A New Class of Protein Biomarker Candidates: Identification of Novel Alternative Splice Isoforms using Proteomic Informatics with a Modified ECGene Database

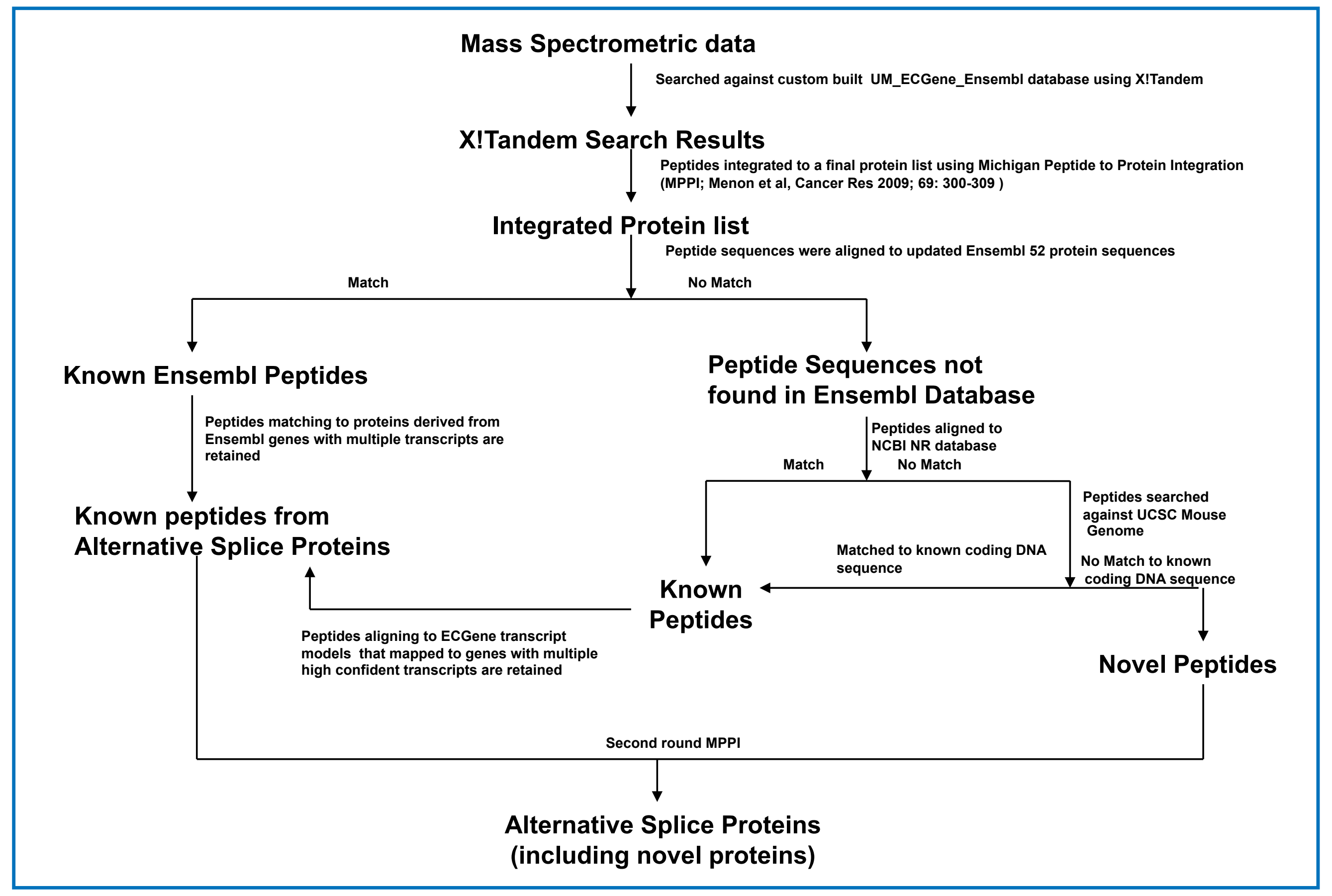
R. Menon, D.J. States, G.S. Omenn

Center for Computational Medicine and Biology and National Center for Integrative Biomedical Informatics, University of Michigan, Ann Arbor, MI, 48109-2218

Overview

Alternative splicing plays a major role in protein diversity without significantly increasing genome size. Aberrations in alternative splice variants are known to contribute to a number of diseases. The several alternative splice databases now publicly available differ in their annotation and modeling methods and contain many transcripts not present in reference resources like Ensembl or Refseq. The ECgene database is one of the largest alternative splice variant databases [Kim P, et al. Genome Research 2005]. Taking alternative splicing events into specific consideration, ECgene combines genome-based EST clustering and the transcript assembly procedure to construct gene models that encompass all alternative splicing events. The reliability of each isoform is assessed from the nature of cluster members and from the minimum number of clones required to reconstruct all exons in the transcript. In this study of potential biomarkers for breast cancer, we have used mass spectrometric data to interrogate a custom-built, non-redundant database created with three-frame translations of mRNA sequences from ECgene and Ensembl to find alternative splice variants. The mass spectrometric files from LC-MS/MS analyses of tumor and normal mammary tissue from a HER2/Neu-driven mouse model of breast cancer [Whiteaker et al, JPR 2007] were downloaded from PeptideAtlas [http://www.peptideatlas.org/repository/]. From our analysis, we identified a total of 584 alternative splice variants, of which peptides from 235 proteins were found only in tumor samples. Included in the 584 proteins, there were 35 proteins which were identified with peptides that did not match completely to any known mouse protein sequence. Novel peptides identified by multiple spectra from 19 proteins were found only from tumor samples.

Determination of Alternative Splice Isoforms



Mass Spectrometric data

Searched against custom built UM_ECGene_Ensembl database using X!Tandem

X!Tandem Search Results

Peptides integrated to a final protein list using Michigan Peptide to Protein Integration (MPPI; Menon et al, Cancer Res 2009; 69: 300-309)

Integrated Protein list

Peptide sequences were aligned to updated Ensembl 52 protein sequences

Match

No Match

Known Ensembl Peptides

Peptides matching to proteins derived from Ensembl genes with multiple transcripts are retained

Known peptides from Alternative Splice Proteins

Peptide Sequences not found in Ensembl Database

Peptides aligned to NCBI NR database

Match

No Match

Peptides searched against UCSC Mouse Genome

Matched to known coding DNA sequence

No Match to known coding DNA sequence

Known Peptides

Peptides aligning to ECGene transcript models that mapped to genes with multiple high confident transcripts are retained

Novel Peptides

Second round MPPI

Alternative Splice Proteins (including novel proteins)

Summary of Michigan Peptide to Protein Integration (MPPI) used to select peptides and build a integrated protein list

1) List of all peptide matches with X!Tandem expect score of <= 0.001 created.
2) Peptides with expect < 0.01 identified from 3 or more spectra added to list (FDR < 0.2%).
3) Peptide list ordered by number of spectra matching each peptide.
4) Peptide with largest number of matching spectra selected.
5) List of all proteins containing this peptide, ranked by decreasing number of total distinct peptides identified, decreasing number of total spectra, increasing expect value, and then increasing protein length.
6) The highest ranking protein was put on the final integrated protein list; if a tie, Ensembl protein chosen over ECGene.
7) All other peptides contained within this protein removed from the peptide list.
8) Steps 3-7 repeated until no peptides remained in the peptide list.

Note:

To achieve a false positive rate of <= 1%, an additional threshold was applied to final integrated protein list. The diagnostic peptide that was used for including the protein to integrated list has to be identified by 3 or more spectra. We identified a total of 1121 distinct proteins of which 9 were from reverse sequence.

Results

Alternative Splice Variant Proteins in Her2/Neu Mouse Model of Human Breast Cancer

Summary of Alternative Splice Variants Identified

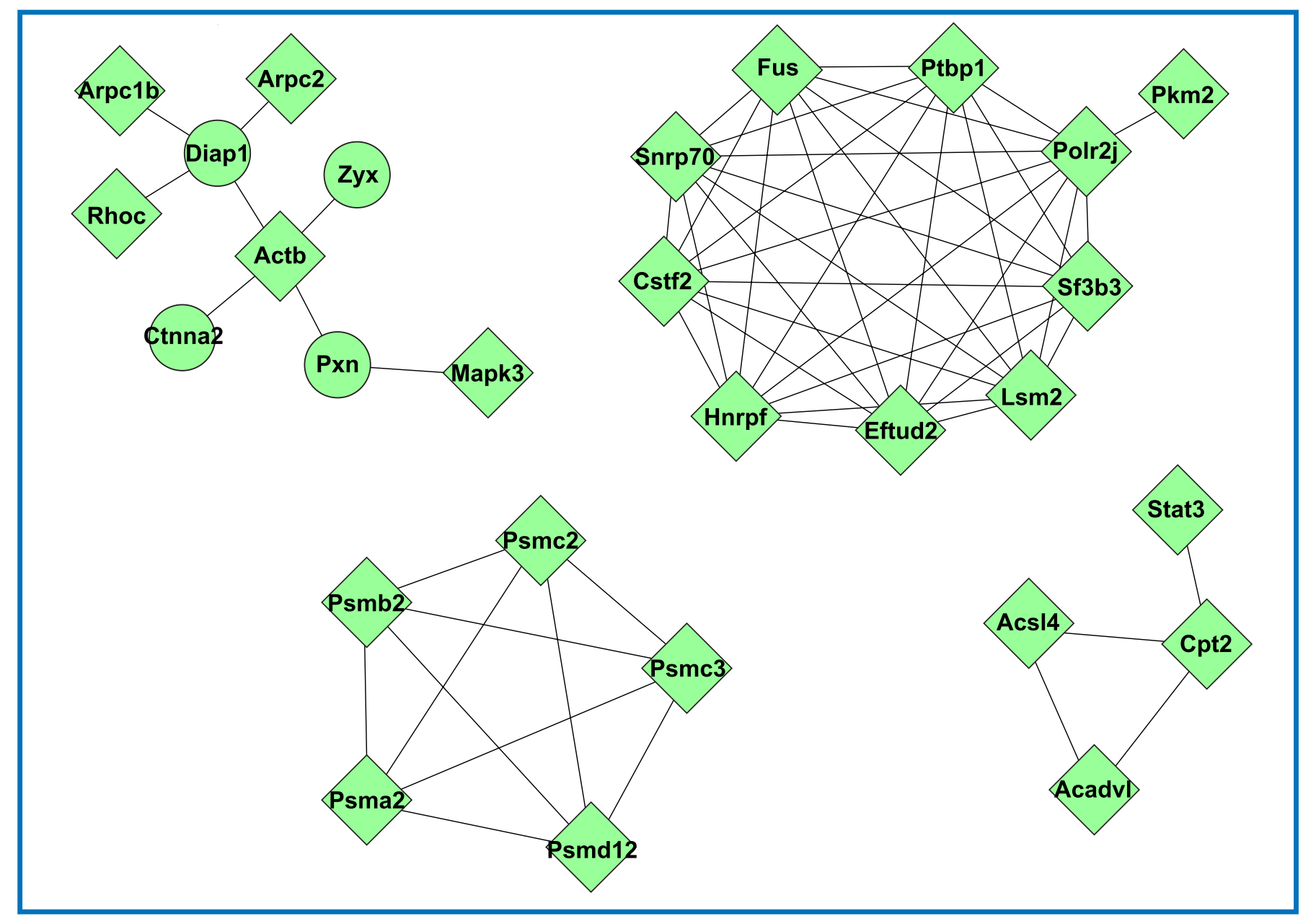
	No. alternative splice variants identified	No. identified only in normal or in tumor
Normal	349	77
Tumor	507	235
common	272	-
Total	584	-

Top 10 Alternative splice variant proteins found only in tumor samples: Ranked by the number of distinct spectra that identified the unique peptide

Protein	Unique Peptide	Gene Symbol	Description
ENSMUSP00000047410 ENSMUSG00000040158	VSEGGPAEIAGLQIGDK	Tax1bp3	Tax1 (human T-cell leukemia virus type I) binding protein 3 Gene
M6C4898_9_s2_e1310_1_rf1_c1_n0|	DELTDLDQSNVTEETPEGEEHPVADTENK	serbp1	SERPINE1 mRNA binding protein 1 isoform 3
ENSMUSP00000021062 ENSMUSG00000020719	QNFTEPTAIQAQGWPVALSGLDMVGVAQTGS-GK	Ddx5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 Gene
ENSMUSP00000029941 ENSMUSG00000028273	SSGTGASVGPPQPSDQDTLVQR	Pdlim5	PDZ and LIM domain 5 Gene
ENSMUSP00000016072 ENSMUSG00000027422	EAEETQNSLQAECDQYR	Rrbp1	ribosome binding protein 1
ENSMUSP00000032992 ENSMUSG00000030738	ILADLEDYLNELWEDK	Eif3c	eukaryotic translation initiation factor 3, subunit C Gene
ENSMUSP00000034524 ENSMUSG00000032026	ESTVTLQQAEYEFLSFVR	Rexo2	REX2, RNA exonuclease 2 homolog
ENSMUSP00000099807 ENSMUSG00000052997	YLFNQLFGEEDADQEVSPDRADPEAAWEPTE-AEAR	Uba2	ubiquitin-like modifier activating enzyme 2 Gene
ENSMUSP00000034215 ENSMUSG00000031765	SCCSCCPVGCSK	Mt1	metallothionein 1 Gene
ENSMUSP00000045073 ENSMUSG00000033732	ELAAEMAAAFLNENLPESIFGAPK	Sf3b3	splicing factor 3b, subunit 3 Gene
ENSMUSP00000044827 ENSMUSG00000040463	AGNALGGVDNEEEEELGDEAMMALDQN-LASLFK	Mybbp1a	MYB binding protein (P160) 1a Gene

Direct interactions observed between the Alternative Splice proteins that were found only in tumor samples using Cytoscape MiMI Plugin

(only the interactions involving three or more input proteins are shown)



Examples of different Alternative Splice Variants from the same gene found in normal and tumor samples

Protein	Sample Type	Gene Description	Domains found only in tumor variant
ENSMUSP00000007814 ENSMUSG00000007670|	tumor	KH-type splicing regulatory protein (khsrp)	pfam_fs:DUF1897 (domain of unknown function); Amino acids 609-686
ENSMUSP00000066416 ENSMUSG00000007670	normal		
ENSMUSP00000044827 ENSMUSG00000040463	tumor	MYB binding protein (P160) 1a (mybbpa1)	pfam_fs:DUF1795 (domain of unknown function); prf:ASP_RICH (Aspartic acid rich region profile)
ENSMUSP00000098459 ENSMUSG00000040463	normal		
ENSMUSP00000030056 ENSMUSG00000028364	tumor	tenascin C (tnc)	prf: multiple fibronectin type-III domain profile
ENSMUSP00000102994 ENSMUSG00000028364	normal		
ENSMUSP00000076801 ENSMUSG00000030795	tumor	fusion, derived from t(12;16) malignant liposarcoma (fus)	prf:Eukaryotic RNA Recognition Motif (RRM) profile
ENSMUSP00000101856 ENSMUSG00000030795	normal		
ENSMUSP00000108561 ENSMUSG00000000568	tumor	heterogeneous nuclear ribonu-cleoprotein D (hnrnpd)	prf:Tyrosine-rich region profile; pfam_fs:CBFNT (NUC161) domain
ENSMUSP00000072533 ENSMUSG00000000568	normal		

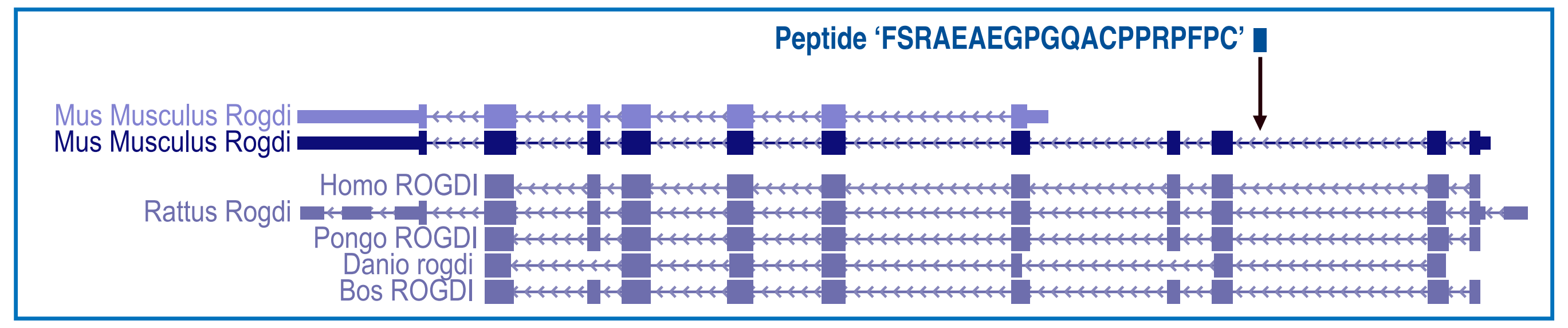
Proteins identified with novel peptides that did not match any known mouse protein sequence and found only from tumor samples

Protein	Novel Peptide	Gene Symbol	Description
ENSMUSG00000050867|ENSMUST00000052606|NULL_s2_e722_1_rf0_c1_n0|	RARLAEQASAMKAVTELNEP	ywhah	tyrosine 3-monooxygenase/tryptophan 5-monooxy-genase activation protein, eta polypeptide;EST BX523414
M4C12080_1_s2_e284_1_rf1_c1_n0|	VTEDENDEPIEIPSEDDGTVLLSTVTAQFPGGSMQR	tardbp	Tar DNA binding protein
M14C5830_1_s431_e746_1_rf0_c1_n0|	GSGLVPTLGRGAETPVSGAGATRGLSR	sox7	transcription factor sox7
ENSMUSG00000022982|ENSMUST00000023707|ENSMUSP00000023707_s2_e479_1_rf0_c1_n0|	QKARPGARGAGRVVLSGQITGLTEG	sod1	superoxide dismutase 1, soluble;
M7C8497_9_s2_e590_1_rf0_c1_n0|	RGQKPPAMPQPVPTA	rps3	ribosomal protein S3
M16C284_1_s56_e302_1_rf0_c1_n0|	FSRAEAEGPGQACPPRPFPC	rogdi	leucine zipper domain protein
M10C5505_7_s2_e1031_1_rf1_c1_n0|	QAHGAGTGDSGAERRAAGEELGLLVS	pfk1	phosphofructokinase, liver, B-type
M8C10692_1_s416_e755_1_rf2_c1_n0|	ANSRTATATQRNYVSTASLFPHPSVGAGEMAQLLR	pard3	par-3 (partitioning defective 3) homolog (C. elegans) (Pard3), transcript variant 3
M12C6304_1_s2798_e2864_1_rf2_c1_n0|	IIYFISVLLPLLKTAFVEKK	nrxn3	novel neurexin III
M10C1728_7_s2951_e3125_1_rf1_c1_n0|	AKLTFVNLPFLDVGGGWGK	l3mbtl3	l(3)mbt-like 3
M7C9387_6_s2_e812_1_rf2_c1_n0|	LFQEEFPGIPYPPDRLEKELG	hpx	hemopexin
M4C916_4_s3683_e3917_1_rf1_c1_n0|	LAAAAAAAAAAAK	fam82b	family with sequence similarity 82, member B
M10C6186_20_s161_e479_1_rf1_c1_n0|	CPPSRTILMMGRYVEPIEDVPCGNIVGLVGVDQFLVK	eef2	eukaryotic translation elongation factor 2
M8C7190_6_s53_e1784_1_rf2_c1_n0|	ELLEITVKLQFGGVKGLFDNTSMSTVIDGVVLP	ces1	carboxylesterase 1
M11C7819_6_s383_e638_1_rf2_c1_n0|	TVIMPHSYPALSAEQKKELSD	aldoc	aldolase 3 c
M5C6439_110_s2_e902_1_rf2_c1_n0|	PNLRENYGELADCYLPAIAADFVEDQEVCK	alb	albumin
M5C6439_135_s278_e851_1_rf0_c1_n0|	SLPPTVTNPFTLFLEISCPAIAADFVEDQEVCK	alb	albumin
ENSMUSG00000057800|ENSMUST00000073485|NULL_s2_e1115_1_rf0_c1_n0|	SFAGDDAPR	actb	beta actin
M7C5448_1_s596_e680_1_rf2_c1_n0|	IYYSFGALKLGCFNFPLLKFL		Mus musculus chromosome 7, clone RP23-49M22

Novel peptide 'FSRAEAEGPGQACPPRPFPC'

- Three distinct spectra
- Only in tumor; two BAC clones; RP24-424L20 and RP23-450O11
- Aligns to intronic region of Rogdi gene (leucine zipper domain)
- Found one predicted donor splice site (with splice prediction score = 0.93) in the intronic sequence using the Splice Site Prediction by Neural Network
- Identified a phosphopeptide motif in the intronic region which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 using ELM motif search.

Genomic structure of the Rogdi gene as shown on the UCSC Genome Browser. The novel peptide identified aligns to the intronic region of the gene.



Novel peptide 'RARLAEQASAMKAVTELNEP'

ref|NP_035868.1| tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide [Mus musculus] (YWHAH gene)

NCBI blastp Score = 58.3 bits (130), Expect = 7e-08, Identities = 20/27 (74%), Positives = 20/27 (74%), Gaps = 7/27

```
Query  1   RARLAEQA---------------SAMKAVTELNEP  20
           RARLAEQA               SAMKAVTELNEP
Sbjct  10  RARLAEQAERYDDMASAMKAVTELNEP  36
```

- Only in tumor; BAC clone RP23-112E4
- A functional motif for tyrosine-based sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex found in the missing amino acid sequence 'ERYDDMA'. This Y based motif determines which vesicular traffic pathway is used to transport a particular molecule and hence determines its final destination.
- Gremlin 1 plays an oncogenic role especially in carcinomas of the uterine cervix, lung, ovary, kidney, breast, colon, pancreas, and sarcoma. Over-expressed gremlin 1 functions by interaction with YWHAH. (Hong, et al. BMC Cancer 2006;6:74)

Conclusion

- The combined proteomic and bioinformatic approach in this study has identified 35 novel splice variants, including 19 found only in tumor samples.
- More analysis is being done to validate these novel peptides using RT-PCR and QT-PCR.
- These data suggest that alternative splice variants play functional roles in tumor mechanisms and are potentially rich sources of candidate biomarkers.

Acknowledgements

Supported by NCI/SAIC 23XS110A on Mouse Models of Human Cancers, MTTC GR 687 for Proteomics Alliance for Cancer Research, U54 DA021519 National Center for Integrative Biomedical Informatics, and P41 RR018627 National Resource for Pathways and Proteomics.