Identifying alternative splicing isoforms in the human proteome with small proteotranscriptomic databases

Edward Lau[1], Maggie Pui Yu Lam[2,*]

1 Stanford Cardiovascular Institute, Stanford University, Palo Alto, CA, USA.
2 Department of Medicine/Cardiology, Consortium for Fibrosis Research and Translation, University of Colorado Anschutz Medical Campus, Aurora, CO, USA.

*** Correspondence**
Maggie Pui Yu Lam, PhD
University of Colorado Denver - Anschutz Medical Campus
Mail Stop B139, Research Complex 2
12700 E. 19th Avenue
Aurora, CO 80045, USA
Email: maggie.lam@ucdenver.edu

Running title: Alternative isoforms in the human proteome
Keywords: Alternative splicing, proteoforms, isoforms, proteogenomics, proteotranscriptomics, proteomics
Abbreviations: A3SS: Alternative 3′ splice site; A5SS: Alternative 5′ splice site; MXE: Mutually exclusive exons; PKA: Protein kinase A; PTC: Premature termination codon; PTM: Post-translational modification; RI: Retained introns; SE: Skipped exons.

Abstract

Advances in next-generation sequencing have led to the discovery of many alternative splice isoforms at the transcript level, but the protein-level existence of most of these isoforms remains unknown. To survey the landscape of protein alternative isoform expression in the human proteome, we developed a proteotranscriptomics tool and workflow, which filters RNA sequencing data by junction reads before translating splice junctions into amino acid sequences. We further limit in silico sequence translation strictly to a single phase to reduce false positives in splice junction identification at the protein level. In total, we re-analyzed public RNA sequencing datasets and constructed custom FASTA databases from 10 human tissue types (heart, lung, liver, pancreas, ovary, testis, colon, prostate, adrenal gland, and esophagus). We used the custom database to identify splice junction peptides in proteomics datasets from the same 10 human tissues as well as 19 cardiac anatomical regions and cell types. We identified a total of 1,984 protein isoforms including 345 unique splice-specific peptides not currently documented in common proteomics databases. The proteotranscriptomics approach using restricted sequence databases described here may help reveal previously unidentified alternative protein isoforms, and aid in the study of alternative splicing at the proteome level.

Introduction

Alternative splicing constitutes a mechanism whereby multiple protein products may be created from a single gene [1], and is implicated in the regulation of development [2], aging, and multiple diseases including in the heart [3]. The human genome appears to be particularly enriched in the number of alternative splice isoforms, with RNA sequencing (RNA-seq) experiments uncovering over 100,000 alternative transcripts from virtually all multi-exonic genes in the human genome [4, 5]. Paradoxically however, relatively few alternative transcripts have been identified and confirmed at the protein level, and as a result, the biological significance of the majority of alternative isoforms remains obscure to-date. This is a critical knowledge gap that hampers our investigations into the functions of alternative isoform proteins, and the continued lack of protein-level evidence has even prompted some to question the extent to which alternative splicing influences cellular and physiological functions [6, 7].

We ask whether large-scale transcriptomics and protoemics data can be integrated in a multi-omics approach to reveal protein isoform expression in the human proteome across tissues and anatomical regions. Mass spectrometry is the preeminent tool for unbiased identification of protein sequences from complex biological samples, but faces several technical challenges for identifying alternative protein isoforms, including the low abundance of alternative exons [8] and incompatibility of many splice junction sequences with trypsin digestion [9]. Arguably most importantly, the omission of isoform peptides from precompiled sequence databases precludes their identification from peptide-spectrum matching in shotgun proteomics which typically relies on such databases. To circumvent this limitation, a number of approaches have been proposed including the use of curated splice variant databases [10, 11], or searching peptide spectra against all theoretical exon-exon junctions [12] or six-frame translation of the entire genome [13]. However, these approaches typically adulterate sequence databases with numerous imprecise sequences that do not exist in the biological samples, which inflate the false positive rate of identification and limit their utility [14]. Alternatively, RNA-seq data have also been leveraged to create sample-specific peptide databases translated from expressed transcripts in a sample, including splice variant transcripts [15, 16]. Initial works showed that this approach can identify novel peptides not found in annotated sequence databases [17], hinting at its potential utility for discovering protein isoforms. Thus far however, most studies of this type have only been performed in animal models or transformed human cell lines in place of primary tissues [15, 17, 18], the latter of which are known to express aberrant splice variants. Moreover, many custom transcript-guided databases remain imprecise and contain large proportions of unidentifiable translated sequences that may not exist in nature. Hence there is a need for methods for more precise in silico translation and evaluation of isoform proteins in human tissues.

Here we describe a new workflow that specifically targets translatable splice junction peptides from RNA-seq data to identify alternative protein isoforms. Our approach is distinguished by additional constraints of sequence translation to create a sparse sample-specific sequence database. Splice junctions are detected at the transcriptomics level, then filtered by expression level. Only non-constitutive splice junctions where both alternative splicing events are detected at the transcript level are considered. Moreover, we strictly enforce translation frame to restrict the sizes of the resulting databases such that they are smaller rather than larger than canonical sequence databases, in order to limit false positives. With the custom databases, we find that 86% of the genes expressed above threshold from RNA-seq data are identifiable with at least one isoform at the protein level at 5% global FDR per experiment. Among all computationally translated transcript isoforms (whether canonical or alternative), 17% are uniquely identifiable

by a unique junction peptide or splice-specific peptide. This approach supports both annotated protein isoforms and uncharacterized novel junction peptides to be detected from mass spectrometry experiments including in previously unidentified spectra.

Results

Generation of custom protein sequence databases

To generate custom protein sequence databases, we retrieved total RNA-seq datasets from ENCODE from the human heart, lungs, liver, pancreas, transverse colon, ovary, testis, prostate, and adrenal gland from the GTEx tissue collection (Methods). The transcripts were mapped to a reference human genome using STAR and analyzed to identify transcript reads spanning splice junctions (Figure 1). On average, we identified 73,111 alternative splicing events from RNA-seq data per tissue, with fewest in the adrenal gland (66,160) and most in the testis (91,895). By far the most common type of alternative splicing events identified was skipped exon, accounting on average for two-thirds (67.7%) of all identified events, followed by alternative $3'$ splice sites (11.5%), then retained introns (8.1%), alternative $5'$ splice sites (7.3%), and mutually exclusive exons (5.4%) (Figure 1b).

We wrote a custom computational workflow to translate the splice junctions to protein sequences in silico. Bearing in mind that large protein sequence databases would inflate false positives of detection, we decided on stringent criteria to enrich for splice transcript pairs that are likely to exist at the protein level. Because we are interested in non-constitutive exons for protein isoforms, we considered only alternative junction pairs where more than one splicing patterns can be detected at the transcript level. We then adopted a multi-staged filtering strategy (Figure 1a), first to remove low abundance transcripts with average read counts below 4 per sample that are more likely to be transcriptional noise (Figure 1), and secondly using the statistical model implemented in rMATS to remove transcripts that exhibit significant differences across technical and biological replicates in the same tissue ($P \leq 0.05$). Thirdly, we prioritized transcripts with known annotated transcriptional start sites and frame and could be translated in-frame without premature termination codons (PTC). Given the improbability of two functional protein products encoded by frame-shifted reading frames of the same sequence, we make the assumption that only one translation frame may be externally valid for any splice junction and hence we strictly enforce one-frame translation even when translation frame is not annotated, by prioritizing the sister frame that results in the longest translatable sequence with no PTCs. Lastly, to ensure that reliable junction peptides can be identified that span constitutive and alternative exons, we require both translated slices in a sequence pair to be able to be joined end-to-end back to full length canonical sequences from UniProt. Each translated sequence slice containing one upstream exon, the alternative exons, and one downstream exon where applicable was stitched back to SwissProt canonical sequences through a 10-amino-acid joint, and redundant entries were combined. Orphan splices that were not stitchable back to canonical sequences were discarded from further considerations.

The filtering strategy reduced the number of entries in the resulting sequence database. For instance, the human heart-specific database contains a representative 11,987 entries following filtering, as compared to 26,774 from unfiltered RNA-seq data. For comparison, the commonly used UniProtKB/SwissProt [19] sequence database catalogs 42,259 protein entries in the human reference proteome from 20,226 coding genes (20,226 canonical + 22,033 isoform sequences) as of November 2017. The TrEMBL component of the UniProt databases, containing un-reviewed sequences, contained additionally 93,583 sequences

including 53,500 with PE=1. Besides TrEMBL, larger databases exist from automatic annotation of genomic and transcriptomic sequences, but likewise it is unclear whether the majority of sequences are bona fide isoforms, fragments, polymorphisms, or redundant entries. Hence the generated database from our workflow is markedly sparser than the human canonical + isoform SwissProt database (42,259 entries), TrEMBL (93,555 entries), RefSeq (109,706 entries), and indiscriminate six-frame translation of transcripts (157,544 entries). This is expected because each tissue is expected to only express a subset of genes in the human genome due to tissue-specific epigenetic landscape and gene regulation, and RNA-seq data provide information only on the specific genes that are expressed above threshold levels in the sample being analyzed (e.g., heart tissues). In total, we generated 10 filtered human tissue-specific protein sequence databases (Figure 1c on database size). The databases contain on average 13,249 total splice-junction specific protein sequence entries, with the human pancreas-specific database containing the fewest protein sequences (7,786) and the testis containing the most entries (18,167).

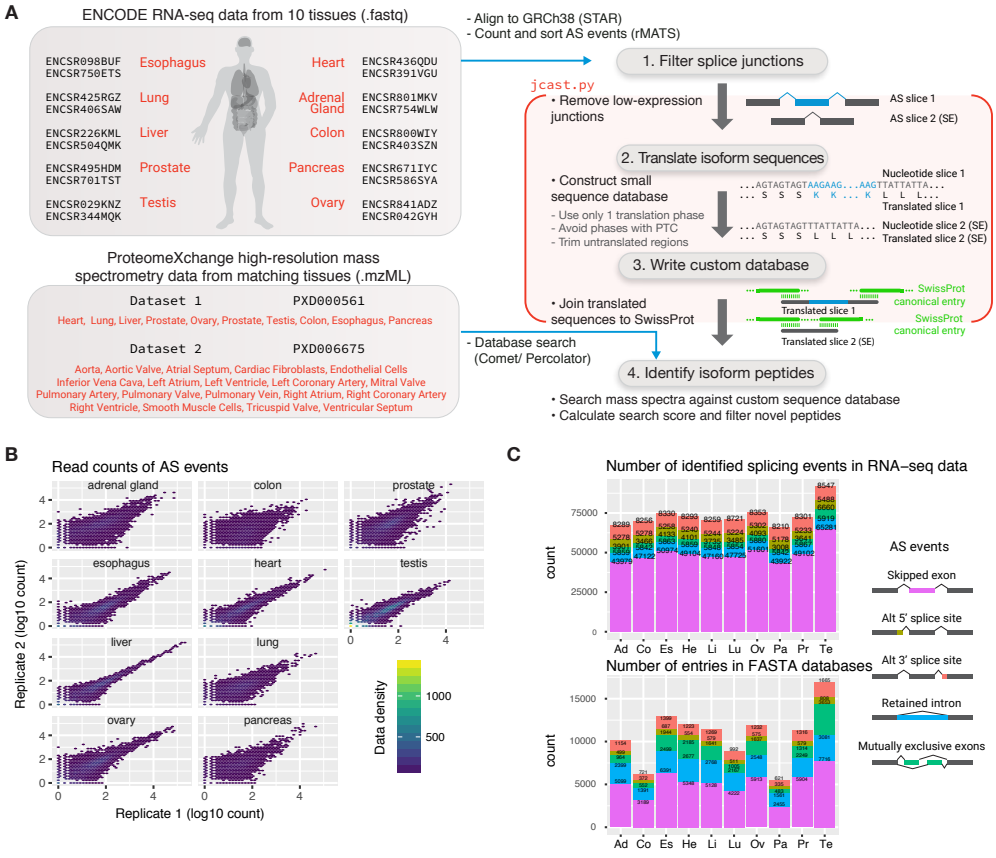

Figure 1. Generation of filtered protein isoform databases from tissue-specific RNA sequencing data. **A** Schematic of database generation workflow. RNA sequencing data from 10 human organs were downloaded from ENCODE. **B** Density scatter plot showing the distribution of log10 read counts in splice junctions of replicate RNA-seq data (SJC_SAMPLE_1 and SJC_SAMPLE_2). Splice junctions with read counts fewer than 10 in both samples are excluded from in silico translation. **C** Number of identifiable splicing events in the RNA sequencing data. **D** Number of translated isoform entries in the generated FASTA database.

Identification of splice junction peptides in proteomics datasets

Using the 10 tissue-specific custom databases, we re-analyzed a publicly available set of shotgun proteomics data on 10 matching human tissues [20] ("Dataset 1" hereafter). In total, we identified 112,625 peptides associated with 6,189 gene entries (SwissProt canonical entries) at a false discovery rate of 5% (Figure 2). Moreover, because in our databases we translated splice junction in pairs (i.e., both sequences within a binary alternative splicing event were always translated together), we were able to determine which identified sequences were splice junction sequences or splice-specific sequences based on their assignment to unique FASTA entries. In other words, isoform-specific peptides that are matched to only one singular entry within the custom database are from splice-specific exons or splice junctions. In total, we identified 7,649 unique sequences from 1,855 genes that denote splice-junction specific peptides, which include also annotated canonical junctions, when distinguishable against translated alternative junctions. The number of uniquely identifiable splice junctions is likely an underestimate, because a proportion of alternative junction peptides would appear in multiple custom translated forms due to the combinatorials of splice junctions, rendering them non-unique in the database. This will likely require full transcript models such as from long-read RNA-seq to resolve.

Building on this result, we further analyzed a more recent and comprehensive mass spectrometry dataset on the human heart [21]. We used the custom generated heart isoform protein database as a reference database for re-searching raw mass spectrum data from 16 sub-anatomical regions and 3 isolated cell types from human donor hearts. From this dataset ("Dataset 2" hereafter) we identified 114,733 peptides belonging to 3,573 cardiac genes, including 5,875 unique peptides from 1,984 isoforms belonging to 1,430 cardiac genes that denote splice-junction peptides or splice-specific peptides (Figure 2 on identification rate in each region). We note that the heart-specific database contains 4,110 unique genes, hence over 86% of all translated genes were detectable at the proteome level which suggests the RNA-seq database was precise with regard to the gene loci that were translated; Combining the Dataset 1 and Dataset 2, we identified 11,593 isoform-specific sequences from 2,572 genes. We found that, unsurprisingly, the majority of the identified splice peptides originated from alternative splicing events where both splice variants are in frame, with the exception of RI peptides encountering PTC frameshifted peptides are more common, and A5SS, where a number of identified peptide sequences encountered frameshifts (Figure 2).

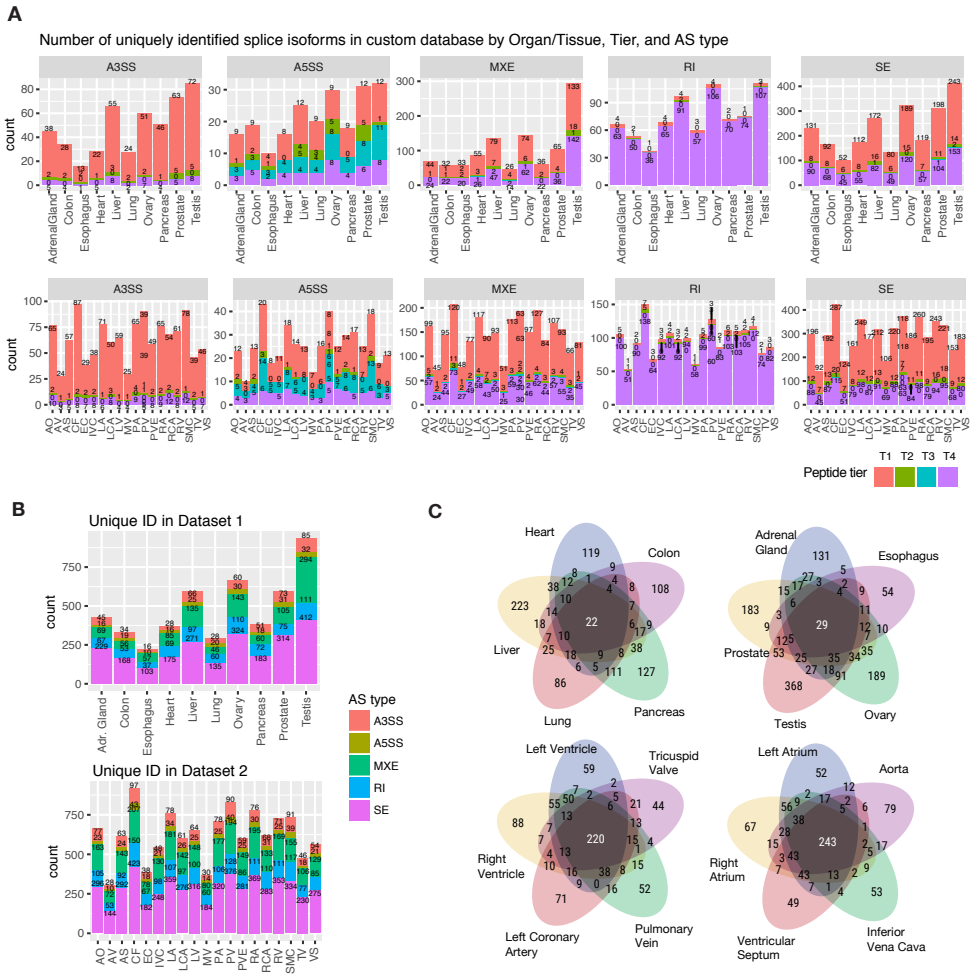

Figure 2. Unique peptide ID by junction type. **A** The number of uniquely-identified splice peptides in Dataset 1 (upper) and Dataset 2 (lower) are grouped by alternative splicing type (A3SS, A5SS, MXE, RI, and SE). Each bar chart is further subcategorized by peptide translation tiers. Tier 1 (red) represents peptides that are translated in-frame by the annotated translation frame in Ensembl GRCh38.89 GTF successfully without encountering a frameshift or premature termination codon (PTC); tier 2 (green) peptides are those translated without PTC using the annotated frames but for which one of the spliced pairs encountered a frameshift; tier 3 peptides (cyan) are translated without frameshift or PTC but using a different translation frame than annotated in Ensembl; tier 4 peptides encountered PTC in one of the two splice pairs. All translated peptides are stitchable back to SwissProt canonical sequences by a 10-amino-acid joint. **B** Total number of isoform-specific peptides identified in each sample from Dataset 1 (upper) and Dataset 2 (lower). Each stacked barchart is further partitioned based on the alternative splicing type of the identified peptide. **C** Venn diagrams showing differential overlaps in the identified splice isoform peptides in selected tissues and cardiac cell types. AO: aorta; AV: atrial valve; AS: atrial septum; CF: cardiac fibroblasts; EC: endothelial cells; IVC: inferior vena cava; LA: left atrium; LCA: left coronary artery; LV: left ventricle; MV: mitral valve; PA: pulmonary artery; PV: pulmonary valve; PVE: pulmonary vein; RA: right atrium; RCA: right coronary artery; RV: right ventricle; SMC: smooth muscle cells; TV: tricuspid valve; VS: ventricular septum.

The plurality of identified junctions were canonical or alternative variants that spanned SE events (47%), followed by MXE (23%), RI (16%), A3SS (10%) and A5SS (4%). Compared to the proportional representation of each AS type in the translated database, the identification result suggests that in our analysis, MXE has higher identification rates than other splice types (23% identified isoform-unique peptides vs 14% translated database entries) whereas RI has a lower-than-expected identification rate (16% identified sequence vs 21% translated database entries). The identified splice junctions are highly organ-specific with relatively few overlaps, suggesting many of the proteins with identified splice junctions are expressed primarily in specific organs (Figure 2). The anatomical regions and cell types of the heart in Dataset 2 show substantially higher overlap in identified isoforms than those across the organs from Dataset 1. Anatomically related regions, i.e., the left and right ventricles and the left and right atria, in particular share proteins with identified isoform junctions.

Identification of novel splice junction peptides

We next compared the identified splice-specific peptides to protein sequences found in the protein knowledgebase UniProt, which was chosen for the comparison because it is widely utilized and is produced by high-quality curation. UniProt contains two components – SwissProt, which contains manually curated non-redundant protein entries, and the larger TrEMBL, which contains computationally analyzed records [19]. First we determined whether we identified peptides that are not currently catalogued in the SwissProt canonical and isoform sequences, and hence may represent potential novel protein isoforms. From all confidently identified peptides in Dataset 1 and Dataset 2 combined, we found 1,878 peptides identified at $q \leq 0.05$ that are not matched to any entry in either SwissProt canonical or isoform sequences, belonging to 1,098 genes. Moreover, 1,227 peptides (65%) belonging to 826 proteins are also not matched to any entry in the TrEMBL isoform and canonical sequence database, which encompasses all SwissProt entries plus computationally annotated and unreviewed sequences (Supplementary Table 1). This result suggests that a number of TrEMBL sequences are likely uncharacterized splice isoforms that are expressed at the protein level, and provides tentative evidence for the protein-level existence of hundreds of isoform sequences not documented in UniProtKB. TrEMBL-unique isoforms can be found in every tissue and sub-anatomical regions analyzed in Dataset 1 and Dataset 2, with particular enrichment in the testis, which is known to differ more markedly from other tissues in alternative splicing pattern.

We observed that the novel peptides tended to have lower spectral identification scores compared to non-novel spectra at percolator false discovery rate (q-value) ≤ 0.05 (Figure 3). This may not be totally unexpected, in part because of the lower proportional abundance of some alternative isoforms, but also because the genome-wide enrichment of lysines at splice junctions leads to a higher proportion of miscleavages for identifiable junction peptides, which is a feature that is penalized in the scoring algorithm of Percolator [9]. Nevertheless, because the data do not rule out the possibility of inflated false discovery rates among the novel peptides, we adopted additional filtering strategies to help determine whether the identified novel peptides from the database search workflow represent bona fide peptide-spectrum matches.

Table 1. Number of identified novel peptides in Dataset 1 and Dataset 2. Number of identified peptides in Dataset 1 and Dataset 2 and their associated genes that are not found in SwissProt (SP0_pep, SP0_prot), not found in TrEMBL (TR0_pep, TR0_prot), and not found in TrEMBL after allowing one mismatch (TR1_pep, TR1_prot).

Sample	Dataset	SP0_pep	SP0_prot	TR0_pep	TR0_prot	TR1_pep	TR1_prot
AdrenalGland	1	80	75	55	54	47	46
Aorta	2	187	138	116	99	100	86
AorticValve	2	46	43	29	28	23	22
AtrialSeptum	2	230	133	115	89	102	79
CardiacFibroblast	2	299	192	143	130	124	113
Colon	1	64	59	44	44	40	40
EndothelialCells	2	71	64	39	36	28	26
Esophagus	1	35	28	19	18	12	11
Heart	1	144	77	67	50	55	41
InferiorVenaCava	2	140	114	101	86	84	70
LeftAtrium	2	273	160	128	99	105	81
LeftCoronaryArtery	2	169	135	114	100	96	85
LeftVentricle	2	262	140	117	89	98	78
Liver	1	114	105	75	68	55	49
Lung	1	44	39	29	27	23	21
MitralValve	2	89	78	59	57	49	47
Ovary	1	129	104	82	72	73	64
Pancreas	1	57	50	31	28	26	24
Prostate	1	100	95	68	68	51	51
PulmonaryArtery	2	201	132	109	88	95	74
PulmonaryValve	2	80	68	47	46	40	40
PulmonaryVein	2	199	117	94	74	82	64
RightAtrium	2	279	160	121	93	104	81
RightCoronaryArtery	2	177	144	113	101	95	84
RightVentricle	2	297	169	139	108	119	93
SmoothMuscleCells	2	186	138	89	85	76	72
Testis	1	206	164	117	103	97	83
TricuspidValve	2	97	86	59	58	49	48
VentricularSeptum	2	222	124	107	81	88	69

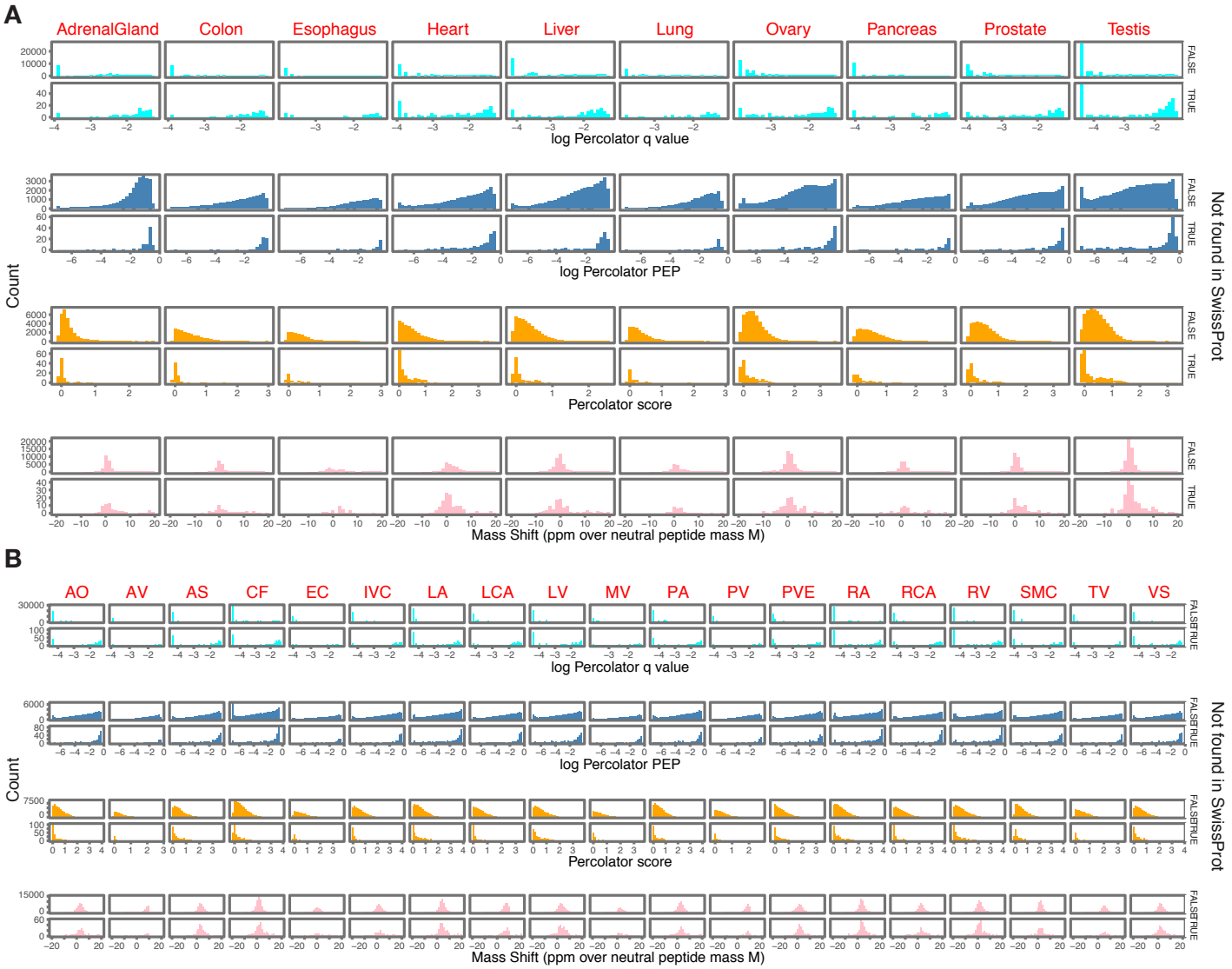

Figure 3. - Score distributions of novel and SwissProt peptides. For (**A**) Dataset 1 and (**B**) Dataset 2, the histograms of Percolator log_{10} q-value, posterior error probability (PEP), peptide score, and mass shift (ppm) for peptides are shown. Data are categorized by whether the peptide sequence is found in SwissProt canonical and isoform database (top of each strip) or not found in SwissProt canonical and isoform database (bottom of each strip). AO: aorta; AV: atrial valve; AS: atrial septum; CF: cardiac fibroblasts; EC: endothelial cells; IVC: inferior vena cava; LA: left atrium; LCA: left coronary artery; LV: left ventricle; MV: mitral valve; PA: pulmonary artery; PV: pulmonary valve; PVE: pulmonary vein; RA: right atrium; RCA: right coronary artery; RV: right ventricle; SMC: smooth muscle cells; TV: tricuspid valve; VS: ventricular septum.

To achieve this, we manually inspected tandem mass spectra of potentially novel junction peptides. We note that in the scoring of the database search algorithm, some peptide-spectrum matches contain relatively fewer matched fragmentation peaks but are nevertheless identified at high confidence, possibly due to the peak distributions in the available competitor theoretical spectra candidates within the precursor mass window. For stringency we therefore counted the peptide-spectrum matches of novel spectra by heuristic criteria on tandem mass spectrum quality alone, and prioritized spectra that pass a threshold of fragment ion count as determined by sequence length (Figure 4).

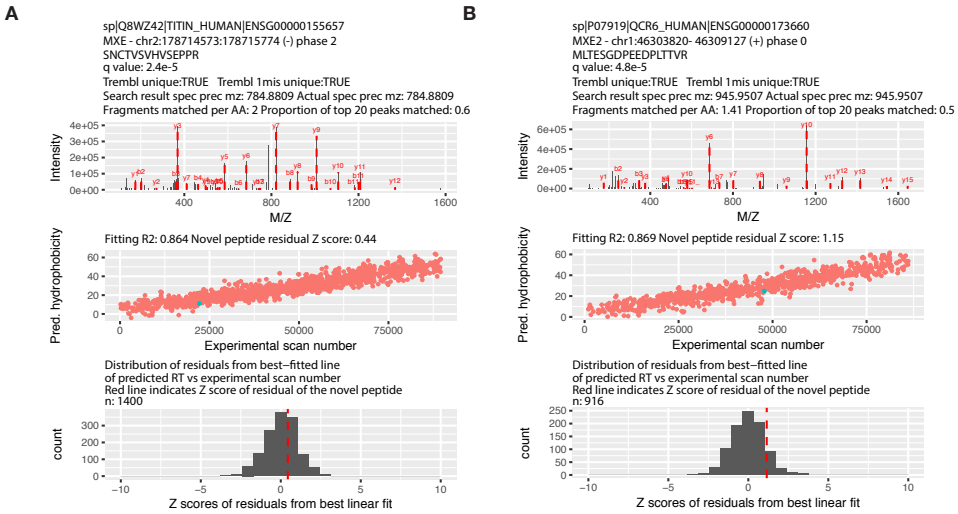

Figure 4. Spectral evidence of identified novel splice junction peptides. Tandem mass spectrum and predicted elution time of two identified junction peptides are shown; the remaining peptides can be found in Supplementary Data 1 and Supplementary Data 2. **A** Peptide SNCTVSVHVSEPPR from protein TTN and **B** Peptide MLTESGDPEEDPLTTVR from protein UQCRH. Neither peptide is matched to SwissProt or TrEMBL isoform sequences. Searching the database allowing one mismatch determines they are unlikely to be due to single amino-acid variants or unknown modifications at a single residue. (Upper) Visual inspection shows tandem mass spectra with matched peaks; (middle and bottom) predicted hydrophobicity compared to high-quality spectra from the same fraction suggests the peptide eluted at expected retention time.

In addition, we determined whether the identified novel isoform peptides eluted in expected positions in the liquid chromatograph gradient based on the predicted hydrophobicity of their assigned peptide sequences. We used an existing algorithm to determine the hydrophobicity coefficient of each novel sequence [22], and matched them against the order in spectrum scan number for which the peptide was identified to determine whether the peptides eluted at the right time. We benchmarked the novel peptides against high-confidence peptides within the same experimental mass spectrum file i.e, the same fraction of the same sample. High-confidence peptides are defined as those found among reviewed SwissProt human sequences, contain no post-translational modification, and have Percolator posterior error probability (PEP) of identification of less than 0.05. We prioritize novel peptides whose retention time residual to the best-fitted curve between predicted hydrophobicity and empirical retention time is within three standard standard deviations from the mean of that of high-quality peptides (Figure 4).

Among the identified proteins with alternative isoforms across the analyzed datasets

here, protein kinase AMP-activated non-catalytic subunit beta 1 (PRKAB1) has the most number of novel isoform peptides not found in UniProt (29) followed by titin (TTN) (18) (Table 1). The latter is by far the largest protein encoded by the human genome as well as consisting of the greatest number of exons, the splicing of which has been increasingly implicated with dilated cardiomyopathy [23]. Proteins with more than one novel peptide can be found across diverse pathways including muscle contraction (MYBPC3, MYL2), metabolism and energetics (IDH3A, CKMT2, NDUFA13), as well as signaling (MAP3K3, SORBS1). The identified novel peptides are mapped to genes spanning all 23 human chromosomes in the human genome, suggesting uncharacterized protein isoforms at the proteome level are pervasive (Figure 5).

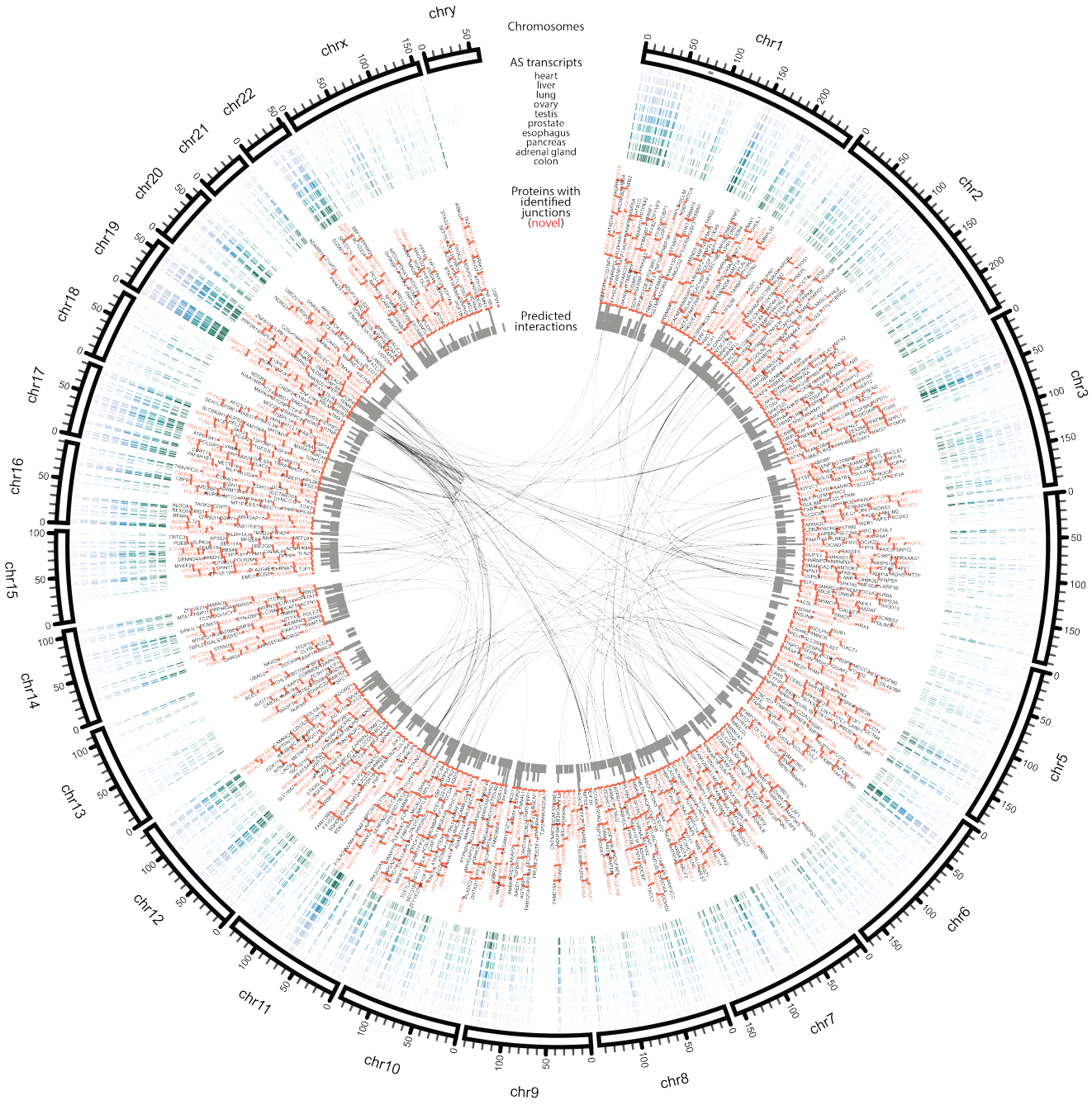

Figure 5. Landscape of protein alternative isoforms across human chromosomes From outer to inner rings, ideogram of 23 human chromosomes, followed by 10 tracks showing the positions of genes containing the top 10,000 alternative splice junctions based on splicing junction read counts in 10 human tissues. Text tracks show the symbols of proteins with identified splice-specific peptides (i.e., splice junction peptides or spliced exon specific peptides). Red symbols denote proteins for which junction peptides have been identified that are not catalogued in the SwissProt human canonical and isoform database. Inner links show predicted protein-protein interactions (StringDB combined score ≥ 500) among proteins with known isoform-specific peptides in this study. The identified novel splice isoforms are widespread and span the human genome.

Table 2. Proteins with most number of identified novel peptides. List of top 20 proteins with the most number of identified peptides not found in the SwissProt human canonical and isoform database are shown.

Symbol	Gene Name	#_novel_peps
PRKAB1	protein kinase AMP-activated non-catalytic subunit beta 1	29
TTN	titin	18
SORBS1	sorbin and SH3 domain containing 1	14
MYBPC3	myosin binding protein C, cardiac	6
HEG1	heart development protein with EGF like domains 1	5
KIAA1210	KIAA1210	5
BBS1	Bardet-Biedl syndrome 1	4
EEF1A2	eukaryotic translation elongation factor 1 alpha 2	4
NEXN	nexilin F-actin binding protein	4
TAF1C	TATA-box binding protein associated factor, RNA polymerase I subunit C	4
TNS1	tensin 1	4
TIAL1	TIA1 cytotoxic granule associated RNA binding protein like 1	4
ABHD17B	abhydrolase domain containing 17B	3
DPP7	dipeptidyl peptidase 7	3
NRAP	nebulin related anchoring protein	3
PALLD	palladin, cytoskeletal associated protein	3
SRRM1	serine and arginine repetitive matrix 1	3
SNX5	sorting nexin 5	3
SVIL	supervillin	3

To further evaluate the quality of protein identification, we note that among the 1,227 peptide sequences not found in Trembl, 1,043 (85.0%) also could not be matched to any Trembl human sequences even when allowing a residue mismatch at any position, suggesting the absolute majority of identified spectra are unlikely to arise from single amino acid variants (SAAV) differing from the reference proteome, or from an unaccounted-for mass shift at a single residue. We also compared the search results in one cardiac anatomical region (left ventricle) to database search on identical data using first a non-filtered translated database containing all RNA-seq sequences regardless of read counts and without stitching back to UniProt, and secondly a database containing six-frame translation of transcripts to mimic a genomic database. In instances where ground-truth protein expression may be approximated, six-frame translation is known to improve recall of protein identification but at the severe expense of precision and computational time [16]. Indeed, we observed that unfiltered RNA-seq database and six-frame translation both led to substantial increase in search time coupled to a sharp decrease in the proportion of identifiable unique splice sequences by database size from identical spectra (Left Ventricle from Dataset 2) at a uniform false discovery rate of 1%, from 4.3% of isoform fasta entries (513/11,987) to 2.7% (711/26,774) and 0.4% (669/157,544), respectively. This result suggests the databases in our workflow show excellent recall in capturing true translated isoform sequences, and is consistent with previous observation that indiscriminate translation led to modest increase in recall but heavy penalty in precision [16]. Combining the additional filtering criteria, we prioritized 424 novel isoform peptide candidates from 345 genes whose peptide-spectrum matches and retention time estimates are shown in Supplementary Data 1 and Supplementary Data 2.

A splice junction peptide overlapping PKA regulatory sites on MYBPC3

We next inspected the potential proteome impact of the identified splice isoforms. Here we zoomed in on a splice variant for myosin-binding protein C3 (MYBPC3). MYBPC3 is a $\approx$ 140 kDa cardiac protein, which forms an important sarcomeric component in

cardiomyocytes responsible for maintaining cardiac sarcomere structure, and which is one of the most commonly mutated genes in human familial hypertrophic cardiomyopathy. From a translated skipped exon event, we identified the splice junction peptide RTD-SHEDTGILDFSSLLK which is not found in UniProtKB/TrEMBL even allowing one mismatch, as well as its sister peptide TDSHEDTGILDFSSLLK (Figure 6, panel a). The sequences are identified widely in the tissues within the datasets analyzed here, including whole heart, left atrium, and left ventricle. Both alternative sister peptides are also not documented on PeptideAtlas which catalogues identified peptide sequences as of the time of query (2018-07-19) and are furthermore not identically matched (100% sequence identity and coverage) to any sequences of any taxonomy in RefSeq via BLASTP [24].

UniProtKB/SwissProt catalogues two isoform protein entries for MYBPC3, including the SwissProt canonical entry with 1,274 residues and an alternative entry with 1,273 residues, in which the serine 408 and lysine 409 of the canonical entry are replaced by a single arginine. Neither entry encompasses the alternative isoform sequence we identified, which omits the segment of amino acid sequence SLAGGGRRIS from residues 275 to 284 encoded by exon 8, and which contains instead the identified peptide matching to residues 273 to 274 (RT-) of the canonical sequence, then resuming after the splice junction at residues 285 to 300 (-DSHEDTGILDFSSLLK) (Figure 6, panel b). This skipped exon falls within a region in between two well-defined motifs, namely an Ig-like C2-type 1 motif before and the following Ig-like motif, but is itself predicted to be within a disordered region (Figure 6, panel c). It has been suggested that the majority of alternative splicing events do not alter conserved protein motifs which has been cited as evidence against the importance of alternative splicing to the proteome [6, 7]; however, intrinsically disordered regions can also contain regulatory hotspots for interaction surfaces and post-translational modifications [25], hence we asked whether the excised region overlapped with other structural features of interest. Interestingly, we found that the skipped exon falls on two of three successive phosphorylation sites (S275, S284, S304) that are known to be key regulatory sites in MYBPC3 by protein kinase A (PKA). The phosphorylation of S275, S284, and S304 in MYBPC3 by PKA as well as by other kinases is known to cause the MYBPC3 N-terminal domain to dissociate from myosin heavy chain and hence an increase in cardiac crossbridge formation [26]. Mutagenesis experiments in animal models replacing these serines with the phospho-negative mimetic alanine have found that the resulting hearts had abnormal relaxation velocity but not ejection fraction [26], suggesting these sites exert diastolic regulations on the sarcomere and may be associated with diastolic dysfunction. Taken together, the result provides evidence of an protein alternative isoform that overlaps with known regulatory post-translational modification sites and a potential mechanism through which it may regulate cellular function.

We further compared the relative abundance of the SwissProt canonical sequence of MYBPC3 and the identified alternative splice junctions across cardiac regions, and found that the integrated abundance of the alternative sister peptides to be significantly enriched in atria over the ventricles (two-tailed Student's t-test P value 0.03). Although determining the exact splice ratios will likely require isotope standards due to the difference in ionization efficiency of the peptides, the data suggest that the alternative isoform displays cardiac chamber-specific expression (Figure 6, panel d).

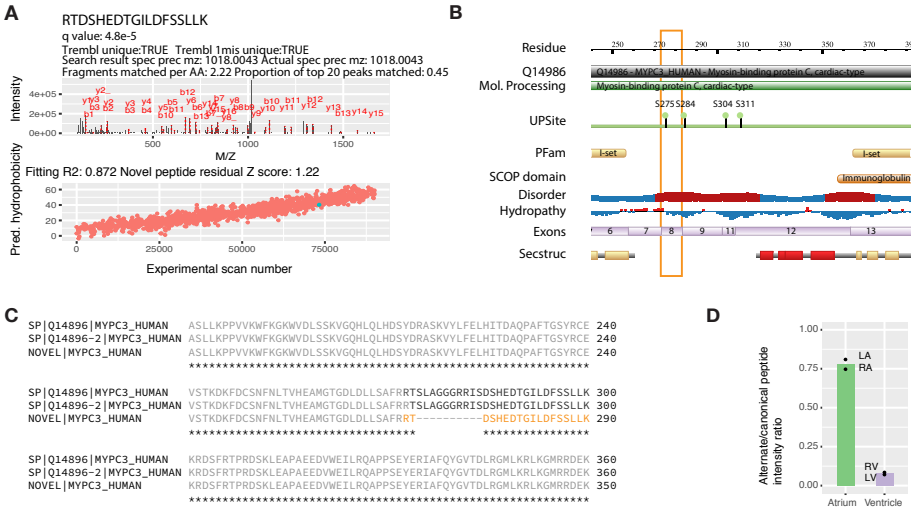

Figure 6. An alternative protein isoform distinguishes a protein kinase A regulatory site on MYBPC3. A Mass spectrum and elution time of peptide RTDSHEDTGILDFSSLLK belonging to the protein MYBPC3. **B** Amino acid sequence tracks showing the region of MYBPC3 in exon 8 affected by isoform expression (orange box) overlaid on annotation tracks including known PKA-phosphorylation sites, sequence disorder, hydropathy, and known domains. **C** Amino acid residues of the alternative regions in the canonical UniProt sequence (top), the documented alternative isoform of MYBPC3 on UniProt (middle), and the identified novel sequence from RNA-translated databases (bottom). **D** Relative label-free quantity ratios of peptides consistent with alternative vs. canonical isoform in atrial vs. ventricular samples in the heart suggesting the novel isoform is enriched in the atrium.

Splice isoform peptides explain some common unidentified spectra

Lastly, we asked whether some commonly unidentified spectra from shotgun proteomics experiments might represent unannotated splice junctions. The Proteomics Identification database (PRIDE) Cluster [27] (2015-04 release) contains a spectral library of commonly unidentified spectra from uploaded proteomics datasets on the PRIDE database operated by the European Bioinformatics Institute (EBI), including a collection of 22,344 high-quality unidentified clustered spectra in the human_100 collection, each of which was clustered from over 100 uploaded and unidentified spectra.

We queried this spectrum library with the custom databases for heart, liver, and testis against PRIDE cluster (2015-04 release) and were able to salvage 76 unique peptide sequence identifications at percolator q $\leq$ 0.05 (Table 3) . The identified peptides include eight peptide sequences from three proteins that uniquely define splice junctions in the databases, including one that is not found in Trembl (ERCC2 IEQIAQQSWSPGDALR from the liver database). Hence, at least some commonly unidentified peptides are likely due to splice junction peptides not present in employed sequence databases. We note that the result here is likely a marked underestimate of the number of unidentified splice isoforms, given that PRIDE cluster documents only clustered unidentified MS2 spectra from over 100 experiments, and hence is biased against human tissues currently less commonly studied by proteomics, and further would not include alternative splice isoforms expressed only in specific cell types or specific developmental stages. As the number

of proteomics experiments searched using conventional sequence database continues to increase, we foresee that the number of recoverable isoform sequences will increase as well.

Table 3. Commonly unidentified spectra include Isoform-specific sequences. We queried clustered unidentified spectra from 100+ human experiments on PRIDE cluster, using the heart, liver, and testis-specific isoform databases. Peptides identified at Percolator peptide q-value lower than 0.05 are included. Z: charge of mass spectra.

Z	m/z	q-value	Sequence	DB	TrNovel	UniProt	Symbol	Gene Name
2	648.88	0.0417	PVGSLAGIGEVLGK	Heart	FALSE	O75531	BANF1	barrier to autointegration factor 1
2	751.402	0.0417	VINGNPITIFQER	Heart	FALSE	P04406	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
2	807.946	0.0417	LVINGNPITIFQER	Heart	FALSE	P04406	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
2	751.402	0.0435	VINGNPITIFQER	Liver	FALSE	P04406	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
3	600.306	0.0435	IEQIAQQSWSPGDALR	Liver	TRUE	P18074	ERCC2	ERCC excision repair 2, TFIIH core complex helicase subunit
2	648.88	0.0345	PVGSLAGIGEVLGK	Testis	FALSE	O75531	BANF1	barrier to autointegration factor 1
2	807.946	0.0345	LVINGNPITIFQER	Testis	FALSE	P04406	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
2	751.402	0.0345	VINGNPITIFQER	Testis	FALSE	P04406	GAPDH	glyceraldehyde-3-phosphate dehydrogenase

Discussion

Recent studies show that alternatively spliced exons might be enriched in flexible regions and interaction surfaces of proteins, suggesting alternative splicing may rewire protein interaction networks [28, 29], whereas differential alternative splicing may also remodel the proteome by regulating the abundance of canonical isoforms [30]. The ability to discriminate which isoform transcripts exist at the protein level is required to better understand the biochemical and signaling functions of alternative isoforms. It is now appreciated that only a subset of expressed transcripts have the potential to be translated, whereas the rest may be removed by nonsense-mediated decay or co-translational proteasomal degradation [31]. Approximately one-third of skipped exon events have been estimated to preserve translation frame and protein structure, and hence may be assumed to have higher translational potential [32]; nevertheless, empirical evidence of their existence as isoform proteins has remained scarce in part due to difficulty in analyzing mass spectrometry data.

Using a custom RNA-seq translated database method, we showed that it was feasible to improve on the identification of alternative protein isoforms from public mass spectrometry data over the more commonly employed canonical or isoform sequence databases. Our workflow combines transcriptomic level modeling of splice junctions followed by shotgun proteomics analysis, a strategy we recently termed a "proteotranscriptomics" approach [33]. One key feature of the approach is a highly restricted protein sequence database generated from organism-specific and tissue-specific RNA-seq data and following the removal of transcripts that are expressed at minuscule levels in the tissue. For instance, the human heart specific database contains only 12.8% as many sequences as there are in TrEMBL and 7.6% compared to a six-frame translation database. The use of RNA-seq reads from matching samples to filter out untranslated proteins has been previously demonstrated to improve confidence assignment [34]. By combining RNA-seq data with statistical comparisons of splice events in multiple RNA-seq replicates, we were able to confidently assign over a thousand alternative isoform peptides and hence provide evidence of their existence at the protein level. Mapping the landscape of isoform expression would be of relevance to research aiming to uncover the translation status of all proteins including "missing proteins" from each chromosome in the human genome, such as in the Human Proteome Organization (HUPO) Chromosome-based Human Proteome Project (C-HPP) initiative (Figure 5).

The identified isoform peptides include a splice junction in MYBPC3, which differs from the canonical version by 10 amino acids and occupies a disordered region between two well-defined Ig-like motifs. It has been suggested that the enrichment of splice isoforms in disordered regions and the preservation of domains argue against the functional significance of alternative isoforms. Many disordered regions on protein molecules serve important regulatory functions, but in practice it is challenging to ascertain the functions of alternative protein regions without first individually establishing the existence of isoform transcripts at the protein level. Biochemically, it is clear that even changing a small number of amino acids can be sufficient to critically alter protein function, with single-residue point mutations causally linked to diverse human diseases from cystic fibrosis to hypertrophic cardiomyopathy. Although the MYBPC3 isoform differs from the canonical by only in 10 of the 1,274 residues in the protein sequence, it is located at a crucial phosphorylation region, which suggests it may impact the functional regulation of the protein.

In summary, we describe a proteotranscriptomics approach to create precise databases for protein isoform analysis. We suggest that with continued refinement predicting translated transcripts, the approach demonstrated here will help elucidate the role of protein isoforms in development and disease.

Materials and Methods

Public RNA-seq and mass spectrometry datasets RNA-seq datasets were retrieved from ENCODE at the following accessions: heart (ENCSR436QDU, ENCSR391VGU), liver (ENCSR226KML, ENCSR504QMK), lung (ENCSR425RGZ, ENCSR406SAW), pancreas (ENCSR671IYC, ENCSR586SYA), adrenal gland (ENCSR801MKV, ENCSR754WLW), transverse colon (ENCSR800WIY, ENCSR403SZN), ovary (ENCSR841ADZ, ENCSR042GYH), esophagus (ENCSR098BUF, ENCSR750ETS), testis (ENCSR029KNZ, ENCSR344MQK), and prostate (ENCSR495HDM, ENCSR701TST). RNA-seq data from at least two biological replicates from each tissue were used. All data were 101nt paired-end total RNA-seq generated on Illumina Hi-Seq 2500 and passed ENCODE quality control unless specified. RNA-seq read [.fastq] files were manually retrieved on 2017-11-12. Proteomic datasets were retrieved from ProteomeXchange/PRIDE [35] at the following accessions: "A draft map of the human proteome" (PXD000561)[20] generated on Thermo Orbitrap Velos and Orbitrap Elite mass spectrometers; and "Region and cell-type resolved quantitative proteomic map of the human heart and its application to atrial fibrillation" (PXD006675) [21] generated on a Thermo Q-Exactive HF mass spectrometer. Thermo [.raw] mass spectra files were manually retrieved on 2018-03-15.

Custom sequence database generation To align the retrieved RNA-seq data, we used STAR v.2.5.0a [36, 37] on a Linux 4.10.0-32-generic Ubuntu x86_64 workstation. We mapped .fastq sequences to Ensembl GRCh38.89 STAR indexed genomes and GTF annotations with the following options (–sjdbOverhang 100, –outSAMtype BAM SortedByCoordinate). To extract splice junctions from the mapped and compare splice levels across biological replicates, we used rMATS-Turbo v.0.1 [38] on the mapped bam files with the following options (–readLength 101 –anchorLength 1).

To generate an accurate protein sequence database requires finding the set of isoforms that are as close as possible to that which actually exists in a particular sample at a level detectable by the mass spectrometry experimental setup, without relying on prior mass

spectrometry data. Here we use public RNA-seq reads from matching samples, aligned to the reference human genome to denote a subset of detectable isoform transcripts present. Except in cases where the proteins are imported from another tissue, it may be safely assumed that the existing protein isoform is a subset of the transcript isoform, because the majority of mass spectrometry data sets remain less sensitive than RNA-seq and also due to the fact that some isoform transcripts may not lead to translatable products. Accordingly, we filter first by transcript level of an isoform i in a tissue k so that transcript level t_i^k is above a defined threshold $t_i^k >= \theta$, which is adjusted according to the specific RNA-seq data sets used. In addition we assume that the isoform is reliably observed across multiple runs. Here we employ the model implemented in rMATS and exclude significantly differential splice junctions at $P < 0.05$, but other similar methods may be employed.

Secondly, to ensure included isoform transcripts are likely translatable requires estimating the translation frame and filtering of isoform transcript sequences that encounter PTC. Previous studies have opted to forego filtering (e.g., three-frame translation) or attempted to estimate the true translation frame from RNA-seq reads, such as by choosing the frame with the highest PSM score [16]. Our approach here is to make the strong assumption that for the most part there ought to be one true translatable frame for all isoforms within the majority of portions of the gene barring some exceptions like truncation. This can be retrieved in the GTF file, by matching exon coordinates in all coding sequence (CDS) annotated lines in the Ensemble GTF file.

To carry out these two tasks, we developed a software tool jcast (junction-centric alternative splicing translator) written in Python v.3.6.1, which tabulates alternative splicing events from each tissue and filters out ineligible splice pairs by virtue of read count threshold or significant inter-sample differences. The script then automatically retrieves nucleotide sequences from each splice pair based on the recorded genomic coordinates using the Ensembl REST web application programming interaface (API), then identifies the appropriate translation frames, transcription start sites, and transcription end sites of each splice pair from the Ensembl GRCh38.89 annotation GTF file. The retrieved qualifying nucleotide sequences are further translated into amino acid sequences using the annotated phase and frame. For analysis purpose, we divide the translated peptides into four tiers. Tier 1 peptides are translated in-frame by the annotated translation frame in Ensembl GRCh38.89 GTF successfully without encountering a frameshift or PTC; tier 2 peptides are those translated without PTC using the annotated frames but for which one of the spliced pairs encountered a frameshift event; tier 3 peptides are translated without frameshift or PTC but using a different translation frame than that which the Python script read from Ensembl GTF using matching exon coordinates; tier 4 peptides do encountered PTC in one of the two splice pairs but are retained here for analysis purpose; the majority of translated peptides do not encounter PTC and in any case only one translation frame is used. Finally, all translated peptides are required to be stitchable back to SwissProt canonical sequences by a 10-amino-acid joint. "Orphan" peptides that are translated but not stitchable back to SwissProt are separately recorded to analyze script performance but are not included in the translated sequence database and are not included in the analysis presented here.

Proteomics data analysis Mass spectrometry raw spectrum files were converted to open-source [.mzML] formats using ProteoWiazrd msconvert v.3.0.11392 [39] with the following options (–filter "peakPicking vendor"). Database search against custom databases were performed using the SEQUEST algorithm implemented in Comet v.2017.01 rev.0 [40] with the following options (–peptide_mass_tolerance 10 –peptide_mass_unit 2 –

isotope_error 2 –allowed_missed_cleavage 2 –num_enzyme_termini 1 –fragment_bin_tol 0.02). Search data were filtered using Percolator [41] in the Crux v.3.0 Macintosh binary distribution [42] with the following options (–protein T –fido-empirical-protein-q T –decoy-prefix DECOY_).

Statistical analysis and data visualization Data analysis and visualization were performed in R v.3.4.4 (2018-03-15 release) x86_64-apple-darwin15.6.0 (64-bit) with Bioconductor v.3.6 [43], with the aid of the MSnBase v.2.4.2 [44] and protViz v.0.2.459 packages. String occurrence of identified peptide sequences in SwissProt and UniProt fasta databases with 0 or 1 mismatch tolerance were assessed using the BioStrings v.3.7.0 package. Circular ideograms were produced with the aid of Circos v.0.69.6 [45].

Acknowledgments

This work was supported in part by National Institutes of Health (NIH) research grants F32 HL139045 to E.L. and R01 GM117624, R01 HL141278, and The University of Colorado Consortium for Fibrosis Research and Translation Pilot Grant to M.P.L.

Supporting Information

Supplementary Data 1

Spectra of Novel Peptides (Dataset 1). Spectral and chromatographic evidence of identified novel peptides in 10 human tissues (Dataset 1). Accessible on figshare at `https://doi.org/10.6084/m9.figshare.6493247.v1`.

Supplementary Data 2

Spectra of Novel Peptides (Dataset 2). Spectral and chromatographic evidence of identified novel peptides in 19 cardiac regions (Dataset 2). Accessible on figshare at `https://doi.org/10.6084/m9.figshare.6493262.v1`.

References

1. R. Aebersold, J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia, Y. Ge, J. Gunawardena, R. C. Hendrickson, P. J. Hergenrother, C. G. Huber, A. R. Ivanov, O. N. Jensen, M. C. Jewett, N. L. Kelleher, L. L. Kiessling, N. J. Krogan, M. R. Larsen, J. A. Loo, R. R. Ogorzalek Loo, E. Lundberg, M. J. MacCoss, P. Mallick, V. K. Mootha, M. Mrksich, T. W. Muir, S. M. Patrie, J. J. Pesavento, S. J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schluter, S. Sechi, S. A. Slavoff, L. M. Smith, M. P. Snyder, P. M. Thomas, M. Uhlen, J. E. Van Eyk, M. Vidal, D. R. Walt, F. M. White, E. R. Williams, T. Wohlschlager, V. H. Wysocki, N. A. Yates, N. L. Young, and B. Zhang. How many human proteoforms are there? *Nat. Chem. Biol.*, 14(3):206–214, Feb 2018.

2. F. E. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.*, 18(7):437–451, 07 2017.

3. M. M. van den Hoogenhof, Y. M. Pinto, and E. E. Creemers. RNA Splicing: Regulation and Dysregulation in the Heart. *Circ. Res.*, 118(3):454–468, Feb 2016.

4. Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, Dec 2008.

5. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

6. M. L. Tress, F. Abascal, and A. Valencia. Most Alternative Isoforms Are Not Functionally Important. *Trends Biochem. Sci.*, 42(6):408–410, 06 2017.

7. M. L. Tress, F. Abascal, and A. Valencia. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem. Sci.*, 42(2):98–110, 02 2017.

8. B. J. Blencowe. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem. Sci.*, 42(6):407–408, 06 2017.

9. X. Wang, S. G. Codreanu, B. Wen, K. Li, M. C. Chambers, D. C. Liebler, and B. Zhang. Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. *Mol. Cell Proteomics*, 17(3):422–430, Mar 2018.

10. R. Tavares, N. de Miranda Scherer, B. A. Pauletti, E. Araujo, E. L. Folador, G. Espindola, C. G. Ferreira, A. F. Paes Leme, P. S. de Oliveira, and F. Passetti. SpliceProt: a protein sequence repository of predicted human splice variants. *Proteomics*, 14(2-3):181–185, Feb 2014.

11. F. Mo, X. Hong, F. Gao, L. Du, J. Wang, G. S. Omenn, and B. Lin. A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. *BMC Bioinformatics*, 9:537, Dec 2008.

12. K. A. Power, J. P. McRedmond, A. de Stefani, W. M. Gallagher, and P. O. Gaora. High-throughput proteomics detection of novel splice isoforms in human platelets. *PLoS ONE*, 4(3):e5001, 2009.

13. D. Fermin, B. B. Allen, T. W. Blackwell, R. Menon, M. Adamski, Y. Xu, P. Ulintz, G. S. Omenn, and D. J. States. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.*, 7(4): R35, 2006.

14. J. A. Alfaro, A. Sinha, T. Kislinger, and P. C. Boutros. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nat. Methods*, 11(11):1107–1113, Nov 2014.

15. G. M. Sheynkman, M. R. Shortreed, B. L. Frey, and L. M. Smith. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol. Cell Proteomics*, 12(8):2341–2353, Aug 2013.

16. F. Zickmann and B. Y. Renard. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics*, 31(12):i106–115, Jun 2015.

17. K. Ning and A. I. Nesvizhskii. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics*, 11 Suppl 11:S14, Dec 2010.

18. A. Koch, D. Gawron, S. Steyaert, E. Ndah, J. Crappe, S. De Keulenaer, E. De Meester, M. Ma, B. Shen, K. Gevaert, W. Van Criekinge, P. Van Damme, and G. Menschaert. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*, 14(23-24):2688–2698, Dec 2014.

19. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 46(5):2699, Mar 2018.

20. M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A draft map of the human proteome. *Nature*, 509 (7502):575–581, May 2014.

21. S. Doll, M. Dressen, P. E. Geyer, D. N. Itzhak, C. Braun, S. A. Doppler, F. Meier, M. A. Deutsch, H. Lahm, R. Lange, M. Krane, and M. Mann. Region and cell-type resolved quantitative proteomic map of the human heart. *Nat Commun*, 8(1): 1469, Nov 2017.

22. O. V. Krokhin, R. Craig, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, and J. A. Wilkins. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell Proteomics*, 3(9):908–919, Sep 2004.

23. W. Guo, S. Schafer, M. L. Greaser, M. H. Radke, M. Liss, T. Govindarajan, H. Maatz, H. Schulz, S. Li, A. M. Parrish, V. Dauksaite, P. Vakeel, S. Klaassen, B. Gerull, L. Thierfelder, V. Regitz-Zagrosek, T. A. Hacker, K. W. Saupe, G. W. Dec, P. T. Ellinor, C. A. MacRae, B. Spallek, R. Fischer, A. Perrot, C. Ozcelik, K. Saar, N. Hubner, and M. Gotthardt. RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. *Nat. Med.*, 18(5):766–773, May 2012.

24. G. M. Boratyn, C. Camacho, P. S. Cooper, G. Coulouris, A. Fong, N. Ma, T. L. Madden, W. T. Matten, S. D. McGinnis, Y. Merezhuk, Y. Raytselis, E. W. Sayers, T. Tao, J. Ye, and I. Zaretskaya. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, 41(Web Server issue):29–33, Jul 2013.

25. M. M. Babu. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.*, 44(5): 1185–1200, 10 2016.

26. P. C. Rosas, Y. Liu, M. I. Abdalla, C. M. Thomas, D. T. Kidwell, G. F. Dusio, D. Mukhopadhyay, R. Kumar, K. M. Baker, B. M. Mitchell, P. A. Powers, D. P. Fitzsimons, B. G. Patel, C. M. Warren, R. J. Solaro, R. L. Moss, and C. W. Tong.

Phosphorylation of cardiac Myosin-binding protein-C is a critical mediator of diastolic function. *Circ Heart Fail*, 8(3):582–594, May 2015.

27. J. Griss, Y. Perez-Riverol, S. Lewis, D. L. Tabb, J. A. Dianes, N. Del-Toro, M. Rurik, M. W. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, and J. A. Vizcaino. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods*, 13(8):651–656, Aug 2016.

28. X. Yang, J. Coulombe-Huntington, S. Kang, G. M. Sheynkman, T. Hao, A. Richardson, S. Sun, F. Yang, Y. A. Shen, R. R. Murray, K. Spirohn, B. E. Begg, M. Duran-Frigola, A. MacWilliams, S. J. Pevzner, Q. Zhong, S. A. Trigg, S. Tam, L. Ghamsari, N. Sahni, S. Yi, M. D. Rodriguez, D. Balcha, G. Tan, M. Costanzo, B. Andrews, C. Boone, X. J. Zhou, K. Salehi-Ashtiani, B. Charloteaux, A. A. Chen, M. A. Calderwood, P. Aloy, F. P. Roth, D. E. Hill, L. M. Iakoucheva, Y. Xia, and M. Vidal. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, 164(4):805–817, Feb 2016.

29. J. D. Ellis, M. Barrios-Rodiles, R. Colak, M. Irimia, T. Kim, J. A. Calarco, X. Wang, Q. Pan, D. O'Hanlon, P. M. Kim, J. L. Wrana, and B. J. Blencowe. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell*, 46(6):884–892, Jun 2012.

30. Y. Liu, M. Gonzalez-Porta, S. Santos, A. Brazma, J. C. Marioni, R. Aebersold, A. R. Venkitaraman, and V. O. Wickramasinghe. Impact of Alternative Splicing on the Human Proteome. *Cell Rep*, 20(5):1229–1241, Aug 2017.

31. R. J. Weatheritt, T. Sterne-Weiler, and B. J. Blencowe. The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.*, 23(12):1117–1123, Dec 2016.

32. Y. Hao, R. Colak, J. Teyra, C. Corbi-Verge, A. Ignatchenko, H. Hahne, M. Wilhelm, B. Kuster, P. Braun, D. Kaida, T. Kislinger, and P. M. Kim. Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Proteins. *Cell Rep*, 12(2):183–189, Jul 2015.

33. L. Lin, P. Jiang, J. W. Park, J. Wang, Z. X. Lu, M. P. Lam, P. Ping, and Y. Xing. The contribution of Alu exons to the human proteome. *Genome Biol.*, 17:15, Jan 2016.

34. S. R. Ramakrishnan, C. Vogel, J. T. Prince, Z. Li, L. O. Penalva, M. Myers, E. M. Marcotte, D. P. Miranker, and R. Wang. Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics*, 25(11): 1397–1403, Jun 2009.

35. E. W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, D. S. Campbell, M. Bernal-Llinares, S. Okuda, S. Kawano, R. L. Moritz, J. J. Carver, M. Wang, Y. Ishihama, N. Bandeira, H. Hermjakob, and J. A. Vizcaino. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.*, 45(D1):D1100–D1106, Jan 2017.

36. S. Ballouz, A. Dobin, T. R. Gingeras, and J. Gillis. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Res.*, 46(10): 5125–5138, Jun 2018.

37. A. Dobin and T. R. Gingeras. Optimizing RNA-Seq Mapping with STAR. *Methods Mol. Biol.*, 1415:245–262, 2016.

38. S. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.*, 111(51):E5593–5601, Dec 2014.

39. R. Adusumilli and P. Mallick. Data Conversion with ProteoWizard msConvert. *Methods Mol. Biol.*, 1550:339–368, 2017.

40. J. K. Eng, M. R. Hoopmann, T. A. Jahan, J. D. Egertson, W. S. Noble, and M. J. MacCoss. A deeper look into Comet–implementation and features. *J. Am. Soc. Mass Spectrom.*, 26(11):1865–1874, Nov 2015.

41. M. The, M. J. MacCoss, W. S. Noble, and L. Kall. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.*, 27(11):1719–1727, Nov 2016.

42. S. McIlwain, K. Tamura, A. Kertesz-Farkas, C. E. Grant, B. Diament, B. Frewen, J. J. Howbert, M. R. Hoopmann, L. Kall, J. K. Eng, M. J. MacCoss, and W. S. Noble. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.*, 13(10):4488–4491, Oct 2014.

43. W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole?, H. Pages, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, 12(2):115–121, Feb 2015.

44. L. Gatto and K. S. Lilley. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2):288–289, Jan 2012.

45. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res.*, 19(9):1639–1645, Sep 2009.