Disease Markers 28 (2010) 241–251
DOI 10.3233/DMA-2010-0702
IOS Press

Alternative splice variants, a new class of protein cancer biomarker candidates: Findings in pancreatic cancer and breast cancer with systems biology implications

Gilbert S. Omenn[a,*], Anastasia K. Yocum[a,c] and Rajasree Menon[a]
[a]*University of Michigan Center for Computational Medicine and Bioinformatics and Michigan Proteomics Alliance for Cancer Research, Ann Arbor, MI, USA*
[b]*Departments of Internal Medicine, Human Genetics and School of Public Health, University of Michigan, Ann Arbor, MI, USA*
[c]*Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI, USA*

Abstract. Alternative splicing plays an important role in protein diversity without increasing genome size. Earlier thought to be uncommon, splicing appears to affect the majority of genes. Alternative splice variants have been detected at the mRNA level in many diseases. We have designed and demonstrated a discovery pipeline for alternative splice variant (ASV) proteins from tandem MS/MS datasets. We created a modified ECgene database with entries from exhaustive three-frame translation of Ensembl transcripts and gene models from ECgene, with periodic updates. The human database has 14 million entries; the mouse database, 10 million entries. We match MS/MS findings against these potential translation products to identify and quantify known and novel ASVs. In this review, we summarize findings and systems biology implications of biomarker candidates from a mouse model of human pancreatic ductal adenocarcinoma [28] and a mouse model of human Her2/neu-induced breast cancer [27]. The same approach is being applied to human tumors, plasma, and cell line studies of other cancers.

Keywords: Alternative splicing, splice variants, protein isoforms, breast cancer, pancreatic cancer, mouse models, proteomics, protein interaction networks, systems biology

1. Introduction

Identifying, confirming, validating, and commercializing biomarkers for heterogeneous common diseases, especially cancers, has been a very challenging task for the biomedical research community. Surprisingly few individual marker candidates or panels of marker candidates have been confirmed by the many research groups dedicated to this effort; even fewer have reached clinical usefulness over the past decade. New kinds of markers may help not only by expanding the array of candidates but also by linking statistical associations with systems biology of the disease process, as this special issue of Disease Markers seeks to stimulate. Our current work on alternative splice variants is a promising example.

Alternative splicing increases protein diversity without significantly increasing genome size. It is now recognized to be very common throughout the human genome. Cancer-specific splicing events have been reported at the mRNA level in colon, bladder, and

*Corresponding author: Gilbert S. Omenn, M.D., Ph.D., Professor of Internal Medicine, Human Genetics, Bioinformatics and Public Health; Director, Center for Computational Medicine and Bioinformatics, University of Michigan, 2017F Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA. Tel.: +1 734 763 7583; Fax: +1 734 615 6553; E-mail: gomenn@umich.edu.

ISSN 0278-0240/10/$27.50

prostate tissues, with diagnostic and prognostic implications [34]. There are examples of every kind of splicing in cancers, including alternative individual splice sites, alternative exons, and alternative introns [28]. Splice events that affect the protein coding region of the mRNA give rise to proteins differing in sequence and activities; splicing within the noncoding regions can result in changes in regulatory elements, such as translation enhancers or RNA stability domains, which may dramatically influence protein expression [6]. Several databases with alternatively spliced transcripts are available [23]. We have utilized the ECgene database [21], which is based on evidence collected from clustering of ESTs, mRNA sequences, and gene model predictions.

We summarize here our approach to identification and quantification of alternative splice variants (ASVs) in cancers and our results for genetically-defined mouse models of human pancreatic and breast cancers. The National Cancer Institute has made such mouse models of human cancers a major thrust (http://mouse.ncifcrf.gov) in the search for biological understanding of mechanisms of cancer initiation and cancer progression and the companion search for biomarkers for diagnosis, prognosis, and response to therapy [25].

2. The modified ECgene database of potential translation products

We generated our alternative splice variant database from the modified ECgene database to include three-frame translations of cDNA sequences [28]. ECgene combines genome-based EST clustering and a transcript assembly procedure to construct gene models that encompass all alternative splicing events. The reliability of each isoform was assessed from the nature of cluster members and from the minimum number of clones required to reconstruct all exons in the transcript [20]. We combined Ensembl (version 40) with ECgene database (mm8, build 1); the transcript sequences were translated in three reading frames; within each dataset, the first instance of each protein sequence of 14 amino acids or longer was recorded. The resulting proteins from both database translations were combined and filtered for redundancy, with preference given to Ensembl. We added a collection of common protein contaminant sequences, and then generated and added a set of reversed sequences as an internal control for false identifications. The total for the mouse was 10.4 million protein sequence entries. A comparable process generated a human modified ECgene database with 14.2 million entries, which we are using for other studies not included in this report. The mzXML files containing the mass spectral information are searched against the modified ECgene database using X!Tandem software [10]. Peptides are integrated to a list of proteins using TransProteomic Pipeline and/or the Michigan Peptide to Protein Integration workflow, and further analyzed as described previously [26–28]. Peptides that were identified by X!Tandem search with X!Tandem expect value < 0.001 or with three or more spectra with expect value < 0.01 had a false discovery rate (FDR) $< 1\%$, based on peptides identified from reverse sequences compared to total peptides identified after applying the threshold. To characterize alternatively spliced peptides and proteins, we used InterProScan and Motif Scan [31], Gene Ontology, and FuncAssociate, and displayed protein-protein interactions with the Cytoscape plug-in for MiMI (Michigan Molecular Interactions) [14].

3. Identification of splice variant peptides in plasma of mice with pancreatic cancer

Pancreatic ductal adenocarcinoma (PDAC) is the most lethal of human cancers, due to absence of methods for early diagnosis and chemoresistance of advanced disease. Five-year survival is < 5 percent of patients, with 31,000 deaths per year in the United States [19]. The $KRAS^{G12D}$ activation and p16/Ink4a and p19/Arf-p53 deletions mouse model of PDAC was genetically engineered by DePinho and Bardeesy to match the molecular lesions of human PDAC; it recapitulates the histopathologic progression and clinical effects of the human disease in a highly reproducible and synchronous fashion. The tumors express pancreatic ductal markers (CK-19) and apical mucins (Muc1, Muc5AC), show activation of Hedgehog, Notch, and EGFR developmental signaling pathways, harbor genomic alterations syntenic to human PDAC, and exhibit proliferative stroma [1,3,13]. We exploited this model to test our hypothesis that cancer-specific ASVs could be identified in mass spectrometric analyses of plasma proteins from mice carrying these molecular lesions, compared with wild-type mice [28].

Plasma samples were processed by the Intact Protein Fractionation and Analysis System [36] after immunodepletion of the three most abundant proteins – albumin, immunoglobulins, and transferrin, which account

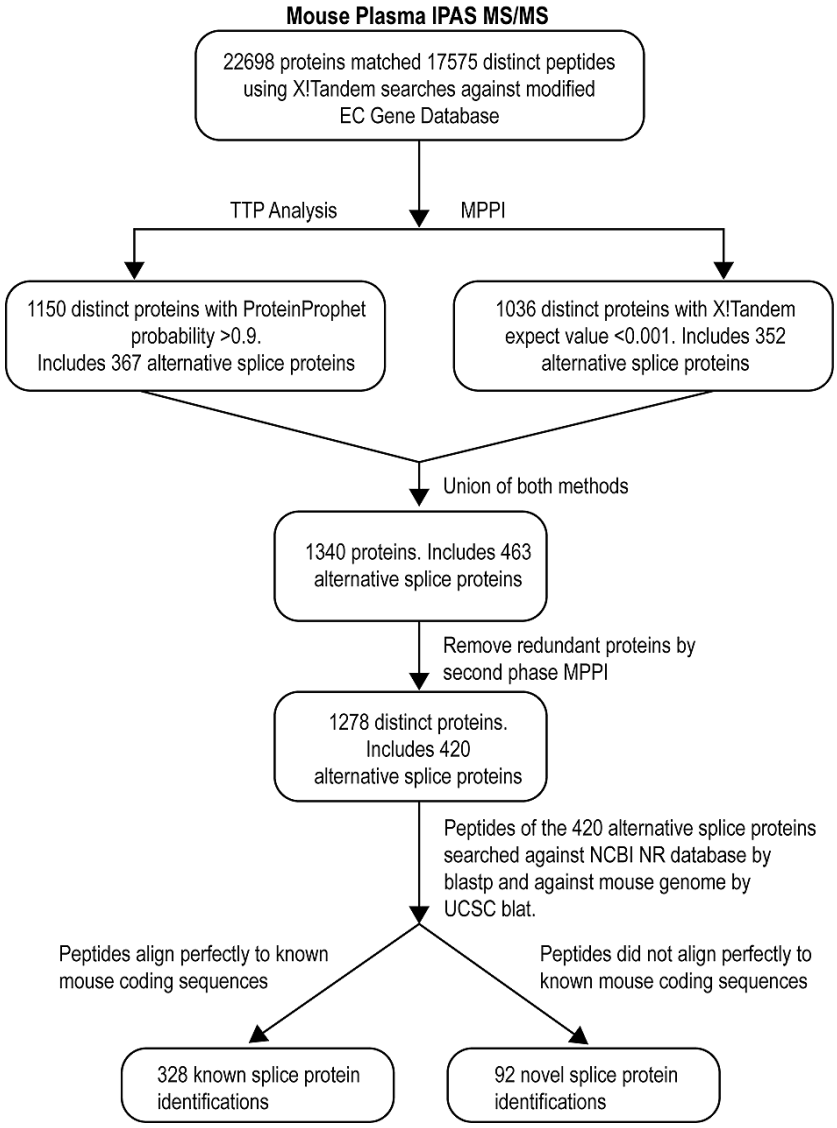

Fig. 1. Workflow of multi-step analysis of X!Tandem search results from Intact Protein Analysis System MS/MS, combining TransProteomicPipeline and Michigan Peptide to Protein Integration algorithms, leading to 420 alternatively spliced proteins, 328 previously identified and 92 novel. [From Cancer Research, Menon et al., 69: (1), 2009, p. 301].

for 90 percent of protein mass. After acrylamide labeling, the combined tumor and wild-type mouse plasma samples were fractionated into 12 anion exchange and then 13 or more reverse phase fractions, yielding a total of 163 fractions, which were digested with trypsin and analyzed with a ThermoFinnigan LTQ-FT mass spectrometer.

As outlined in Fig. 1, our integrated analysis revealed 420 distinct splice isoforms, of which 92 were novel, not matching any previously annotated mouse protein sequence. For seven of those novel variants, we prepared primers and validated the predicted sequences in the mRNA with qRT-PCR for all seven. Isotopic labeling of cysteine-containing peptides with D3 vs D0 acrylamide for the tumor-bearing mice and wild-type controls, respectively, permitted relative quantitation of 28 of the 92 novel proteins (those whose ASV peptides contained cysteine). Differential expression was demonstrated for peptides from novel variants of muscle-type pyruvate kinase, malate dehydrogenase 1, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), proteoglycan 4, minichromosome maintenance complex component 9, high mobility group box 2, and hepatocyte growth factor activator. Upon annota-

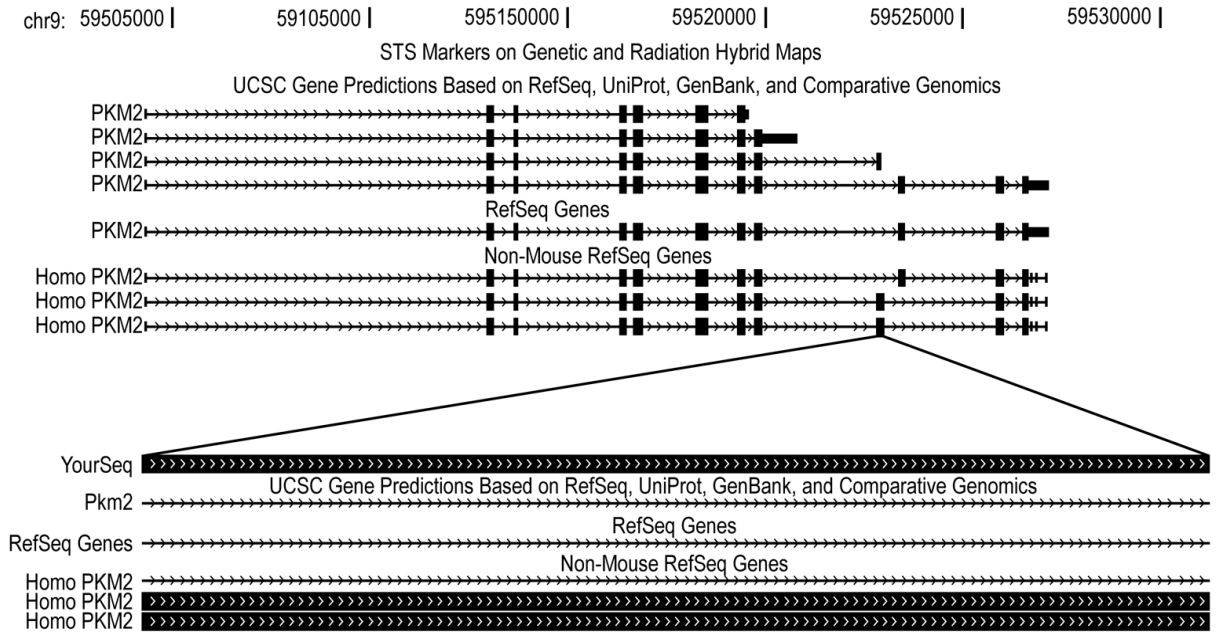

Fig. 2. View from the UCSC Genome Browser of the genomic structure of mouse muscle pyruvate kinase gene and the homologous human gene, with expansion of the 42bp region from 59,522,962 to 59,523,003 on chromosome 9 to map a 14 amino acid peptide CLAAALIVLTESGR that was the clue to identification of the alternatively spliced *pkm2* variant, identified with a total of 25 distinct peptides [From Cancer Research, Menon et al., 69: (1), 2009, p. 304].

tion, we presented [28] literature evidence that many of these ASVs may well be involved in pancreatic cancer, including alpha-fetoprotein, apolipoprotein E, ceruloplasmin, fibronectin, glyceraldehyde-3-phosphate dehydrogenase, hemopexin, peptidyl-prolylisomerase, and tubulin alpha among the novel ASVs, and acyl coA acetyl-transferase, chromograinin b, granulin, insulin-like growth factor binding protein 2, and regenerating islet-derived 3alpha among the known ASVs that also had significant differential expression (up-regulation).

From a systems biology point of view, one of the most interesting proteins is pyruvate kinase, which is the critical enzyme in the metabolic switch to aerobic glycolysis in cancers known since 1929 as "the Warburg effect" [8,37]. Aerobic glycolysis refers to persistence of high lactate production in the presence of oxygen. PK catalyzes transfer of a phosphoryl group from phosphoenolpyruvate to ADP, generating ATP. Most adult tissues express the M1 isoform, whereas tumors (and embryonic tissues), rapidly growing tissues, express the M2 isoform. Our analyses showed positively correlated up-regulation of peptide and mRNA expression of the novel variant of *pkm2* in plasma from tumor-bearing mice, compared with the wild-type.

Figure 2 shows the genomic structure of the mouse muscle-type pyruvate kinase, *pkm2*, with a 42 bp region highlighted that corresponds to the peptide CLAAALIVTESGR (aa 482 to 495) using UCSC Blast. This peptide was identified from five different spectra. It had never before been reported in the mouse, but it aligns perfectly with the homologous portion of the human, rat, and chicken muscle PKM1 isoform. This alternatively spliced protein variant of *pkm2* had 25 distinct peptides, covering $\sim$52% of the sequence, 22 of which were shared with the known mouse *pkm2* gene (homologous with M2 of human and rat *pkm2*). We also detected the known PKM2, as completely aligned with the known mouse coding sequence.

Kumar et al. [22] and Ventrucci et al. [35] both reported tumor type PKM2 as a metabolic marker specifically for pancreatic cancers. We are now exploring computerized structural prediction algorithms to characterize the effects of phosphorylation and of alternative splicing. With regard to glycolysis, we also found multiple novel splice variants of GAPDH. *Gapdh* mRNA levels are over-expressed in many cancers; Chen et al. [7] reported $>$2-fold increase of GADPH in pancreatic cancer tissue using isotope-coded affinity tag (ICAT) labeling, followed by mass spectrometry.

Warburg [37] noted that cancer cells take up glucose at higher rates than normal tissue, but use a smaller fraction of the glucose for oxidative phosphorylation,

even when oxygen is not limited. Such aerobic glycolysis is due to reprogramming of metabolic genes to permit a greater fraction of glucose metabolites to be incorporated into macromolecule synthesis rather than burned to carbon dioxide. Hitosugi et al. [18] showed that oncogenic forms of fibroblast growth factor receptor type 1 inhibit the pyruvate kinase M2 isoform by direct phosphorylation of PKM2 tyrosine residue 105 (Y105). This site-specific phosphorylation inhibits the formation of active, tetrameric PKM2 from less active dimers by disrupting binding of the PKM2 cofactor fructose-1,6-bisphosphate. They found that phosphorylation of PKM2 Y105 is common in human cancers; a PKM2 mutant with phenylalanine substituted for this tyrosine (Y105F) in cancer cells leads to decreased cell proliferation under hypoxic conditions, increased oxidative phosphorylation with reduced lactate production, and reduced tumor growth in xenografts in nude mice.

Growth factor signaling pathways activate protein tyrosine kinases and decrease the specific activity of pyruvate kinase (measured without regard to isoforms). Christofk et al. [9] used a proteomic screen with an immobilized phosphotyrosine (pTyr)-peptide library affinity matrix and SILAC labeling to identify novel pTyr-binding proteins from HeLa cell lysates. Binding of pTyr-peptides to PKM2 releases the allosteric activator fructose-1,6-bisphosphate, leading to inhibition of PKM2 enzyme activity. pTyr signaling stimulated by several growth factors diverts glucose metabolites from energy production to anabolic processes to support rapid growth of cancer cells, including nucleic acid and fatty acid biosynthesis. The M2 isoform is the only pyruvate kinase isoform that binds pTyr peptides (M1 liver, and red blood cell isoforms do not). PKM1 and PKM2 are identical in sequence except for a 56-amino acid stretch encoded by an alternatively spliced region involving exon 10 in PKM2, which forms an allosteric pocket unique to PKM2 in which the FBP activator can bind. Mutation of lys-433 at the lip of the pocket to glutamate interferes with the binding. Boxer et al. have now generated substituted N,N'-diarylsulfonamide molecules that activate the PKM2 and alter the Warburg effect, a new antiproliferation therapeutic strategy [5]. The glycolytic enzymes enolase and lactate dehydrogenase also are phosphorylated on tyrosine residues [9].

Another interesting splice variant that showed $>$2-fold increase in expression in tumor samples is high mobility group box 2, which is involved in DNA repair. While it has not previously been reported as up-regulated in pancreatic cancer, it is located at 4q32-34, a region identified as a potential locus for familial pancreatic cancer in a large Caucasian family [12]. The serine protease hepatocyte growth factor activator, which converts hepatocyte growth factor to its active heterodimer in response to tissue injury, was found to have an ASV; extracellular receptor kinase activation by hepatocyte growth factor has been reported to activate the MAP kinase pathway in human pancreatic cancer via MEK/ERK and p38 MAP kinase interaction [24].

We also searched our peptide findings for variants of proteins chosen as possible pancreatic cancer biomarkers, from among 1442 proteins identified, in a parallel study of this same mouse model that was the source of the data for this analysis [13]. We found variants of three of their nine proteins assayed by ELISA in humans (see next paragraph) in our list of 420 splice variant proteins: lipocalin 2 (LCN2), regenerating islet-derived 3 (REG3A), and tumor necrosis factor receptor superfamily member 1A (TNFRSF1A); according to our quantitative expression analysis, the TNFRSF1A showed $>$2-fold increase in expression in plasma from the tumor-bearing mice compared with wild-type mouse plasma.

Faca et al. [13] had first chosen 45 proteins, from a set of 165 up-regulated in plasma, that had $>=$ 1.5-fold increased expression, had corresponding ortholog gene in humans, were not known to represent acute-phase reactants, complement, or coagulation proteins according to Ingenuity Pathway Analysis, and had increased expression of their transcripts in pancreatic cancers. A much smaller number had antibodies available for ELISA and/or immunohistochemistry. Faca et al. were able to obtain immunohistochemistry staining for six in mice, but only three – receptor type tyrosine-protein (PTPRG), TNFRSF1 (above), and tenacin C (TNC) – in humans. They generated ELISA results for nine – ALCAM,TIMP1, ICAM1, LCN2, REG1A, REG3, IGFBP4, TNFRSF1A, and WFDC2 (HE4) – plus the clinical gold standard, CA 19-9. Eight of the nine new biomarker candidates tested by ELISA were statistically higher in cancer patient sera than at least one of the two controls in their study of serum from 30 newly-diagnosed patients with PDAC, 15 patients with chronic pancreatitis, and 20 healthy individuals at the University of Michigan; 5/7 were significantly increased against both controls. None of the candidates individually or collectively came close to matching CA 19-9's AUC of 0.98 for cancer versus normal; however, the full panel outperformed CA19-9 with AUC 0.96 vs 0.79 for cancer versus pancreatitis, and ICAM1 and TIMP1

individually outperformed CA 19-9 for PDAC versus with pancreatitis.

They then tested five proteins selected on the basis of their increased expression at early PanIN stage (5.5 weeks versus 7.0 weeks for the advanced stage) in the mouse model on serum from 26 participants in the lung cancer chemoprevention trial CARET (the Beta-Carotene and Retinol Efficacy Trial [29] cohort. The panel (LCN2, REG1A, REG3, TIMP1, IGFBP4) discriminated between serum specimens from 13 individuals who 7 to 13 months later would be diagnosed with pancreatic cancer and 13 individuals who were matched controls and did not develop cancer in a subsequent four-year follow-up period. Individually, only IGFBP4 and TIMP1 showed significance at 0.05 and 0.04, respectively, compared with CA 19-9 at 0.04; however, as a panel, the five proteins achieved an AUC of 0.82, compared with 0.74 for CA 19-9. The combination of the five plus CA 19-9 gave an AUC of 0.91, which is a promising result.

4. Compendium of potential biomarkers for pancreatic cancer

We are submitting our findings to the new pancreatic cancer data repository created by Harsha et al. [16], a compendium for biomarker candidates from published microarray and proteomic datasets from both exocrine and endocrine neoplasms of the pancreas obtained from GEO, ArrayExpress, and Oncomine [4,30,32]. They annotated the lists with evidence of these molecules in pancreatic juice, plasma, or serum and on plasma membranes of cells. They also compared results for pancreatitis. The manual curation of the literature consumed 7000 person-hours. Proteins are included if quantitative methods of ICAT, iTRAQ, or SILAC were used on tissues or cell lines. A total of 1868 genes was reported as over-expressed only in mRNA analyses, with 441 over-expressed in both mRNA and protein studies; 207 molecules were over-expressed only at the protein level, amang 648 proteins altogether. The compendium lists 166 membrane molecules over-expressed in pancreatic cancers in both mRNA and protein levels. There are 372 molecules over-expressed in chronic pancreatitis, with CCL3 and CCL4 cited as potential markers for pancreatitis not (yet) reported in PDAC. They also note two proteins, CECAM1 and MUC1, as outperforming CA 19-9 [2,33]. A major use of this compendium is the choice of 60 targets in PDACs for which Lustgarden Consortium investigators are developing antibodies for a range of studies [16].

5. Identification of splice variant peptides in tumor tissue of mice with HER2/neu breast cancer

In this study, we analyzed LC-MS/MS datasets from tumor and normal mammary tissue from a mouse model of HER2/neu-driven breast cancer [27]. Whiteaker et al. [38] identified 6758 peptides, representing $>$700 proteins in this well-established Chodosh model; their mzXML dataset was downloaded from their submission to PeptideAtlas [11]. The lysates from individual tissue specimens were pooled from 5 tumor-bearing mice and from 5 normal mice and analyzed with LC-MS/MS. We modified the workflow shown in Fig. 1; in the absence of acrylamide labeling, MPPI was sufficient without TPP (and its Q3Ratio and XPRESS features). Peptides that were identified by X!Tandem search with false discovery rate $<$ 1% (based on peptides identified from reverse sequences) were used in our MPPI analysis.

We found a total of 608 distinct alternative splice variants, 540 known and 68 novel; there were 216 more from the tumor sample than the normal sample (505 vs 289), reflecting greater cellularity and higher expression per cell. Of the 68 novel ASV proteins, 54 were from the tumor and 23 from the control sample, with 9 in common. Of the 15 biomarker candidates Whiteaker et al. [38] were able to confirm as over-expressed in tumor lysates with quantitative MRM-MS, we found that 10 had splice variants in our analysis; of course, we do not know the activities of different isoforms of these or any other proteins from proteomics analyses.

Among these 68 novel proteins we demonstrated variants resulting from new translation start sites, new splice sites, extension or shortening of exons, deletion or switch of exons, intron retention, and translation in an alternative reading frame. To validate the protein findings, we were able to design optimal primers for qRT-PCR analysis for 32 of the 45 novel peptides found only in the tumor sample. Each was amplified successfully; 31 of the 32 were validated, and 29 of the 31 showed increased mRNA expression [27].

In our annotations, 16 of the novel peptides found only in the tumor sample and with increased mRNA expression by PCR were highlighted because of functional motifs potentially significant in cancers. There were two variants with interesting annotations for BRCA. The peptide sequence 'FSRAEAEGPGQACPPRPFPC" is in the second intronic region of leucine-zipper-containing LZF (*rogdi*) gene (Fig. 3a). Using Splice Site Prediction by Neural Network from the Berkeley Drosophila Genome

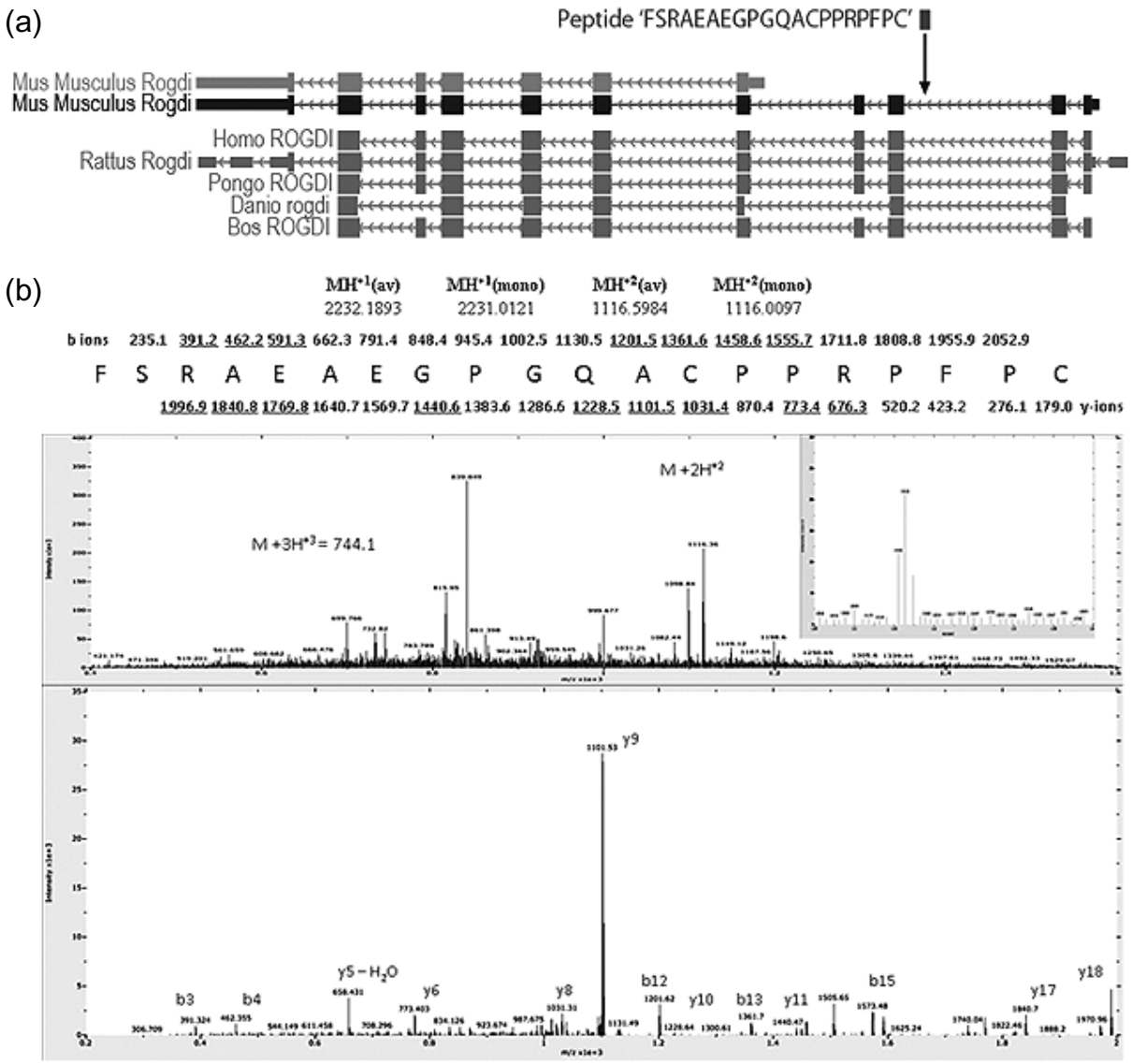

Fig. 3. **a:** A novel variant for leucine zipper domain (*Rogdi*) was identified by a unique peptide 'FSRAEAEGPGQACPPRPFPC' in the tumor sample. This figure shows the alignment of the novel peptide to the second intronic region of the mouse *Rogdi* gene by UCSC blat (chr 16: 5,012,685-5,012,744). The peptide also aligns perfectly with the homologous portion of the human, rat, orangutan, dog, cow and zebra fish *Rogdi* gene. The wide blocks are the exonic regions; the narrow blocks are the UTR regions; the line with arrows denotes the non-coding intronic regions. **b:** MS1 (top) and MS/MS (bottom) spectrum images of the novel peptide identified from the *Rogdi* gene. The inset of the MS1 spectrum shows a clearly defined precursor ion isotope envelope.

Project [http://www.fruitfly.org/seq_tools/splice.html], we found a predicted donor splice site 'gactgaggtgaggtg' where the novel peptide was identified as coding sequence with a splice site prediction score of 0.93. Functional motifs identified in this section of intronic sequence include LIG_BRCT_BRCA1_1, a phosphopeptide motif which interacts directly with the carboxy-terminal domain of BRCA1. The peptide 'GSGLVPTLGRGAETPVSGAGATRGLSR' aligned to the first intronic region of transcription factor *sox7*; the same LIG_BRCT_BRCA1_1 motif was found in this intronic region. We intend to model interactions with BRCA1 for both of these protein domains.

Two variants were annotated with a tyrosine-based sorting signal motif, which is particularly interesting due to tyrosine-based internationalization of the *neu* proto-oncogene product [15]. One is a novel variant of tyrosine-3-monooxygenase/tryptophan 5-

Table 1
Manual inspection of the spectra that identified the five novel peptides highlighted in the Her2/Neu breast cancer section. The six criteria used in manual inspection are: 1) the peptide sequence generated is biologically and chemically plausible; 2) there are two or fewer chemically significant tryptic missed cleavages; 3) the hydrophobicity of the peptide corresponds to the elution time in the HPLC reverse phase gradient; 4) the measured precursor ion is above noise with a clearly defined, accurate isotope envelope and appropriate alternative charge states; 5) the predominant fragments generated in the MS2 spectra are clearly defined above noise and are of predicted m/z for the peptide sequence, without prominent unpredicted fragment ions; and 6) exactly one charge is always lost for each subsequent stage of collision energy, e.g., a triply charged precursor will generate primarily doubly charged fragment ions and, conversely, a doubly charged precursor will not generate doubly charged fragment ions

Peptide sequence	Gene symbol	Criteria					
		1	2	3	4	5	6
FSRAEAEGPGQACPPRPFPC	*Rogdi*	x	x	x	x	x	x
GSGLVPTLGRGAETPVSGAGATRGLSR	*RSox7*	x	x	x	x	x	x
GHPGPEVWGGAGCGHGVCIFPAAVGAVEASFK	*RPkm2*	x	x	x		x	
RARLAEQASAMKAVTELNEP	*Ywhah*						
IYYSFGALKLGCFNFPLLKFL	No known gene	x	x	x	x	x	x

monooxygenase activation protein (*ywhah*), identified by peptide 'RARLAEQASAMKAVTELNEP', which is missing the 7 amino acids 'ERYDDMA' from its middle (amino acids 9–15). That missing sequence has the tyrosine-based sorting signal motif. Another peptide 'IYYSFGALKLGCFNFPLLKFL' aligns perfectly to a region in mouse chromosome 7 with sequence conservation in five other species, including human; two functional motifs link this unnamed protein to a tyrosine-based sorting signal TRG_ENDOCYTIC_2 responsible for interaction with the mu subunit of Adaptor Protein (AP) complex and to a MAP kinase docking function via LIG_MAPK_2. Then there are 12 variants with casein kinase II (CK2) phosphorylation, protein kinase phosphorylation (PKC), or N-myristoylation sites [27]. One of these is a new variant of pyruvate kinase muscle type (*pkm2*) identified by the peptide 'GHPGPEVWGGAGCGHGVCIFPAAVGAVEASFK'; the first 20 amino acids are from the middle section of exon 6 and the remaining 12 amino acids are from the middle section of exon 9. Two N-myristoylation sites and one PKC phosphorylation sites were found in this peptide sequence.

Employing spectral counting, we found 53 known splice variants differentially expressed. Using MotifScan with prosite patterns and prosite profiles as search parameters, we focused on the top 5 frequently occurring prosite patterns; CK phosphorylation, PKC phosphorylation, and n-myristoylation sites were found 1.5 times more frequently in these 53 variants than in 53 randomly selected normal proteins. We refer to these 53 known alternative splice variants and the 45 novel proteins found only in tumor sample as "tumor-associated splice variants".

We published spectra for each of these five peptides highlighted here and all 16 peptides extensively annotated in the original article (27, see Supplementary Material, Fig. 1). We also submitted the spectra to independent expert manual inspection (by Dr. A.K. Yocum, with re-review by all of us). See Table 1 for criteria of manual inspection of MS/MS spectra. Examination of the peptide sequence provides clues why a non-tryptic peptide sequence or a sequence that has missed cleavages is identified by database search methods with high scores. As an example, the peptide analysis for ROGDI in Fig. 3b contains two missed tryptic cleavage sites, one of which is followed by a proline, and a non-tryptic C-terminus which can be generated by common cysteine proteases. In the full-scan MS1 spectrum (Fig. 3b, top), both the doubly and triply charged precursor ions are found well above background noise levels. In the inset of the MS1 spectrum, clearly visible is the doubly charged isotope envelope in the proper relative distribution for the mass value of the analyte. Finally, the fragmentation spectrum (Fig. 3b, bottom) is of high intensity with each of the predominant peaks matching identifiable predicted peptide backbone fragments.

Three of the five highlighted novel spliced peptides met all six criteria of manual inspection in Table 1. The novel peptide for PKM2 met 4 of the 6 criteria; the novel peptide for YWHAH was retained because two known ywhah peptides were also identified, YWHAH (like PKM2) has been reported as involved in various cancers, and the novel peptide was identified only from tumor samples in this X!Tandem analysis. Thus, these novel peptides were submitted to validation using qRT-PCR; qRT-PR analyses showed specific amplification of the mRNA sequence corresponding to the novel peptides, together with increased expression of those mRNA sequences in the tumor sample relative to expression in the normal sample. If the peptide does not satis-

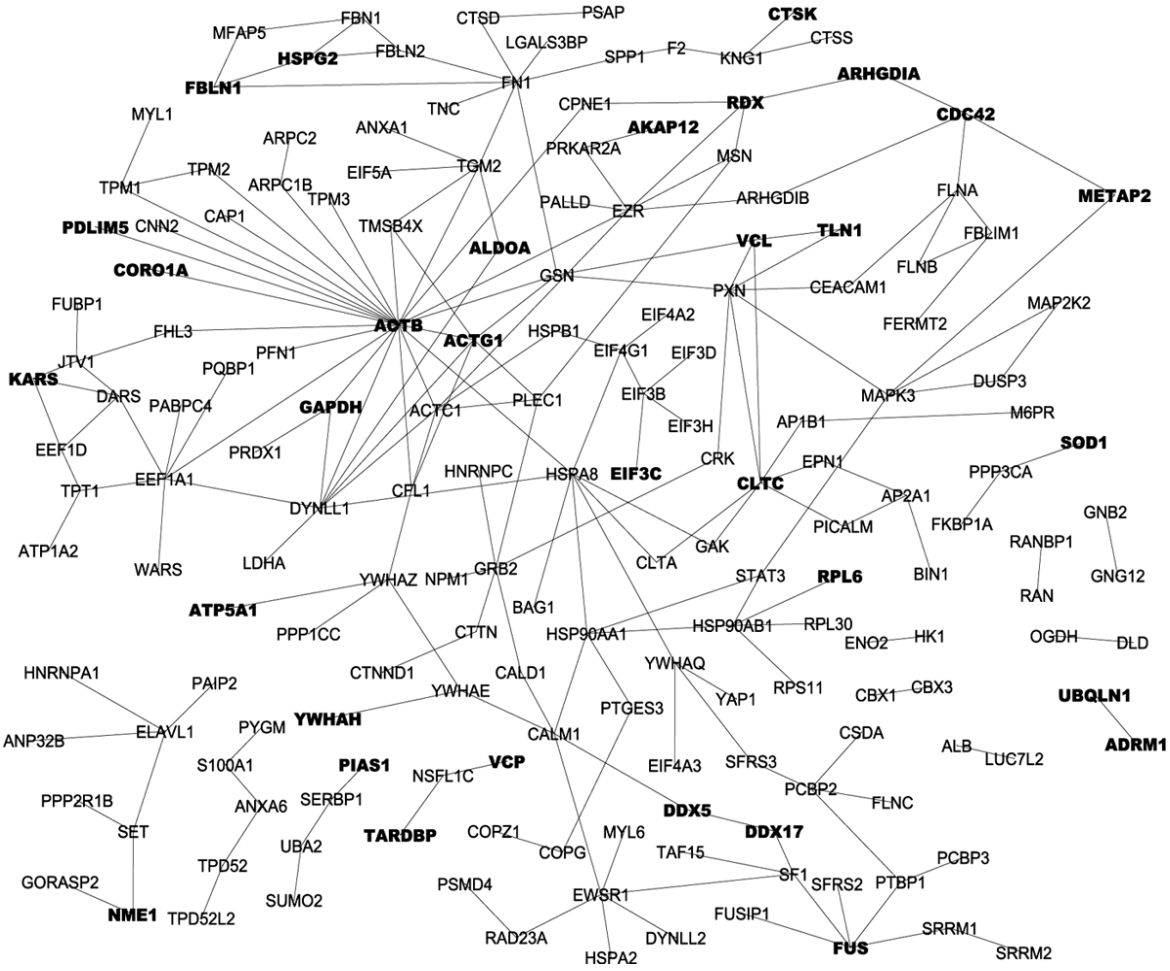

Fig. 4. Protein interaction network displayed in Cytoscape with Michigan Molecular Interactions (MiMI) Plug-in. The input gene list contained the alternative splice variants found only in the tumor sample. Only direct interactions between input genes are shown. The gene symbols in bold are those annotated in depth in the original publication. [From Cancer Research, Menon & Omenn, 2010 (in press)].

fy all six criteria unambiguously, the investigator must weigh all of the evidence, interpret with prudence, and follow up with orthogonal verification for the proposed alternative splice translated product, as was done in this study with qRT-PCR of the mRNA

Finally, we utilized GeneGo Metacore^{TM} software to characterize significant biological process networks. Cytoskeletal rearrangement, integrin-mediated cell adhesion, and translation initiation were found in common among the top-ranking networks from both all tumor-associated variants and the variants identified in the tumor sample. Figure 4 shows the direct protein interactions displayed by Cytoscape with MiMI plug-in; 177 of 460 input gene symbols are interacting. The gene names in bold denote differentially expressed alternative splice variants, including many of those we had already annotated as candidates for a role in the systems biology of breast cancer. CDC42, ARHGDIA, and RDX are among these proteins previously implicated in breast cancer mechanisms, as annotated extensively in our original publication [27]. Proline-rich motifs in specific splice variants of RDX and other proteins involved in extracellular matrix and cell motility can be contrasted with other known isoforms of these genes, which do not contain the proline-rich region. These motifs in these proteins participate in delivering actin monomers to cellular locations where ruffles, filopodia, and microspikes – actin-rich membrane protrusions – are formed.

6. Conclusion

We are in the earliest stages of identifying and evalu-

ating alternative splice variants for their potential roles in systems biology of cancers, especially critical features like initiation, progression, cell motility, invasiveness, and metastasis. Further research is needed to delineate major subtypes of common cancers, model the likely changes in structure and function of splice variants compared with prevalent isoforms, and propose and validate uses of these proteins as targets for therapy and biomarkers for diagnosis, prognosis, and response to treatments. In addition, the coding functions of the reverse transcriptome [17] and the regulatory functions of both strands warrant investigation, accelerated by next-generation sequencing methods.

References

[1] A.J. Aguirre, N. Bardeesy, M. Sinha et al., Activated Kras and Ink4a/Arf deficiency cooperate to produce metastatic pancreatic ductal adenocarcinoma, *Genes & Development* **17** (2003), 3112–3126.

[2] P. Argani, C. Rosty, R.E. Reiter et al., Discovery of new markers of cancer through serial analysis of gene expression: prostate stem cell antigen is overexpressed in pancreatic adenocarcinoma, *Cancer Res* **61** (2001), 4320–4324.

[3] N. Bardeesy, A.J. Aguirre, G.C. Chu et al., Both p16Ink4a and the p19Arf-p53 pathway constrain progression of pancreatic adenocarcinoma in the mouse, *Proc Natl Acad Sci USA* **103** (2006), 5947–5952.

[4] T. Barrett, D.B. Troup, S.E. Wilhite et al., NCBI GEO: mining tens of millions of expression profiles–database and tools update, *Nucl Acids Res* **35** (2007), D760–D765.

[5] M.B. Boxer, J.-k. Jiang, M.G. Vander Heiden et al., Evaluation of substituted N,N'-diarylsulfonamides as activators of the tumor cell specific M2 isoform of pyruvate kinase, *Journal of Medicinal Chemistry* **53** (2009), 1048–1055.

[6] L. Bracco and J. Kearsey, The relevance of alternative RNA splicing to pharmacogenomics, *Trends in Biotechnology* **21** (2003), 346–353.

[7] R. Chen, T.A. Brentnall, S. Pan et al., Quantitative proteomics analysis reveals that proteins differentially expressed in chronic pancreatitis are also frequently involved in pancreatic cancer, *Mol Cell Proteomics* **6** (2007), 1331–1342.

[8] H.R. Christofk, M.G. Vander Heiden, M.H. Harris et al., The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth, *Nature* **452** (2008), 230–233.

[9] H.R. Christofk, M.G. Vander Heiden, N. Wu et al., Pyruvate kinase M2 is a phosphotyrosine-binding protein, *Nature* **452** (2008), 181–186.

[10] R. Craig and R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, *Bioinformatics* **20** (2004), 1466–1467.

[11] E.W. Deutsch, The PeptideAtlas Project, in: *Proteome Bioinformatics*, S. Hubbard and A.R. Jones, eds, Methods in Molecular Biology (Totowa, NJ: Humana Press, 2010).

[12] J. Earl, L. Yan, L.J. Vitone et al., Evaluation of the 4q32-34 locus in European familial pancreatic cancer, *Cancer Epidemiology Biomarkers & Prevention* **15** (2006), 1948–1955.

[13] V.M. Faca, K.S. Song, H. Wang et al., A mouse to human search for plasma proteome changes associated with pancreatic tumor development, *PLoS Med* **5** (2008), e123.

[14] J. Gao, A.S. Ade, V.G. Tarcea et al., Integrating and annotating the interactome using the MiMI plugin for cytoscape, *Bioinformatics* **25** (2009), 137–138.

[15] L. Gilboa, R. Ben-Levy, Y. Yarden et al., Roles for a cytoplasmic tyrosine and tyrosine kinase activity in the interactions of neu receptors with coated pits, *Journal of Biological Chemistry* **270** (1995), 7061–7067.

[16] H.C. Harsha, K. Kandasamy, P. Ranganathan et al., A compendium of potential biomarkers of pancreatic cancer, *PLoS Med* **6** (2009), e1000046.

[17] Y. He, B. Vogelstein, V.E. Velculescu et al., The antisense transcriptomes of human cells, *Science* **322** (2008), 1855–1857.

[18] T. Hitosugi, S. Kang, M.G. Vander Heiden et al., Tyrosine phosphorylation inhibits PKM2 to promote the Warburg effect and tumor growth, *Sci Signal* **2** (2009), ra73–.

[19] A. Jemal, R. Siegel, E. Ward et al., Cancer statistics, 2009, *CA Cancer J Clin* **59** (2009), 225–249.

[20] N. Kim, S. Shin, and S. Lee, ECgene: Genome-based EST clustering and gene modeling for alternative splicing, *Genome Research* **15** (2005), 566–576.

[21] P. Kim, N. Kim, Y. Lee et al., ECgene: genome annotation for alternative splicing, *Nucl Acids Res* **33** (2005), D75–D79.

[22] Y.M. Kumar, K.M. Gurusamy, V.M. Pamecha et al., Tumor M2-pyruvate kinase as tumor marker in exocrine pancreatic cancer: a meta-analysis, *Pancreas* **35** (2007), 114–119.

[23] T.P. Larsson, C.G. Murray, T. Hill et al., Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery, *FEBS Letters* **579** (2005), 690–698.

[24] K. Lee, M. Hyun, and J.-R. Kim, Growth factor-dependent activation of the MAPK pathway in human pancreatic cancer: MEK/ERK and p38 MAP kinase interaction in uPA synthesis, *Clinical and Experimental Metastasis* **20** (2003), 499–505.

[25] C. Marks, Mouse Models of Human Cancers Consortium (MMHCC) from the NCI, *Disease Models & Mechanisms* **2** (2009), 111–111.

[26] R. Menon and G.S. Omenn, Identification of Alternatively Spliced Transcripts using a Proteomic Informatics Approach, in: *Data Mining in Proteomics*, M. Hamacher, C. Stephan and M. Eisenacher, eds, Totowa, New Jersey: Humana Press, 2010.

[27] R. Menon and G.S. Omenn, Proteomic characterization of novel alternative splice variant proteins in HER2/neu-induced breast cancers, *Cancer Res* **70** (2010), 3440–3449.

[28] R. Menon, Q. Zhang, Y. Zhang et al., Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer, *Cancer Res* **69** (2009), 300–309.

[29] G.S. Omenn, G.E. Goodman, M.D. Thornquist et al., Effects of a combination of beta-carotene and vitamin A on lung cancer and cardiovascular disease, *N Engl J Med* **334** (1996), 1150–1155.

[30] H. Parkinson, M. Kapushesky, M. Shojatalab et al., ArrayExpress – a public database of microarray experiments and gene expression profiles, *Nucl Acids Res* **35** (2007), D747–D750.

[31] E. Quevillon, V. Silventoinen, S. Pillai et al., InterProScan: protein domains identifier, *Nucl Acids Res* **33** (2005), W116–W120.

[32] D.R. Rhodes, J. Yu, K. Shanker et al., ONCOMINE: a cancer microarray database and integrated data-mining platform, *Neoplasia* **6** (2004), 1–6.

[33] D.M. Simeone, B. Ji, M. Banerjee et al., CEACAM1, a novel serum biomarker for pancreatic cancer, *Pancreas* **34** (2007), 436–443.

[34] K. Thorsen, K.D. Sorensen, A.S. Brems-Eskildsen et al., Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis, *Mol Cell Proteomics* **7** (2008), 1214–1224.
[35] M. Ventrucci, A. Cipolla, C. Racchini et al., Tumor M2-pyruvate kinase, a new metabolic marker for pancreatic cancer, *Digestive Diseases and Sciences* **49** (2004), 1149–1155.
[36] H. Wang, S.G. Clouthier, V. Galchev et al., Intact-protein-based high-resolution three-dimensional quantitative analysis system for proteome profiling of biological fluids, *Mol Cell Proteomics* **4** (2005), 618–625.
[37] O. Warburg, On the origin of cancer cells, *Science* **123** (1956), 309–314.
[38] J.R. Whiteaker, H. Zhang, L. Zhao et al., Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer, *J Proteome Res* **6** (2007), 3962–3975.