peptides uniquely mapped to gene EIF4H’s splice
variant (NM_022170) have correlated quantitative pattern. As shown in Fig. 6, three peptides
(4, 5 and 6) were clustered together and showed down-regulated at 24h compared to the other
peptides. As seen by the transcript positions of the peptides, peptide 4 (DDFNSGFR) and
peptide 5 (DDFNSGFRDDFLGGR) are unique to the splice variant NM_022170. Although
peptide 6 (DDFLGGR) is not uniquely mapped to NM_022170, it could be derived from this
splice variant based on its quantitative pattern. The other splice variant shared peptides show
no significant regulation; implying that splice variant NM_022170 of EIF4H is differentially
regulated. Since all peptides are assigned with genomic coordinates by SpliceVista, it is also
possible to compare protein level data with RNA level data. As shown in Supplementary file
2, Fig. S5, splice variant specific change of gene EIF4H identified at protein level is compared
with RNA-seq data.

Suggested position for Fig. 5

Suggested position for Fig. 6

Discovery and visualization of novel splice variant peptides in A431 data

SpliceVista can also be used to visualize novel protein isoforms not yet reported in the EVDB
database. To exemplify this feature we searched the A431 data against the ECgene database
which contains predicted splice variants based on EST data. In total, 31,985 and 31,023
unique peptides were identified at 1% FDR in the ECgene high (high evidence level)
concatenated with Ensembl 72 database and ECgene low (low evidence level) concatenated
with Ensembl 72 database, respectively. Of these, 30,708 peptides were identified in both searches. Since more or less the same number of peptides was identified in both searches, we focused on peptides identified from the high evidence level database. In that search 223 unique peptides were exclusively identified (with Xcorr>2) in ECgene high and not in Ensembl 72 database. SpliceVista was used to map the 223 peptides to their genomic positions (a list of these 223 peptides and their genomic coordinates is provided in Supplementary file 5). We then searched these 223 peptides in the NCBI human non-redundant protein sequence database using BLASTP (21). Of these, 157 peptides had at least one mismatch to the sequences in the database and were therefore considered novel. One of the peptides mapping to a novel splice variant of gene PLCB2 is shown in Fig.7.

**Suggested position for Fig.7**

**Splice variant specific analysis on human 4Skin hiPS cells, parental fibroblast cell lines and human ESCs**

To demonstrate SpliceVista’s compatibility with other proteomics datasets, the Munoz et al. 2011 stem cell dataset (22) was downloaded from ProteomeXchange (PXD000134). MSF files for two experiments in which 4Skin hiPS cells were compared to parental fibroblast cell line and hESCs were downloaded and opened in Proteome Discoverer (version 1.3, Thermo Electron). PSMs with 1% FDR cutoff were extracted for each experiment. Then the peptides were mapped to splice variants in EVDB (see statistics of SVSP identifications and comparison to A431 cell line dataset in Supplementary file 2 Table S3). The number of splice variants reported in the two mass spectrometry experiments was 390 and 397 respectively. On average, we have found SVSPs for 7% of the identified genes in this dataset. This is comparable to the A431 cell line dataset, in which we found SVSPs for 9% of the identified genes.
The number of overlapping proteins with SVSPs identified in both MS experiments from Munoz study was 296. Protein centric analysis of these proteins was compared to splice variant specific analysis. As shown in Fig.8A, for most proteins, protein centric analysis of Fibro/hiPS ratio is consistent with splice variant specific results. However, some proteins marked as red showed large differences of Log2(Fibro/hiPS) value between protein centric and splice centric analysis, indicating that peptides assigned to the protein may be derived from differentially regulated splice variants. The top 10 proteins ranked by Fibro/hiPS ratio differences between splice variant and protein centric analysis are shown in Fig.8B.

Suggested position for Fig.8

Discussion

Shotgun proteomics is being widely applied in large-scale characterization of proteins due to its intrinsic ability to systematically profile the entire protein complement in the sample, both qualitatively and quantitatively. However, the lack of tools for shotgun proteomics data analysis limits the exploration of the large amount of data generated, leading to many missed protein isoforms that are relevant to specific biological events.

The herein presented tool, SpliceVista, simplifies detailed analysis of known and predicted splice variants by enabling their visualization and by mapping peptide evidence to these variants. Splice variant specific peptides (SVSPs) are the key for both identification and quantification of splice variants by SpliceVista. Hence, splice variant specific analysis is limited to the genes with SVSPs identified. Increased protein sequence coverage obtained by pre-fractionation methods at subcellular, protein or peptide level will increase the likelihood of identifying more SVSPs. However, we acknowledge the limitation of using shotgun
proteomics to identify splice variants since at most one out of four splice variants can be uniquely identified even assuming 100% sequence coverage in the MS analysis.

In MS-based proteomics, the unique peptides for a gene could come from several splice variants that usually possess high sequence similarity. Unless the gene has only one splice variant, it should be noted that in gene centric analysis a gene’s quantitative pattern is a mixed outcome of all present splice variants. Consequently, the quantitative pattern can thus be dominated by one or a few highly abundant splice variants, since most copies of the peptides identified come from those abundant protein variants. As shown in this study, the same problem can occur in protein centric analysis due to ambiguous assignment of peptides to protein variants. SVSPs mapped by SpliceVista provide the possibility to do splice variant centric analysis at the protein level. The quantification of a splice variant is then done by quantification of its SVSPs. In practice, we most often have unique peptides from only one splice variant of the gene. Nevertheless, differentially regulated hidden splice variants can still be detected indirectly by comparing gene centric or protein centric to splice variant centric analysis as here illustrated.

SpliceVista also enables remapping of peptide data to splice variant database, such as EVDB, and thus splice variant specific analysis can be performed on already generated peptide data. This makes SpliceVista compatible for analysis and re-analysis of most MS-based proteomics datasets. Additionally, SpliceVista can be used to map and visualize novel splice variant peptides identified via searching MS data against the ECgene database. When including predicted splice variants in the peptide search space, it is important to be aware of the following potential problems: the increased search space tends to increase false discoveries; also the expected low occurrence of findings in the novel (predicted) part of the database can
lead to increase FDR. It is therefore important to validate these novel splice variant identifications with independent experimental evidence.

In summary, the presented program, SpliceVista, can assist users in identification and quantification of splice variants in MS based proteomics. First, the program reports the number of known splice variants of the gene and aligns identified peptides to their transcript positions. Given this information, users can easily screen out the peptides unique to splice variants. Second, the given genomic coordinates of each peptide make it possible to compare with the results from RNA level experiments such as RT-PCR and RNA-seq. Third, the visualization feature of SpliceVista can help users interpret MS based proteomics data for a specific gene associated with splice variant information. If peptide clustering by PQPQ is applied, SpliceVista will also present clustering results and histograms of the different quantitative patterns detected. This information combined and visualized by SpliceVista, enables users to identify and evaluate splice variant specific quantitative patterns and to infer alternative splicing regulation. With these features, SpliceVista will serve as a tool to explore splice variant specific information from high throughput proteomics data and generate alternative splicing associated hypotheses.

Acknowledgements

The authors thank Henrik Johansson for performing the MS experiments on the A431 cell line samples and Mikael Huss for his help in the analysis of RNA sequencing data of A431 cell line samples. This project was financed with support from Karolinska Institutet (KID-funding), Swedish research council, Swedish Cancer Society, the Cancer Society in Stockholm, Stockholm County Council and FP7 EU project GlycoHit.
References:

Fig.1. **Workflow of SpliceVista.** The blue boxes explain the four main steps of SpliceVista. The yellow boxes depict the detailed workflow of SpliceVista. Given the peptide data, converter.py assign each protein ID a gene symbol which is used to retrieve known splice variants of this gene and its exon structure in EVDB. Peptide sequences are mapped to the translated sequence of splice variants retrieved from GenBank. Genomic coordinates of the peptides and their transcript positions are reported in the output. Known splice variants are identified by splice variant specific peptides that map uniquely to the splice variants. Quantification of splice variants can then be done by quantification of splice variant specific peptides.

Fig.2. **SpliceVista visualization overview.** Figure showing the SpliceVista output picture of BAX gene detected in A431 whole cell fraction. In the top panel, the exon composition of the gene is depicted and the gene symbol is written at the upper left corner. The white boxes are all sub-exons of the gene present in the database. The beige boxes show the exon compositions of splice variants and their accession numbers are marked on the left. The shadow on the transcript indicates the location of identified peptides. Below the transcript variant, in the mid panel, the colored lines are identified peptides which are aligned to corresponding positions on transcripts. The number underneath the lines is a numbering given to the peptide (the numbering is sorted based on genomic coordinates of peptides). If PQPQ based grouping (18) of the peptides in quantitative clusters has been performed, each peptide is filled by the same color of the cluster that it is assigned to by PQPQ. On the left, the boxes in different colors represent different clusters. After the box, follows the name assigned to the cluster, for example, “cluster1”, and then the number of unique peptides that belongs to this cluster written in brackets. The histogram in the bottom panel is the relative quantitative pattern of each cluster. Each group of bars represents one peptide and the number of bars is equal to the sample size. The height of one bar is the mean of the relative intensity ratio of all PSMs from one peptide in this cluster (in this data set iTRAQ 8-plex was used for relative quantification). The black vertical lines indicate the standard deviation of the intensity ratio of the PSMs connected to the peptide (number in the bracket after each peptide is number of PSMs). For those which do not have black lines, there is only one PSM for the peptide in that cluster. The picture generated by SpliceVista is with high resolution, more detailed and clear view can be achieved by zooming in.

Fig.3. **Theoretical analysis of human splice variant specific peptides.** Pie chart of number of theoretical splice variant specific peptides (SVSPs). 146 818 trypsin digested peptides and 74 798 LysC digested peptides are splice variant specific, corresponding to 16 935 and 13 593 splice variants respectively. 3072 splice variants in the database are from single isoform genes, those generate approximately half of the SVSPs. Combining all
SVSPs 17 562 splice variants in EVDB database can be identified in MS assuming the presence of all splice variants and 100% sequence coverage.

Fig.4. **Identified splice variant specific peptides in the A431 cell line dataset.** Venn diagram comparing the output of whole cell lysate analysis compared with the analysis of the combined subcellular fractions. From left to right are illustrated the number a) unique peptides, b) splice variant specific peptides and c) splice variants.

Fig.5. **Comparison of gene centric, protein centric and splice variant centric quantitative analysis.** Three examples (EIF4H, ITGB4 and SHROOM3) from A431 dataset where discrepancies were observed between gene centric, protein centric and splice variant centric analysis. Fold changes are reported for different time points after treatment with gefitinib. The numbers in the parentheses are the number of PSMs used for quantification in each analysis. In gene centric analysis, all PSMs mapping to that gene were used to calculate fold change. In protein centric analysis, PSM grouping is done by the search engine. In splice variant centric analysis, only PSMs from SVSPs are used for fold change calculation. For further details, see the results section.

Fig.6. **SpliceVista visualization of EIF4H gene.** Figure showing the SpliceVista output picture of EIF4H gene detected in A431 cell line whole cell samples. Gene EIF4H has six exons and four known splice variants (exon 6 is cut out to enable better resolution). Nine unique peptides were identified for EIF4H and grouped in 3 clusters. Cluster 1 (blue) which includes peptide 4, 5 and 6 has a unique pattern which shows clear down-regulation in the last two samples (replicates at 24h after drug treatment). Peptide 4 (DDFNSGFR) and peptide 5 (DDFNSGFRDDFLGGR) are uniquely mapped to splice variant NM_022170. Peptide 6 (DDFLGGR) is not unique to splice variant NM_022170, but it is very likely that this peptide was derived from NM_022170 based on its similar quantitative pattern to peptide 4 and 5. In the mid panel, the number in the bracket after each cluster is the number of unique peptides grouped in this cluster. In the bottom panel, the number in the bracket after each peptide is number of PSMs.

Fig.7. **SpliceVista visualization of a novel PLCB2 variant.** Figure showing a novel peptide (GSAAQNSSFMPVSLQRHQR) identified from a previously unknown splice variant H15C2281.2 of gene PLCB2 in ECgene database. In the figure, introns and exons (beige boxes) are scaled to its size, green and red lines indicates start codon and stop codon respectively. The blue line with number 1 below indicates peptide's genomic position and it is also marked as blue line on the splice variant which the peptide is derived from. Discovery of this novel peptide also complies with RNAseq data (not shown here). Four splice variants are shown with complete structure, parts of other splice variants are cut out.
Fig.8. **Comparison of protein centric and Splice centric quantitative analysis in the stem cell dataset.** A) Comparison of Fibro/hiPS protein ratio between protein centric analysis and splice variant centric analysis on proteins with splice variants identified. 296 proteins with splice variant identified in both MS experiment 1 and 2 (biological replicas) were included. Fibro/hiPS ratio in the figure was calculated as mean of the two replicas. The red dots indicate proteins which show differences of log2(Fibro/hiPS) value larger than 0.5 (1.41 fold) between protein centric and splice variant centric analysis. B) Zoom into the top 10 proteins ranked by differences of log2(Fibro/hiPS) between protein centric and splice variant centric quantification.

**Table 1.** Overview of unique peptide and splice variant identifications in each subcellular fraction and whole cell.

<table>
<thead>
<tr>
<th>Subcellular fraction</th>
<th>Heavy fraction</th>
<th>Medium fraction</th>
<th>Light fraction</th>
<th>Whole cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of identified genes</td>
<td>8015</td>
<td>8062</td>
<td>6562</td>
<td>7762</td>
</tr>
<tr>
<td>Total number of genes found in EVDB and analyzed by SpliceVista</td>
<td>7260</td>
<td>7244</td>
<td>5900</td>
<td>6945</td>
</tr>
<tr>
<td>Number of unique peptides identified</td>
<td>40689</td>
<td>40760</td>
<td>32435</td>
<td>37099</td>
</tr>
<tr>
<td>Number of unique peptides that can be mapped to EVDB and have quantitative data</td>
<td>36077</td>
<td>36417</td>
<td>28726</td>
<td>32969</td>
</tr>
<tr>
<td>Number of splice variant specific peptides (SVSPs)</td>
<td>1963</td>
<td>2027</td>
<td>1309</td>
<td>1680</td>
</tr>
<tr>
<td>Number of splice variants with SVSPs mapped by SpliceVista</td>
<td>672</td>
<td>681</td>
<td>512</td>
<td>607</td>
</tr>
<tr>
<td>Number of genes with splice variants identified</td>
<td>669</td>
<td>678</td>
<td>509</td>
<td>606</td>
</tr>
</tbody>
</table>
Fig. 1

Data preprocessing

- converter.py
- mergepsm.py

peptide data with gene symbol

Downloading

- fetch translated sequences from GenBank
- retrieve exon-composition of splice variant from EVDB

Mapping

- map peptides to splice variants
- mappingout.txt
- genestatistics.txt

Visualization

- visualize transcript position of peptides
- splice variant specific peptides

splice variant identification and quantification
Fig. 2
Fig. 5

A) EIF4H

Log2 FoldChange

0 0.2 0.4 0.6 0.8

gene centric (204)
protein centric (204) ENSP00000285753
splice centric (6) NM_022170

0h 2h 6h 24h

B) ITGB4

Log2 FoldChange

0 0.2 0.4 0.6 0.8

gene centric (227)
protein centric 1 (3) ENSP00000200181
protein centric 2 (1) ENSP00000344079
protein centric 3 (223) ENSP00000200181; ENSP00000344079
splice centric (2) NM_000213

0h 2h 6h 24h

C) SHROOM3

Log2 FoldChange

0 0.2 0.4 0.6 0.8

gene centric (39)
protein centric 1 (2) ENSP00000370111
protein centric 2 (37) ENSP00000298043
splice centric (2) AK127929

0h 2h 6h 24h
Fig. 6
Fig. 7

Fig. 8

A) Log2(Fibro/hiPS) vs ProteinCentric and SpliceCentric

B) Log2(Fibro/hiPS) zoom-in vs ProteinCentric and SpliceCentric