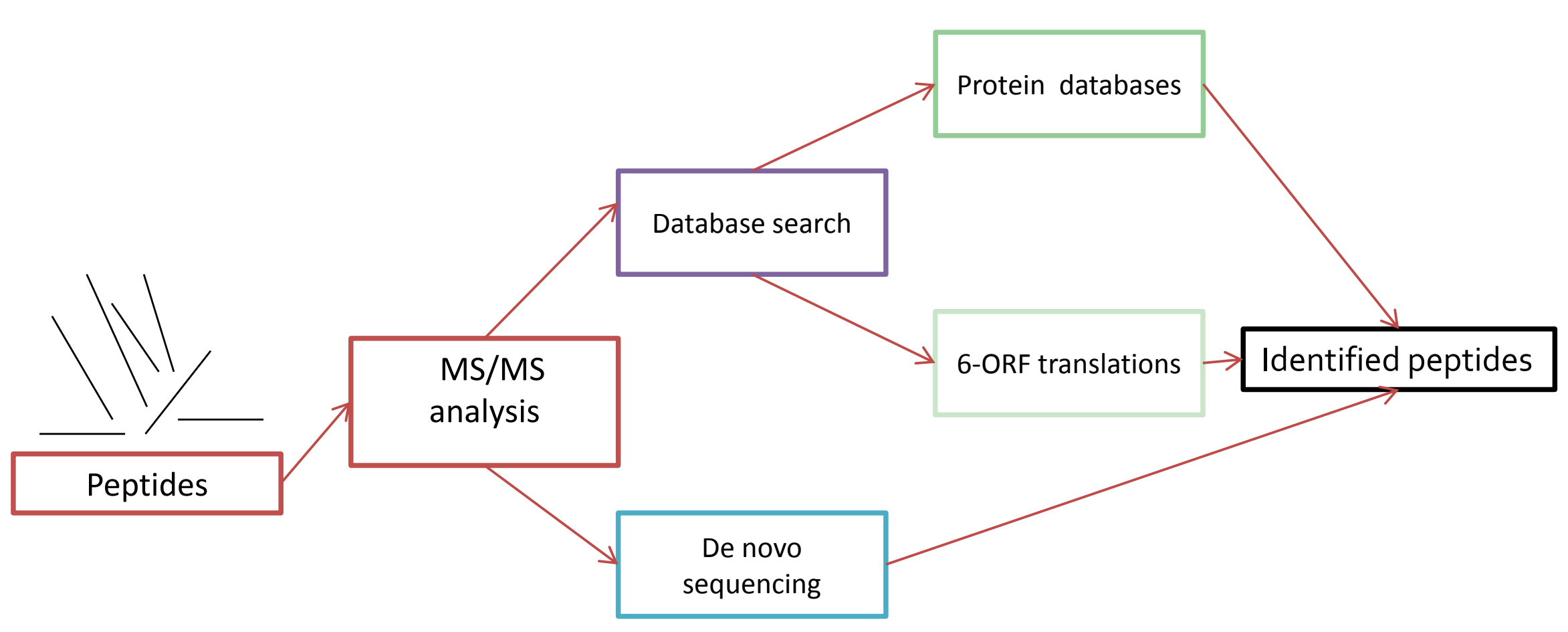
Using Proteomics Data for the Validation of Gene Models

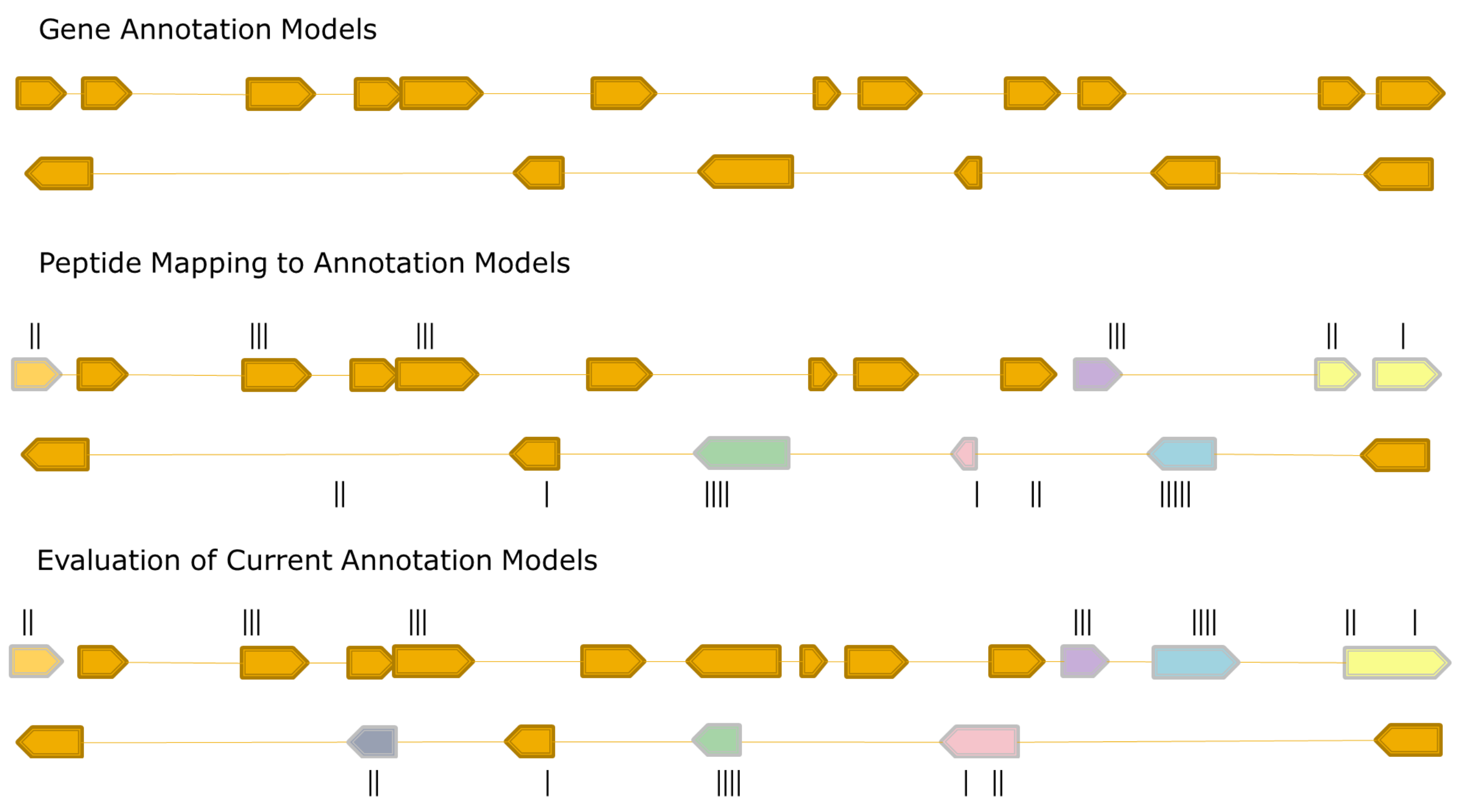
Seçkin BOZ*, Canan HAS*, Jens ALLMER*

*Izmir Institute of Technology, Molecular Biology and Genetics Department, Gulbahce – Urla- Izmir

INTRODUCTION

Gene annotation is often based on purely hypothetical gene models and even if transcripts have been sequenced this is no proof for the existance of a protein. A proteogenomcis approach can be used to map proteins to the genome hence adding experimental evidence to the existence of the annotated transcripts on the protein level. In this study, we used publicly available mass spectrometric data to analyze the gene models for human available at Ensembl. Initially, we directly mapped the identified peptides to the six frame translation of the genome. This map was then intersected with the gene models and we were able to confirm many of the proposed exons. As we expected, many of the peptides did not fully fit to the gene models and we are thus able to suggest changes. Some intergenic peptide clusters were discovered which suggest unknown, but expressed, proteins which need further investigation.




Gene Annotation Models
Peptide Mapping to Annotation Models
Evaluation of Current Annotation Models

Digested peptides are measured by mass spectrometry. After tandem mass spectrometry measurement, tandem mass spectra (MS/MS spectra) are analyzed by database search or de novo sequencing. Database search can be done against either protein database or the six-frame translation of a genome. Identified peptides are mapped to annotated genome. Therefore;

- 🟧 Validation of a current model
- 🟦 Discovery of a new gene model
- 🟩 Determination of exon downstream
- ▨ Determination of exon upstream
- 🟪 Alternative start site selection
- 🟦 Reversal of model
- 🟨 Merging of two exons

METHODOLGY

PeptideAtlas Accession Number	Source	Instrument	Reference
PAe000028	Human serum glycoproteins	LCQ DECA XP	Hui Zhang
PAe000044	Red blood cell lysate	LCQ DECA XP	Dan Martin
PAe000134	Human blood plasma sample	LCQ DECA XP	Omenn et al (2004)

Table 1: PeptideAtlas accession numbers of tandem mass spectra collections used in this study, data sources, instrument type and references are given.

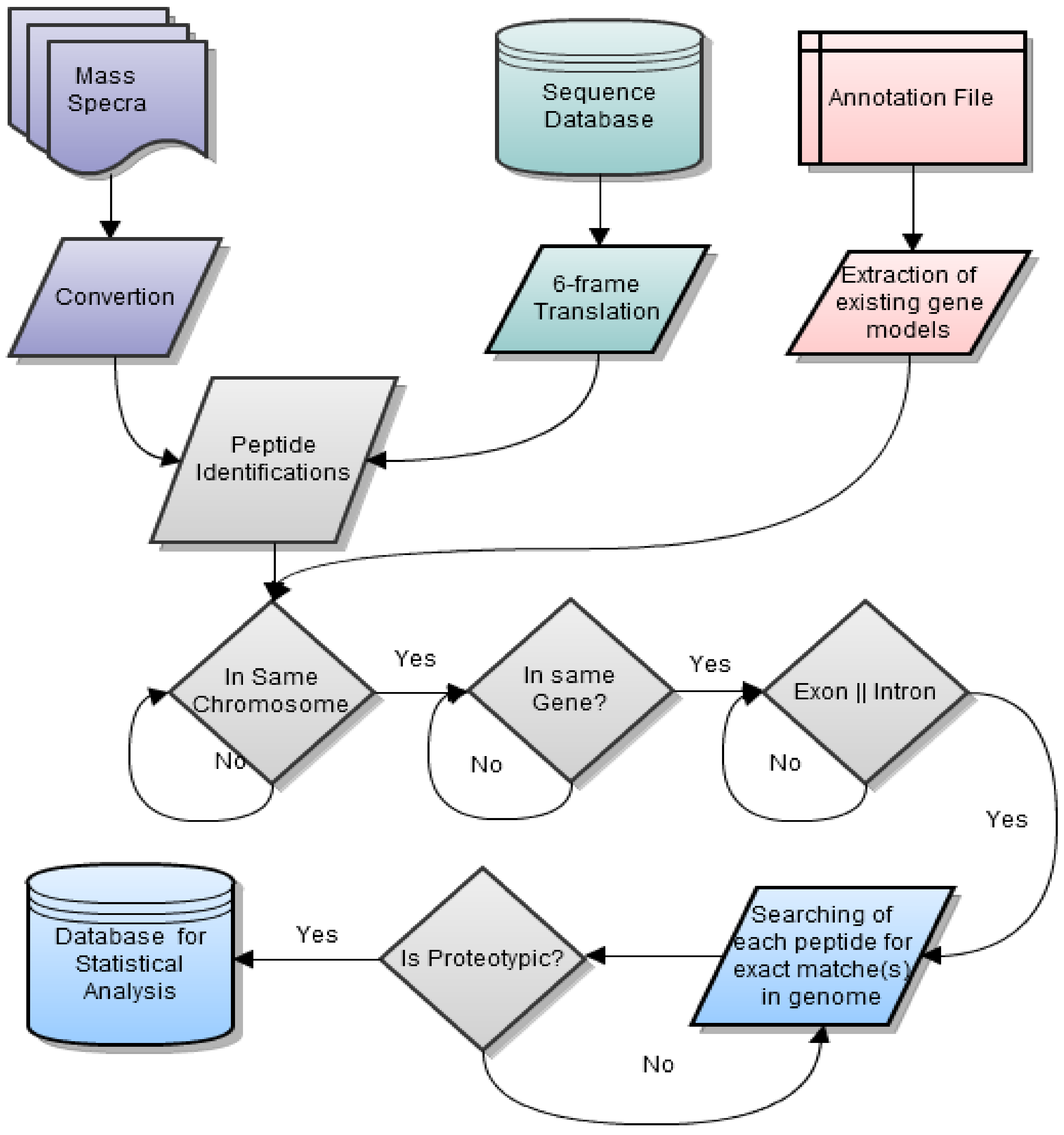


Figure 2: General workflow of human blood proteogenomic study. There are three different tracks that were studied. In the first track, human blood plasma tandem mass spectra collection was processed and peptides were identified by OMSSA. In the second track, currently available human genome annotations by ENSEMBL and HAVANA were parsed. The peptides identified in track 1 were further searched in human genome annotation in third track. Peptides found in exon or exon-intron boundary of same transcript of a gene for a chromosome were stored. This process was repeated for each chromosome. Stored peptides then were blasted aganist six-frame translation of the genome. When a peptide only occurs once, it is considered to be 'proteotypic'. According to BLAST results, proteotypic peptides are stored.

RESULTS

OMSSA Found Chromosome	BLAST Found Chromosome	Peptide Sequences
1	1	EDLVSSWEHIR, FLNVQELAAAHHEK, LTLSHPSDAPQIQEMK
3	3	DIASGLIGPLIIcK, HTFmGVVSLGSPSGEVSHPR, NNEGTYYSPNYNPQSR, TVVQPSVGAAAGPVVPPcPGR
6	6	TTNIQGINLLFSSR
7	7	YAPNFVmSIAYSIIK
9	9	CGFNsWLSFIHSFIFLANIGGPMLCKALCWASGIQR, KPCDPIPAWPVScTFSLR, NYSIRLPICsLQLVQTR
10	10	KLNPSIKPEFGQMsMSK, RDIALPFIAQMLPVLVsK
12	12	LLIYAVLPTGDVIGDSAK, VTAAPQSVcALR
14	14	IQLSHSAcEYTVFQTPYVK
15	15	LLPYSQPRGELISK
17	17	ASGGmVTATFGSLVKSGPWIWR
18	18	SGGQTDLPmSVQLSGHLSWFR
19	19	DWPRTVFVYILALAsVNPAQSQASASGGGEVGK, RPNSDAPLRGGGLEVGGWTQSTLGGSFPSSYLCVYPQIK, TmQALPYSTVGNSNNYLHLSVLR
22	22	ADAYQVTHTHtSPRR, AVGAAAVPHLPAFRPAGGAcWARcR, QHNVIILLLSFALFLLPPQTHTR
X	X	mLAAATSSDGYEGLRR

Figure 3: Group of proteotypic peptides which are seen in genome as unique are given with their corresponding chromosomes. Expectation value returned by BLAST search of each of these peptides in the complete six-frame translation of the genome is zero. Lower case letters show potentially post-translationally modified amino acids. For 14 chromosomes, proteotypic peptides have been identified so far.

# of Spectra	Peptide	Gene	Transcript	Genomic Part	Type	Proteotypic
1	EHWDHLLER	ENSG00000117228.8	ENST00000542693.1	chr1_ENSEMBL_exon_89521699	Exonic	x
1	DLQGVQNLLK	ENSG00000117228.8	ENST00000542693.1	chr1_ENSEMBL_exon_89525000	Exonic	X
2	FLNVQELAAAHHEK	ENSG00000117228.8	ENST00000542693.1	chr1_ENSEMBL_exon_89523675	Overlaps with 5' exon end	√
1	AQLIDER	ENSG00000117228.8	ENST00000370473.4	chr1_HAVANA_exon_89518002	Exonic	X
2	FLNVQELAAAHHEK	ENSG00000117228.8	ENST00000370473.4	chr1_HAVANA_exon_89523675	Overlaps with 5' exon end	√
18	ITDLEHFAESLIADEHYAK	ENSG00000117228.8	ENST00000370473.4	chr1_HAVANA_exon_89518002	Exonic	X
1	EHWDHLLER	ENSG00000117228.8	ENST00000370473.4	chr1_HAVANA_exon_89521699	Exonic	x
1	DLQGVQNLLK	ENSG00000117228.8	ENST00000370473.4	chr1_HAVANA_exon_89525000	Exonic	x
3	EDLVSSWEHIR	ENSG00000202385.1	ENST00000365515.1	chr1_ENSEMBL_exon_89485929	Overlaps with 3' exon end	√
2	LTLSHPSDAPQIQEMK	ENSG00000202385.1	ENST00000365515.1	chr1_ENSEMBL_exon_89485929	Exonic	√

Figure 4: An example is given for chromosome 1 where two different genes GBP1 ENSGO..7228.8 (gray color) and its two transcripts (light green and dark green) and YRNA-ENSGO..2385.1 (blue color) are shown. Moreover, peptides found in either exonic region or 3'/ 5'exon-intron boundary of transcripts of corresponding gene are given with their proteotypic feature. For YRNA, two proteotypic peptides are determined. For GBP1, for each transcript only one proteotypic peptide is found. Nonetheless, one proteotypic peptide uniquely identifies a protein and thus the existence of these proteins has been confirmed.

CONCLUSION and FUTURE STUDIES

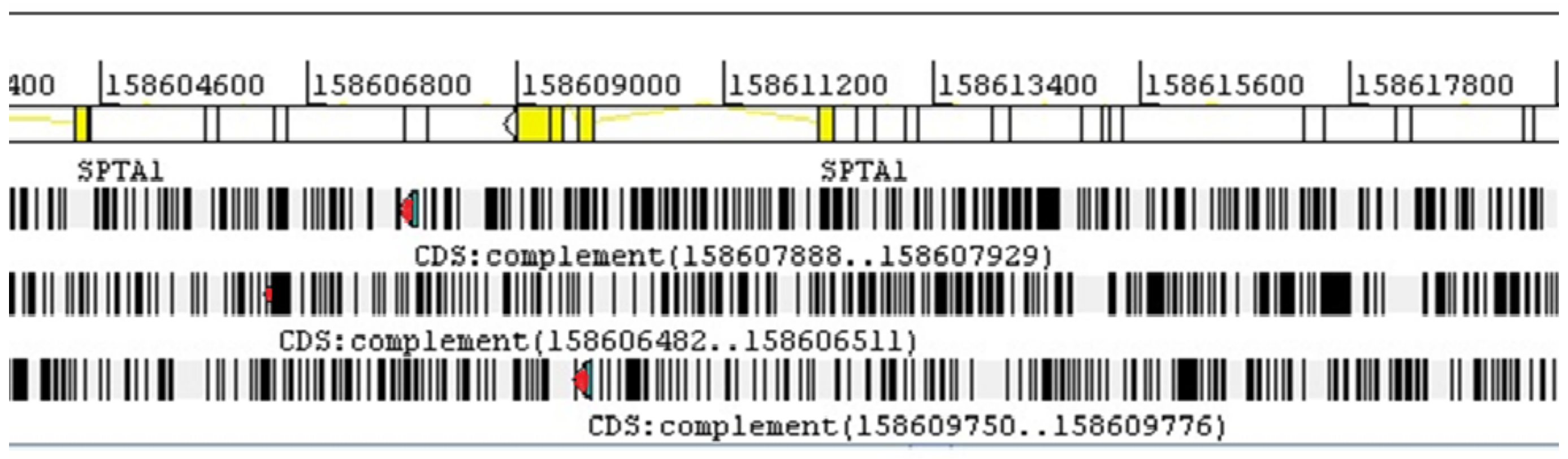

400
158604600
158606800
158609000
158611200
158613400
158615600
158617800
SPTA1
SPTA1
CDS:complement(158607888..158607929)
CDS:complement(158606482..158606511)
CDS:complement(158609750..158609776)

Figure 5: Mapped peptides (red) to SPTA1 gene (yellow) is visualized in Artemis Genome Browser. Proteotypic peptide is shown as red-black and non-proteotypic peptides are shown as red-cyan. According to this model, current exon upstream boundary should be extended.

This study can be considered as an important step on the integration of genomic and proteomic data in order to have a systems point of view. In this study, 1874 mzXML files of three tandem mass spectra collections were searched against six-frame translated human genome. 36456 tandem mass spectra were assigned peptides by OMSSA. After removing redundant peptides, 29555 peptides were collected in total. In order to validate existing gene annotations, peptides, which were found in the exonic regions and exon-intron boundaries of the same transcript, 28 proteotypic peptides were identified on 14 chromosomes. The importance of proteotypic peptide identification is related to the correction of existing gene annotations, discovery of new gene models in the intronic or intergenic regions, alternative splice isoforms and additional reading frames.

Further analysis on proteotypic peptides, and discovery of proteins in intergenic regions via peptide clusters will be performed in the future. Moreover, we aim to identify the use of alternative start site selection using this data.