Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing

Laura M. Agosto,[1,2,3] Matthew R. Gazzara,[2,4] Caleb M. Radens,[2,5] Simone Sidoli,[2,3] Josue Baeza,[2,3] Benjamin A. Garcia,[2,3] and Kristen W. Lynch[2]

[1] *Biochemistry and Molecular Biophysics Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA;* [2] *Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA;* [3] *Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA;* [4] *Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA;* [5] *Genetics and Epigenetics, Cell & Molecular Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

Alternative pre-mRNA splicing has long been proposed to contribute greatly to proteome complexity. However, the extent to which mature mRNA isoforms are successfully translated into protein remains controversial. Here, we used high-throughput RNA sequencing and mass spectrometry (MS)–based proteomics to better evaluate the translation of alternatively spliced mRNAs. To increase proteome coverage and improve protein quantitation, we optimized cell fractionation and sample processing steps at both the protein and peptide level. Furthermore, we generated a custom peptide database trained on analysis of RNA-seq data with MAJIQ, an algorithm optimized to detect and quantify differential and unannotated splice junction usage. We matched tandem mass spectra acquired by data-dependent acquisition (DDA) against our custom RNA-seq based database, as well as SWISS-PROT and RefSeq databases to improve identification of splicing-derived proteoforms by 28% compared with use of the SWISS-PROT database alone. Altogether, we identified peptide evidence for 554 alternate proteoforms corresponding to 274 genes. Our increased depth and detection of proteins also allowed us to track changes in the transcriptome and proteome induced by T-cell stimulation, as well as fluctuations in protein subcellular localization. In sum, our data here confirm that use of generic databases in proteomic studies underestimates the number of spliced mRNA isoforms that are translated into protein and provides a workflow that improves isoform detection in large-scale proteomic experiments.

[Supplemental material is available for this article.]

Eukaryotic proteome diversity arises via multiple mechanisms, including alternative premessenger RNA (pre-mRNA) splicing. Alternative splicing (AS) is a highly regulated process by which a single gene may code for multiple proteins through differential inclusion of alternative exons in the mature mRNA sequence. Transcriptome profiling across tissues has shown that >90% of multiexon genes undergo AS in eukaryotes and that ~80% of these events occur within the protein-coding region of transcripts (Pan et al. 2008; Wang et al. 2008). The splicing pattern for any given gene often differs across cell types and/or tissues, as well as in response to environmental cell stimuli (Cieply and Carstens 2015). Therefore, AS has been proposed to play an important role in shaping protein expression in a condition-specific manner. Studies of individual genes have provided some concrete examples of how AS impacts protein function, such as reduced kinase activity by altering a kinase docking site (e.g., *MAP2K7*) (Martinez et al. 2015), hindered immunotherapy via a truncated extracellular domain of a cell surface signaling molecule (e.g., *CD19*) (Sotillo et al. 2015), and changes in protein subcellular localization that consequently affect enzymatic activity (e.g., *EHMT2*) (Fiszbein et al. 2016). However, global analyses aiming to detect protein isoforms (“proteoforms”) of alternatively spiced mRNAs have been limited. Furthermore, no global studies looking at proteoform localization have been reported.

Recently, several groups have combined RNA sequencing with other sensitive and high-throughput techniques to determine how posttranscriptional processes impact the proteome (Sheynkman et al. 2013; Sterne-Weiler et al. 2013; Floor and Doudna 2016; Weatheritt et al. 2016; Jeong et al. 2018). Overall, these approaches and developments have been valuable for describing the role of AS in modulating proteome complexity, as well as understanding some of the disconnect between transcript and protein abundance. For example, a study using ribosome profiling determined that at least 75% of transcripts with alternative cassette exons are engaged by ribosomes (Weatheritt et al. 2016). However, these studies lack direct evidence showing that alternatively spliced mRNAs are translated into detectable proteins within a cell.

Corresponding authors: klync@pennmedicine.upenn.edu, bgarci@pennmedicine.upenn.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.248435.119.

The most widely used technique to study proteomes is nanoliquid chromatography (LC) coupled with tandem mass spectrometry (nLC-MS/MS), because of its sensitivity, robustness, and throughput. Yet, one of the biggest hurdles for nLC-MS/MS analysis of proteoforms has been increasing proteome coverage to detect low abundance peptides, thus proteins. Depth of coverage is particularly critical for proteoform analysis, as discrimination between proteoforms requires the detection of peptides translated across exon junctions (exon junction peptides [EJPs]) and/or alternative exon peptides (AEPs) that map uniquely to each variant. Additionally, the composition of the database used for searching MS data can impact which and how many isoforms one can confidently identify. In contrast to the findings of Weatheritt et al. (2016), a different research group observed that genes with a dominant splice variant also express a dominant protein isoform by nLC-MS/MS (Ezkurdia et al. 2015). This correlation does not rule out there being more than one protein product per gene, as they were only able to identify main proteoforms for ∼32% of annotated protein coding genes with multiple splice isoforms. The main proteoforms were determined by "counting the total number of peptides that mapped to each splice isoform annotated for a gene," which introduces a bias toward longer protein sequences and does not take into account normalized relative protein abundance values (Ezkurdia et al. 2015). Here, we seek to combine improvements in both transcriptomic and proteomic analysis to more comprehensively assess if differential inclusion of alternative exons is reflected in the proteome.

Results

Alternate exon junction usage and changes in splicing patterns upon T-cell stimulation

We have previously identified hundreds of splicing events that are regulated upon phorbol 12-myristate 13-acetate (PMA) stimulation of the JSL1 Jurkat T-cell line (Rothrock et al. 2003; Ip et al. 2007; Martinez et al. 2012, 2015; Gazzara et al. 2017). The homogeneity of these cells and reproducibility of the splicing patterns under both unstimulated and stimulated conditions (Ip et al. 2007; Martinez et al. 2015) make them an ideal system to study proteoform expression regulated by AS. To identify mRNA isoforms that are most likely to be abundantly represented in the proteome of JSL1 cells, we reanalyzed our previously published RNA-seq data from unstimulated and PMA-stimulated JSL1 cells (Gazzara et al., 2017) with the MAJIQ algorithm to identify local splice variations (LSVs) in which two or more exon junctions were highly used (20%–80% of the reads) in either condition. We used MAJIQ for these studies as it is optimized to detect and quantify both annotated and novel isoforms, including those derived from complex splicing events, intron retention, and novel ends of transcripts (Vaquero-Garcia et al. 2016). Therefore, this algorithm allows us to gain a comprehensive and quantitative view of the transcriptome. The MAJIQ analysis led to the identification of 25,199 "alternative junctions" (AJs), corresponding to 12,435 splicing events in 5578 genes (Fig. 1A). Most of these AJs (∼81%) were detected at a similar level in both the unstimulated and stimulated JSL1 cells, but ∼20% of AJs are used at least 20% more often in one cellular condition relative to the other (Fig. 1B). Examples of different patterns of AJ usage are shown in Supplemental Figure S1.

To address how much the above splicing variability impacts the proteome, we first looked at the distribution of the AJs across transcript regions and found that 73% of the AJs impact splicing events within the coding sequence (CDS) of transcripts, whereas 13% affect the 5′ UTR and 2% the 3′ UTR (Fig. 1C). This observation is in close agreement with previously reported effects of splicing on protein sequences (Wang et al. 2008). Of note, analysis of all junctions (including those in >80% or <20% of transcripts) reveals that 91% fall in the CDS, suggesting that AJs may be somewhat selected against within the CDS. In sum, this analysis identified 3929 genes that express at least two readily detectable mRNA isoforms in the JSL1 cells that are predicted to impact the proteome, with ∼20% of these mRNA isoforms changing in

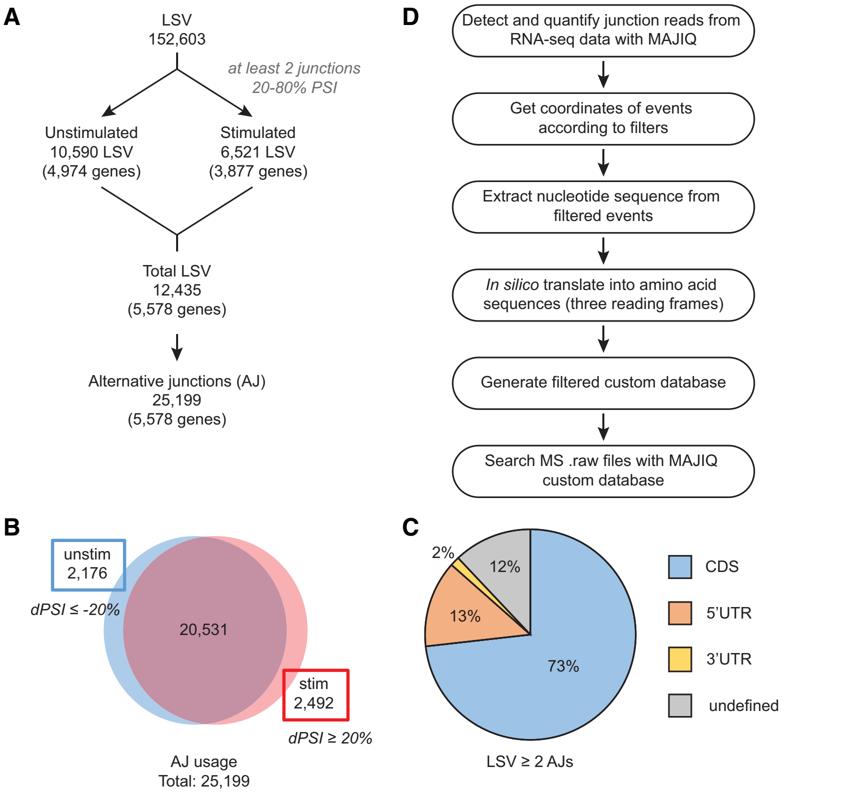


Figure 1. Identification of local splice variations (LSVs) and alternative junctions (AJs). (*A*) Workflow for identifying significantly used AJs in mRNA by identifying LSVs containing two or more exon junctions with support from 20% to 80% of the reads in either condition in previously described RNA-seq data. (*B*) Distribution of AJ usage in unstimulated and stimulated conditions. A shift in AJ usage is defined as the difference in percent spliced in (ΔPSI), set to 20%, between unstimulated and stimulated T cells. (*C*) Distribution of LSVs with two or more AJs throughout transcript regions, namely, the coding sequence (CDS) and untranslated regions (UTRs). (*D*) Workflow for generating MAJIQ custom protein database using AJs from *A*.

abundance upon stimulation of JSL1 cells. Finally, to use the MAJIQ analysis to guide our proteomics analysis, we used the genomic coordinates of these CDS-relevant AJs to extract the nucleotide sequences of the flanking exons and generate a custom database for MS data search (Fig. 1D; Supplemental Fig. S2). For simplicity, this custom database uses only AJs relating to cassette exons (8160 AJs total in 2607 genes), as this type of splicing event is the most common, and the most likely to generate detectable changes to the protein sequence.

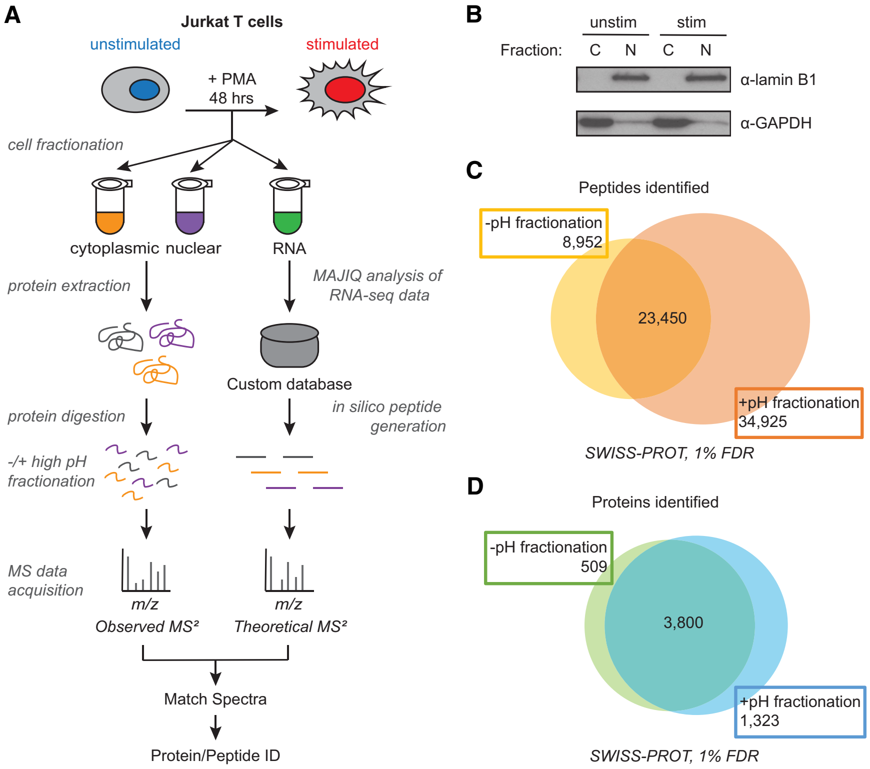


Figure 2. Increased protein and peptide identification achieved by high pH peptide fractionation. (*A*) Sample processing workflow used for integration of mass spectrometry (MS)–based proteomics and RNA-seq data. Jurkat T cells (JSL1) were stimulated for 48 h with phorbol 12-myristate 13-acetate (PMA) before protein or RNA extraction. Cells were harvested and fractionated into cytoplasmic and nuclear subfractions for MS analysis. Additionally, a second fractionation step was introduced after protein digestion to decrease sample complexity (high pH peptide fractionation). MS data were acquired via data-dependent and data-independent acquisition (DDA and DIA). *Right* path in workflow is described in Figure 1. (*B*) Validation of subcellular fractionation by western blot using antibodies for GAPDH and lamin B1. (*C*) Number of peptides identified −/+ high pH peptide fractionation at 1% false-discovery rate (FDR). (*D*) Number of proteins identified −/+ high pH peptide fractionation.

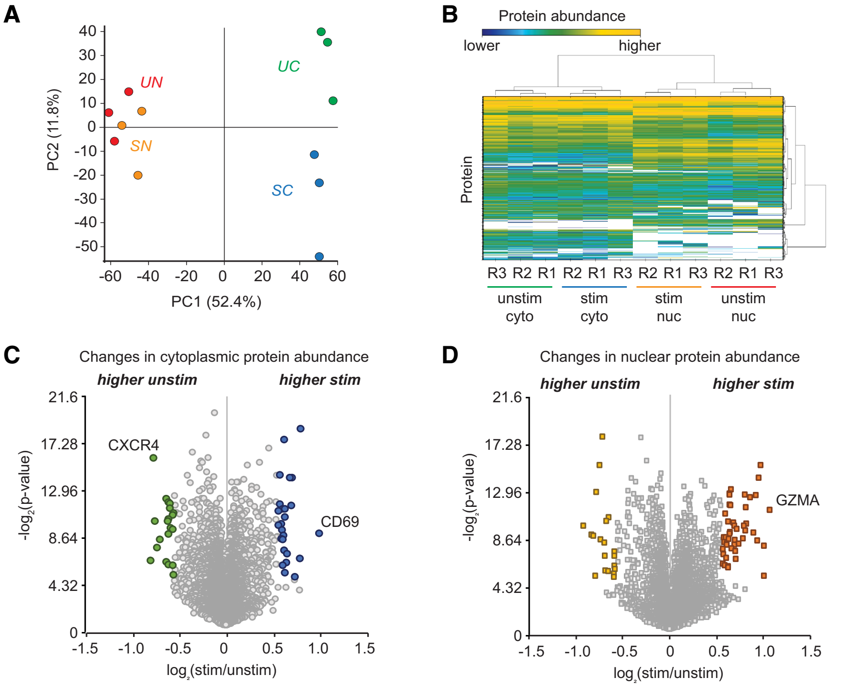
Increase in proteome coverage

Having identified the peptide sequences that are possible to derive from AS in the JSL1 cells, we next turned our attention to optimizing the MS detection of protein isoforms. When using RNA-seq data to analyze AS changes in the transcriptome, it is crucial to sequence samples in great depth so that lower abundance transcripts and junctions are readily detected with higher confidence. Likewise, to achieve identification of low abundance proteins, we need to increase proteome coverage by expanding our ability to separate and detect peptides, which ultimately increases the probability of detecting AEPs and EJPs (see Supplemental Fig. S2). To achieve deep, comprehensive proteome coverage, we used a workflow with multiple fractionation steps designed to decrease sample complexity and increase peptide and protein identification (Fig. 2A, left). Our first step was to fractionate cells into cytoplasmic and nuclear fractions, which we confirmed by western blot analysis (Fig. 2B). The second fractionation step consisted of separating the digested peptides from each compartment by reverse phase high pH fractionation (Supplemental Table S1). We illustrate the peptide fractionation step by overlaying the total ion chromatogram (MS data-dependent acquisition [DDA] mode) of a sample's fractions (Supplemental Fig. S3). Because we fractionated peptides based on hydrophobicity, we can see a shift in the peptide retention time when we separate the fractionated peptide mixtures by reverse-phase liquid chromatography (RPLC). The later fractions contain more hydrophobic peptides, thus elute later in the gradient.

Using the search method described in the section below, the high pH peptide fractionation step resulted in confident identification of 34,925 additional unique peptides compared with samples that were only fractionated by subcellular compartment (Fig. 2C), corresponding to 1323 unique proteins (Fig. 2D), both with a 1% false-discovery rate (FDR). The 8952 unique peptides identified in the –pH peptide fractionation samples can be explained by the analysis of fragment ions with different MS instruments and mass analyzers, which provide different limits of detection and MS/MS acquisition speeds that, although complementary and consistent, provide slightly different overlapping proteomes. Additionally, we used different strategies to remove detergents from our samples after the cell fractionation step (acetone precipitation vs. S-traps; see Methods).

Changes in protein abundance upon T-cell stimulation

After confirming that pH fractionation before nLC-MS/MS analysis greatly increased our ability to identify and quantify peptides, we wanted to look at the proteomic changes occurring upon T-cell stimulation, namely, changes in protein abundance. For the initial peptide analysis, we used Proteome Discoverer, with the Sequest HT search engine, to search the raw data (MS-DDA) and obtain peptide and protein identification with a 1% FDR cutoff applied with Percolator (Supplemental Table S2; Käll et al. 2007; Brosch et al. 2009). Of note, each database search was performed independently, and abundance values were manually normalized against each search. Principal component analysis (PCA) confirmed that the three biological replicates of each condition cluster together (Fig. 3A), with sample variability attributable to whether the samples are nuclear (N) or cytoplasmic (C; PC1, 52.4%) and unstimulated (U) or stimulated (S; PC2, 11.8%). For an unbiased analysis of peptide abundances, we also collected the MS^2 spectra in a data-independent acquisition (DIA) mode. The biggest difference



E

	Enriched gene set	Size	NES	FDR q-value	Rank
unstimulated	Meiotic chromosome segregation	15	2.18	0.044	475
	Anaphase promoting complex dependent catabolic process	66	2.16	0.044	584
	Sister chromatid segregation	112	1.99	0.087	598
	Kinesin complex	25	1.98	0.077	124
	Multi-organism localization	51	1.95	0.084	602
stimulated	Tumor necrosis factor mediated signaling pathway	52	-1.90	0.059	821
	Response to osmotic stress	20	-1.90	0.044	437
	Antigen binding	28	-1.90	0.044	760
	Positive regulation of leukocyte proliferation	39	-2.00	0.044	593
	Cytokine mediated signaling pathway	109	-2.00	0.044	848

F

	Enriched gene set	Size	NES	FDR q-value	Rank
unstimulated	Chromatin condensation	32	2.22	0.049	237
	Condensed nuclear chromosome	35	2.15	0.049	172
	Negative regulation of gene expression epigenetic	47	2.12	0.049	237
	DNA replication	124	2.02	0.054	650
	Mitotic spindle organization	39	2.02	0.053	134
stimulated	Pyruvate metabolic process	35	-1.68	1	862
	Regulation of IFNG production	17	-1.68	1	958
	Antigen processing & presentation of exogenous peptide antigen via MHC Class I	51	-1.68	1	1248
	Regulation of αβ T cell differentiation	15	-1.70	1	864
	Actin cytoskeleton reorganization	26	-1.74	1	793

Figure 3. Proteomic analysis of unstimulated and stimulated Jurkat T cells. (*A*) Principal component analysis (PCA) of three biological replicates (i.e., independent 48-h PMA stimulations and cell fractionations). Data were acquired via MS-DDA and analyzed using Proteome Discoverer 2.2 software (SWISS-PROT database, 1% FDR). (*B*) Heatmap and clustering of protein abundance as quantified by MS-DIA. Data were analyzed and clustered using Spectronaut software (SWISS-PROT database, 1% FDR). (*C*,*D*) Volcano plot of $\log_2$ ratio (stimulated/unstimulated) cytoplasmic (*C*) or nuclear (*D*) protein abundance (MS-DDA) versus two-tailed *t*-test adjusted *P*-value; significance set at 1.5-fold change (>|0.58| when $\log_2$ transformed) versus adjusted *P*-value <0.05 (>4.32 when $-\log_2$ transformed), respectively. Negative $\log_2$ ratio values represent proteins that are more abundant in unstimulated cells, whereas proteins with positive $\log_2$ ratio values are more highly expressed in stimulated cells. Decreased expression of CXCR4 and increased expression of CD69 and GZMA are markers of T-cell activation. (*E*,*F*) Top five most significant functional categories by gene set enrichment analysis (GSEA) of proteins significantly changing upon stimulation in the cytoplasm (*E*) or nucleus (*F*). Metrics: (1) size, number of genes that are categorized under the GO term; (2) NES, normalized enrichment score for the gene set after it has been normalized across analyzed gene sets; (3) FDR *q*-value, estimated probability that the normalized enrichment score represents a false positive finding; and (4) rank, position in the ranked list at which the maximum enrichment score occurred (proteins are ranked in order of positive to negative correlation to enrichment [signal/noise] in unstimulated cells).

between these acquisition modes is that DDA selects a predetermined number of ions for fragmentation (the most abundant ions from a given precursor scan), whereas in DIA, we select mass over charge (*m/z*) windows and fractionate peptides within that prespecified range (see Methods; Supplemental Table S3).

By confidently identifying peptides from DDA data analysis, we generated a T-cell–specific peptide spectral library to search the DIA data using Spectronaut software (Bruderer et al. 2015). Spectral libraries are a repository of fragmentation spectra and information on all the peptides identified by DDA. Although this type of DIA analysis approach allowed us to better quantify peptide abundance, the analysis is limited by the peptide information obtained from the DDA analysis (we cannot identify peptides absent from the spectral library). In agreement with the PCA analysis, a heatmap based on DIA protein abundance also shows coclustering of samples based on T-cell stimulation state (unstimulated, stimulated) and subcellular fraction (cytoplasmic, nuclear) of our three biological replicates (R1, R2, R3) (Fig. 3B). Furthermore, we show that sample clustering also occurs at the high pH peptide fractionation level across replicates (Supplemental Fig. S4).

After analyzing the DIA data, we exported the Spectronaut result file and manually processed it to determine stimulation-induced differences in protein abundance among our sample types. Protein abundance is presented in the form of volcano plots based on subcellular compartment, where the changes in protein abundance between our conditions are plotted in the *x*-axis ($\log_2$ normalized ratios of protein abundance in stimulated/unstimulated conditions), and the *y*-axis corresponds to the adjusted *P*-value calculated based on the three biological replicates ($-\log_2$ transformed) (Fig. 3C,D). *P*-values were obtained by performing a two-tailed homoscedastic *t*-test between unstimulated and stimulated conditions of either cytoplasmic or nuclear fractions. We set our significance cutoff as *P*-value ≤0.05 (or 4.32 when $-\log_2$ transformed). As expected, we observe that proteins known to be induced by T-cell activation such as CD69 and GZMA (Hess et al. 2004; Lieberman 2010; Singleton et al. 2011; Cibrián and Sánchez-Madrid 2017) are significantly up-regulated in stimulated cells (Fig. 3C, D), confirming the efficacy of our PMA stimulation. Gene set enrichment analysis (GSEA) (Subramanian et al. 2005) indicates the top five Gene Ontology (GO) enrichments are cell cycle– and epigenetic regulation–related categories in unstimulated cells, whereas stimulated cells show enrichment for activated immune signaling pathways (cytokines, TNF, IFN) and cell metabolism, proliferation, and differentiation (Fig. 3E,F). Globally, we also observed these trends after comparing unstimulated versus stimulated conditions (Supplemental Fig. S5). Supplemental Tables S4 and S5 contain the full list of identified and quantified proteins and peptides, respectively, in unstimulated versus stimulated subcellular compartments. We note that although this study is not focused on detecting changes in protein localization, we are able to quantify fluctuations in protein abundance across subcellular compartments (Supplemental Table S6).

Identification of EJPs

Having achieved deep detection and quantification of peptides in unstimulated and stimulated Jurkat T cells, we wanted to determine if we could detect splicing-derived proteoforms and if the MAJIQ-trained custom database improved detection of proteoforms. Thirty-nine percent of the AJs within the custom database developed in Figure 1 are predicted to generate peptide insertion or deletion, whereas another 34% are predicted to generate a frame shift (Fig. 4A). The majority of the remaining AJs alter the N (16%) or C (9%) terminus of the protein (Fig. 4A). In all of these cases, AS is predicted to generate AEPs and/or EJPs (Fig. 4B, "unique peptides"). In addition to searching for protein isoform matches using SWISS-PROT and the custom MAJIQ database, we also generated additional databases by translating all RefSeq-assembled transcripts in three reading frames. In total, we identified and quantified EJPs and AEPs that provide evidence for 554 proteoforms (which correspond to 274 genes) by combining the results of the SWISS-PROT, MAJIQ, and RefSeq databases (Fig. 4C; Supplemental Table S7). The majority of the EJPs and AEPs were peptides present in SWISS-PROT; however, 92 proteoforms (∼16% of total) were only identified with high confidence by the database trained on MAJIQ junction usage (Fig. 4C; Supplemental Table S7), whereas ∼10% were uniquely identified with RefSeq. The MAJIQ database also identified 140 proteoforms that are not captured by RefSeq. Therefore, as we predicted, the use of the RNA-seq to train a database increases our isoform discovery, although in no case was detection saturating. Possible reasons for the limited overlap in the proteoforms detected by the three databases are discussed below. Thus, we conclude that RNA isoform data obtained from a specific cell type carry unique information critical to predicting the resulting proteome.

In addition to observing distinct proteoforms, we can also quantify differential expression between alternate proteoforms. A clear example of this is the differential expression of alternate isoforms of the metabolic kinase *PKM*. *PKM* transcripts have been previously shown to be regulated by AS, specifically a mutually exclusive event between exons 9 and 10 (*PKM*-E9 and *PKM*-E10), thus encoding for PKM1 and PKM2 protein isoforms, respectively (Takenaka et al. 1991), which play opposing roles in metabolism (Christofk et al. 2008). Our MS results show that PKM2 (UniProt ID P14618) is more highly expressed than PKM1 (UniProt ID P14618-2) in both unstimulated and stimulated cells (Fig. 4D), consistent with up-regulation of pyruvate metabolic processing in stimulated cells (Fig. 3F). We used quantification of the AEPs (Supplemental Fig. S6) as a proxy for proteoform abundance, as the remainder of the peptides are shared between both proteoforms.

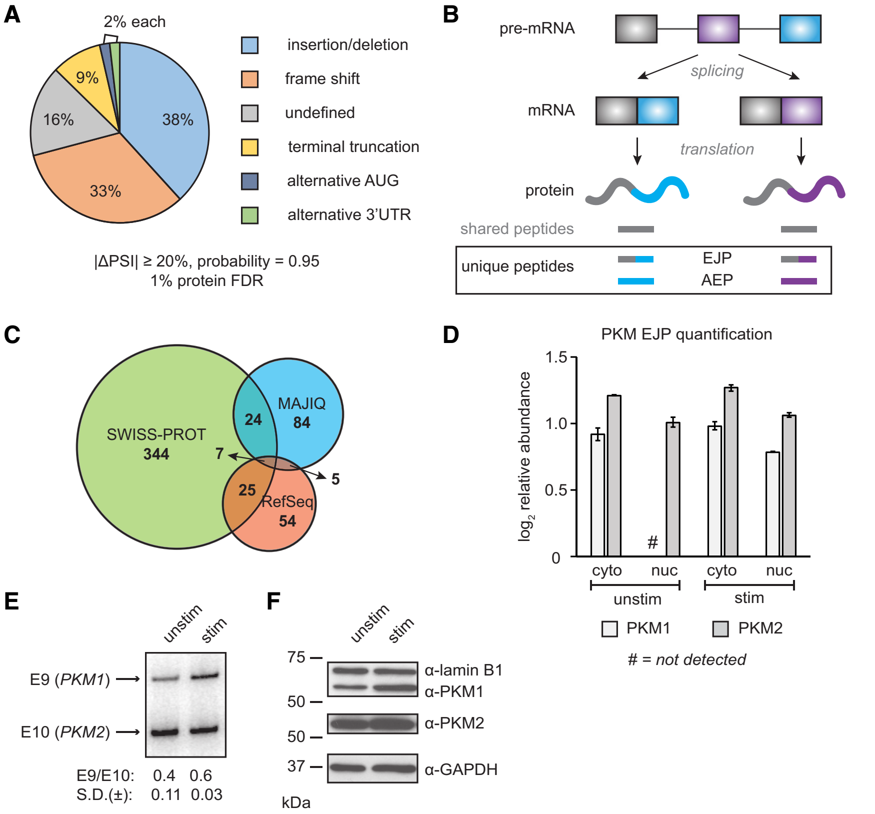


Figure 4. Identification of proteoforms by MS. (*A*) Pie chart of effect of alternative splicing (AS) events on protein sequence, based on the CDS cassette exons from Figure 1. (*B*) Schematic representation of AS leading to exon junction peptides (EJPs) and alternative exon peptides (AEPs), which distinguish AS proteoforms. (*C*) Venn diagram showing the number of peptides that report on alternate proteoforms as identified by SWISS-PROT (canonical + isoforms), our customized MAJIQ junction usage database, and/or RefSeq three-frame translation of transcripts (1% FDR). (*D*) Quantification of PKM EJP relative abundance by MS-DIA. (*E*) RT-PCR analysis of *PKM*-E9 (*upper* band) and -E10 (*lower* band) inclusion upon T-cell stimulation. Bands were detected with a Typhoon Phosphorimager and quantified with ImageQuant software ($n = 3$ per condition). (*F*) Western blot of PKM1 and PKM2 isoforms in unstimulated and stimulated conditions (5 µg whole-cell extract per lane). Lamin B1 and GAPDH were used as loading controls.

To determine if our detection of proteoforms mimicked the mRNA isoform profile, we performed RT-PCR with two forward primers that anneal to the unique exonic sequences of *PKM*-E9 or -E10 (Fig. 4E). Of note, *PKM*-E9 and -E10 are each 167 nt in length and code for a variable segment of 56 amino acids for either isoform. Inclusion of both exons would result in nonsense mediated decay (NMD) of the transcript, and both exons are not observed together in the transcriptome (Chen et al. 2012). Consistent with the MS result, we observed that the overall inclusion of *PKM*-E10 is higher than *PKM*-E9 under both unstimulated and stimulated conditions (Fig. 4E). Moreover, we are able to show differences in expression of PKM1/2 by western blot consistent with the MS and RT-PCR quantification (Fig. 4F). Therefore, at least for this case study, the relative abundance of the proteoform is reflective of the AS preference. We also note that the RT-PCR, western blot, and MS quantification of PKM all show an increase in the PKM1 isoform upon cell stimulation (Figs. 4D–F).

Differential expression of splicing-generated protein isoforms

To determine if other proteoforms showed differential expression in a stimulation-specific manner, we looked at proteoform expression by plotting cytoplasmic and nuclear $\log_2$ abundance ratios (stim/unstim) against $-\log_2$ *P*-values (Fig. 5A,B). We do observe several instances of a specific spliced proteoform that shows increased or decreased expression upon stimulation (Fig. 5A,B, colored circles). Functionally relevant examples of this are alternative proteoforms of the LEF1 transcription factor generated through alternative use of terminal coding exons 11 (E11) or 12 (E12). This AS event generates LQESASGTGPR from *LEF1*-E11 inclusion (E11-EJP) (Fig. 5A) and AATPGPLLEMEAC from *LEF1*-E11 skipping (E12-EJP) (Fig. 5B; Supplemental Fig. S7). These proteoforms are both detected in our data and are differentially distributed between subcellular compartments (Fig. 5C). In addition, although both increase upon stimulation, the proteoform containing the E12-EJP is more enhanced upon stimulation (Figs. 5A–C). Unfortunately, antibodies that efficiently discriminate between LEF1 isoforms are not available. However, as with *PKM*, the change in relative abundance in the E11-EJP and E12-EJP is reflective of the RNA isoform abundance, as RT-PCR detects an increase in the abundance of both the E11 included and skipped product; however, there is a preferential increase in E11 skipping in stimulated Jurkat T cells (Fig. 5D). In sum, these data show our ability to detect and quantify alternate proteoforms derived from AS and highlight that changes in splicing efficiency can impact the proteomic repertoire.

Discussion

We describe here a workflow that allows for increased identification and quantification of peptides and proteins in the interest of improving detection and quantification of peptides that uniquely distinguish protein isoforms. By using a T-cell stimulation system, we were able to quantify changes in alternative pre-mRNA splicing and proteoform expression. We show that LSV and alternative exon junction usage information is sufficient to train a peptide database in order to confidently identify proteoforms. By training a database with exon junction usage evidence for alternatively spliced cassette exons, we increased detection of splicing-derived proteoforms by >16% over those identified with a standard database (SWISS-PROT) and by 28% when combined with RefSeq. In addition to the improved database, we show that increased depth of peptide detection and mining of RNA-seq data improves the discovery of alternative proteoforms. Lastly, although this was not the focus of this study, we are also able to quantify fluctuations in protein abundance across subcellular compartments.

The approach we describe here to identify AS-generated proteoforms adds to a growing number of approaches in the field of proteogenomics (Nesvizhskii 2014). One alternative approach to proteomic discovery that has been described elsewhere (Ma et al. 2018) is the use of Cufflinks (Trapnell et al. 2010) to assemble transcripts from RNA-seq data that then are used to train a database of predicted peptides. This Cufflinks-based approach has proven fruitful in the discovery of microproteins (Ma et al. 2018), something that our approach did not specifically look for and was not designed to detect. In contrast, the MAJIQ-based approach we describe here is optimized to identify alternative proteoforms generated via AS by focusing on AEPs and EJPs. The use of MAJIQ is preferred over Cufflinks for discovery of AEP and EJPs, as the latter does not specifically distinguish alternative transcripts derived from AS as opposed to other gene regulatory mechanisms (Wang and Rio 2018). In addition, the MAJIQ-derived database, which focuses on local splicing variations in a gene, is smaller than a Cufflinks-derived peptide data base; thus, the MAJIQ-derived database involves reduced search space and computational time. In sum, there is unlikely to be a single best approach to the discovery of all of the nuances that may exist in the proteome. Rather, it is valuable to recognize the limits and strengths of each approach and train databases according to the goal of any particular study.

We do note that even though our custom database allowed for detection of AEPs and EJPs not in the SWISS-PROT database, there were other proteoforms only detected by SWISS-PROT. There are several reasons that may explain why we were not able to identify as many proteoforms with our custom junction usage database compared with SWISS-PROT. Filtering LSVs with a highly stringent junction usage cutoff that excluded some detectable AEP/EJP IDs and the limited length of the translated segment in

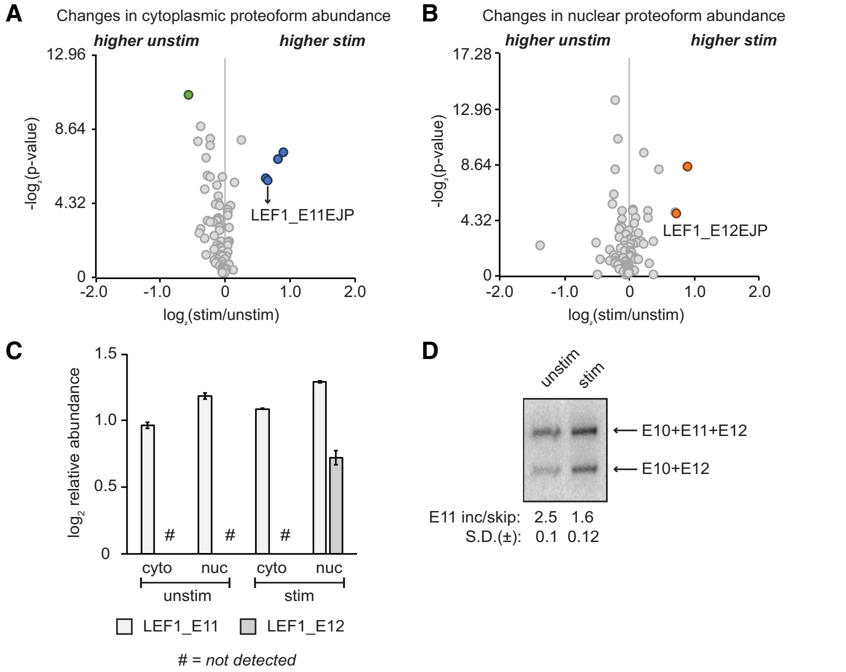


Figure 5. Analysis of differential proteoform expression. (*A*,*B*) Volcano plot of $\log_2$ ratio (stimulated/unstimulated) cytoplasmic (*A*) and nuclear (*B*) proteoform abundance (MS-DIA) versus $-\log_2$ *P*-value; significance was set at 1.5-fold change (>|0.58| when $\log_2$ transformed) versus *P*-value <0.05 (>4.32 when $-\log_2$ transformed), respectively. (*C*) Quantification of LEF1 proteoforms using EJPs generated by AS of *LEF1*-E11. (*D*) RT-PCR analysis of *LEF1*-E11 inclusion (*upper* band) and skipping (*lower* band). Bands were detected with a Typhoon Phosphorimager and quantified with ImageQuant software ($n = 3$).

the MAJIQ database mean many EJPs/AEPs identified by SWISS-PROT are not present in the MAJIQ custom database. Moreover, a limiting factor to our ability of detecting EJPs with either database is the trypsin digestion step, as it has been reported that a fraction of splice sites code for lysine and arginine residues (Wang et al. 2018). Therefore, using an enzyme that cuts at lysines and arginines may reduce the number of peptides that span exon junctions. In addition, not all EJPs or AEPs are detectable in normal nLC-MS/MS pipelines. For example, although our laboratory has previously studied the regulated skipping of exon 6 of *LEF1* (as opposed to exon 11 analyzed in Fig. 5), we could not quantify the proteomic impact of this splicing event as the peptide encoded as a result of exon 6 skipping cannot be analyzed using our nLC-MS/MS method because digestion with trypsin would generate a 55-amino-acid peptide for which the *m/z* is out of our scan range and parameters. We emphasize, however, that despite not detecting all possible proteoforms, our workflow does show the impact of AS on the proteome, as we were able to quantify the relative abundance of many proteoforms, including several with known biologic significance, such as PKM and LEF1.

Pyruvate kinase M1/2 (PKM) is a key enzyme for glycolysis, in which it catalyzes the final step of the glucose conversion pathway in order to produce pyruvate and ATP. *PKM* transcripts have been previously shown to be regulated by AS (Takenaka et al. 1991). PKM1 is constitutively active in healthy adult cells and promotes oxidative phosphorylation, whereas PKM2 is allosterically regulated and leads into aerobic glycolysis, thus driving tumorigenesis (Christofk et al. 2008). Several years ago, it was shown that the regulation of *PKM* splicing is performed by HNRNPA1/A2 and PTBP1 binding to intronic silencing sequences (ISSs) flanking *PKM*-E9, as well as A1 binding within *PKM*-E9 (Clower et al. 2010; David et al. 2010; Chen et al. 2012). Changes in expression levels of these RNA-binding proteins (RBPs) determine the splicing outcome and can revert lactate production in cancer cells (Christofk et al. 2008; Clower et al. 2010; David et al. 2010; Chen et al. 2012).

Similarly, AS has been shown to regulate the activity of Wnt/catenin beta 1–regulated transcription factor lymphocyte enhancer factor 1 (LEF1) (Waterman et al. 1991; Arce et al. 2006). *LEF1* has multiple exons regulated by AS that give rise to various proteoforms that differ in catenin beta 1 binding, the context-dependent repression domain (CDR), DNA-binding, and/or C-termini identity (Arce et al. 2006; Archbold et al. 2012; Nagalski et al. 2013). Our group has extensively studied the regulation of *LEF1* exon 6 (*LEF1*-E6; addition of 28 amino acids within the CDR) (Mallory et al. 2011; Ajith et al. 2016), which increases upon T-cell stimulation and promotes transcription of the T-cell receptor alpha chain (Mallory et al. 2011; Ajith et al. 2016). Additionally, regulation of exon 11 (*LEF1*-E11) alters the C terminus of the protein, as inclusion of E11 introduces an in-frame stop codon that ends the CDS region of the transcript, whereas skipping of E11 results in usage of a stop codon found in exon 12 (*LEF1*-E12). Our finding that the C terminus of LEF1 correlates with subcellular localization provides useful information to guide future studies regarding the functional impact of exon 11 included on LEF1 protein function.

In sum, our work here underscores that AS does contribute to the diversity and regulation of the proteome and is often underestimated owing to incomplete depth of both proteomic and transcriptomic data. In particular, we show that intentional acquisition and focused analysis of nLC-MS/MS data are required to observe the full impact of splicing on protein expression. We propose that a similar workflow will likewise be useful for interrogating the impact of AS on other physiologic systems.

Methods

Cell culture and stimulation

JSL1 cells were cultured in RPMI 1640 (Corning 10-040-CV) supplemented with 5% heat-inactivated fetal bovine serum (FBS; Gibco 16000-044), penicillin, and streptomycin (100 units/mL each) and were grown at 37°C in 5% CO_2. For stimulation, we set up three independent replicates by diluting cells to 3.5×10^5 cells/mL and treating with 20 ng/mL PMA (Sigma-Aldrich) for 48 h before cell harvest (Lynch and Weiss 2000). Unstimulated JSL1 cells were cultured in parallel for each replicate at a seeding concentration of 2×10^5 cells/mL.

Custom database generation

MAJIQ

RNA-seq data and differential splicing analysis by MAJIQ was previously published (GSE93594) (Gazzara et al. 2017). In brief, total RNA was isolated and poly(A) selected from unstimulated and PMA stimulated JSL1 cells using RNA-Bee (Tel-Test) as previously described (Smith and Lynch 2014). RNA-seq data were analyzed with MAJIQ (Vaquero-Garcia et al. 2016) to identify LSVs and quantify changes in exonic splicing patterns between our conditions, which we term difference in percent spliced in (ΔPSI). To generate the peptide custom database to search MS data, we filtered the MAJIQ output for LSVs that had two or more junctions each with 20%–80% of reads going to/coming from a specific junction, hereby termed "alternative junctions" (AJs). This allowed us to filter the data set for junctions that were highly used (by an arbitrary threshold) and increase the likelihood of detecting the isoforms at the protein level. Next, we extracted the flanking exonic sequences upstream of and downstream from a given LSV and translated them in silico using three forward reading frames (SeqinR R package) (Charif and Lobry 2007) to generate peptide sequences that benchmark splicing products (Supplemental Fig. S2; see also Supplemental Code). In the case of novel splice junctions identified by MAJIQ, we used 50 nt upstream of and downstream from the junction to build the sequences. The use of a 100-nt window around the splice junctions is based on the fact that our MS pipeline is optimized for peptides of six to 25 amino acids, thus predicting splice junction peptides of longer than about 30 amino acids provides limited additional usable information. Our database includes translated sequences up to the first stop codon found in a given reading frame, removing any peptides that are fewer than six amino acids long. In total, the MAJIQ database contains 60,894 sequence elements.

RefSeq

We downloaded the FASTA file for the GRCh37 annotated transcripts. This assembly was used as the best comparison for MAJIQ, as MAJIQ is based on the GRC37/hg19 genome assembly. Using GRCh38 (hg38) annotated transcripts, or indeed any other annotation, would not significantly affect the conclusions of this study because MAJIQ identifies splicing isoforms independent of any assembly (i.e., identified de novo isoforms) and thus is more complete than any specific annotation. Similarly to the MAJIQ extracted sequences, we translated the RefSeq transcripts in silico using three forward reading frames to generate peptide sequences. Our database includes translated sequences up to the first stop codon found in a given reading frame, removing any peptides that are less than six amino acids long. In total, the RefSeq database contains 150,004 sequence elements.

RT-PCR

Low-cycle reverse transcription (RT)–PCR analysis of AS events was performed as described previously (Lynch and Weiss 2000; Melton et al. 2007) using sequence-specific primers for individual genes. In brief, we set-up three independent RT-PCR reactions with 1 µg of RNA obtained from unstimulated or stimulated cells. Primer sequences and RT-PCR conditions are provided in Supplemental Table S8. Formamide buffer was used to run samples on 5% denaturing polyacrylamide gels (PAGEs). RT-PCR products were detected by densitometry using a Typhoon Phosphorimager (Amersham Biosciences). Product bands were quantified with ImageQuant software.

Cell fractionation and protein extraction

Unstimulated and PMA-stimulated JSL1 cells were harvested after 48 h of treatment, rinsed with PBS, resuspended in lysis buffer (20 mM HEPES at pH 7.9, 150 Mm NaCl, 0.5 Mm $MgCl_2$, 0.5% NP-40 alternative, 10% glycerol, 1 mM DTT, 1.2 mM AEBSF), and incubated on ice for 5 min. The lysates were centrifuged for 20 min at 4°C at maximum speed to pellet the nuclei (supernatant is the cytoplasmic fraction). To obtain the nuclear fraction, we resuspended the nuclei pellets in 50 mM ammonium bicarbonate (Sigma-Aldrich 11213-1KG-R) and sonicated with short intermittent pulses until the lysate was cleared. We determined protein concentration of each sample by Bradford assay (Bio-Rad 500-0006).

Western blot

For protein analysis, 5 µg of whole-cell protein extract (WCE) was loaded into 8% 37.5:1 bis-acrylamide SDS-PAGE gels. Antibodies for western blot were diluted in 5% BSA-TBST as follows: GAPDH (Abcam ab128915; 1:5000, 5 µg WCE), lamin B1 (Abcam ab133741; 1:5000, 5 µg WCE), PKM1 (Cell Signaling D30G6; 1:1000, 5 µg WCE), and PKM2 (Cell Signaling D78A4; 1:1000, 5 µg WCE).

Protein sample processing for MS

−pH peptide fractionation samples

For each sample, we aliquoted 50 µg of fractionated protein extract for further processing. Samples were denatured (10 mM DTT, 30 min, 55°C), and alkylated (25 mM iodoacetamide, 30 min, room temperature). To remove the detergent used for cytoplasmic fractionation, we used acetone (1:6, sample:acetone) to precipitate the proteins overnight at −20°C, centrifuged at 8000*g* for 10 min, and air-dried the pellets. We resuspended the samples in 50 mM ammonium bicarbonate and added trypsin (Promega, Fisher Scientific PRV5113) at a 1:33 ratio (trypsin:sample) for protein digestion overnight at 37°C. Samples were dried on a SpeedVac and resuspended in 0.1% TFA for stage-tipping.

+pH peptide fractionation samples

For each sample, we aliquoted 50 µg of fractionated protein extract for further processing. To remove salts and detergents from the samples, we used S-trap micro columns (ProtiFi C02-micro-40) and followed the previously recommended protocol (Zougman et al. 2014; HaileMariam et al. 2018). In brief, we added SDS to each sample for 5% final concentration and denatured (20 mM DTT, 10 min, 95°C) and alkylated the proteins (40 mM iodoacetamide, 30 min, room temperature). Next, we applied the samples to S-trap micro columns and used the protocol-indicated buffers for cleanup (Zougman et al. 2014). We used trypsin for overnight digestion at 47°C (1:30 ratio in 50 mM triethylammonium bicarbonate). After eluting the peptides from the S-trap, we dried the solutions on a SpeedVac and resuspended the samples in 0.1% TFA. For further sample fractionation at the peptide level, we applied the peptide mixture to a C18 Micro SpinColumn (Harvard 74-4601) and generated fractions by sequentially eluting peptides off the column with increasing concentrations of acetonitrile in 100 mM ammonium formate (pH 10; Honeywell Fluka 17843-50G) (Supplemental Table S1). Samples were collected in 1.7-mL microtubes, and fractions were paired as shown in Supplemental Table S1 (i.e., high pH fractions A and D were pulled together). The fractions to be combined were chosen this way to allow for peptides having different hydrophobic properties and to cut by half the MS run time. Samples were dried on a SpeedVac and resuspended in 0.1% TFA.

−/+ pH fractionation samples

Peptides were desalted before nLC-MS/MS analysis by using in-house-packed stage-tips assembled by sealing a disk of C18 material at the bottom of a P200 tip. Stage-tips were equilibrated with 100 µL of 0.1% TFA, applied with sample, and washed once with 100 µL 0.1% FA. Elution of peptides was performed by using 50 µL of 70% acetonitrile +0.1% FA (twice). Samples were dried on a SpeedVac and resuspended in 0.1% TFA for subsequent nLC-MS/MS analysis.

nLC-MS/MS

DDA (−pH peptide fractionation samples)

Samples were analyzed by using an EASY-nLC nano HPLC system (Thermo Fisher Scientific) coupled online with a Fusion Orbitrap Tribrid MS instrument (Thermo Fisher Scientific). The nLC was configured with a 75-µm ID × 20-cm Reprosil-Pur C18-AQ (3 µm; Dr. Maisch) reverse-phase capillary column packed in-house. The full HPLC method was 135 min long with a 125-min gradient as detailed: 1% to 28% solvent B (solvent A = 0.1% formic acid; solvent B = 0.1% formic acid in 100% acetonitrile) over 120 min and from 28% to 90% solvent B in 5 min at a flow-rate of 300 nL/min. Data were acquired using DDA and positive polarity modes, consisting on a full-scan MS spectrum (350–1200 *m/z*) performed in the Orbitrap at 120,000 resolution (cycle time = 3 sec), followed by higher energy collision dissociation (HCD) fragmentation of the precursor ions (two to six charge state, intensity threshold minimum = 2×10^4). HCD collision energy was set at 35. MS/MS scans detected on the ion trap (AGC target minimum = 1×10^4). Xcalibur software was used for data collection.

DDA (+pH peptide fractionation samples)

Samples were analyzed by using a Dionex UltiMate 3000 (Thermo Fisher Scientific) LC system coupled online with a Q Exactive HF-X instrument (Thermo Fisher Scientific). The LC was configured with a 75-µm ID by a 20-cm Reprosil-Pur C_{18}-AQ (3 µm; Dr. Maisch) reverse-phase capillary column packed in-house. The full HPLC method was 75 min long with a 67-min gradient as detailed: 5% to 20% solvent B (solvent A = 0.1% formic acid; solvent B = 0.1% formic acid, 80% acetonitrile) over 45 min, from 20% to 40% solvent B in 15 min, and from 40% to 85% solvent B in 7 min at a flow-rate of 300 nL/min. Data were acquired using DDA and positive polarity modes, consisting on a full-scan MS spectrum (300–1100 *m/z*) performed in the orbitrap at 120,000 resolution, followed by HCD fragmentation of the top 20 precursor ions (≥+2 charge state) and MS/MS scans detected on the Orbitrap with 30,000 resolution. HCD collision energy was stepped, with variable energies at 25.5, 27, and 30. Xcalibur software was used for data collection.

DIA (+pH peptide fractionation samples)

As for the DDA data acquisition, we used the Dionex UltiMate 3000 LC system coupled online with the same Q Exactive HF-X instrument. The same solvent gradient and reverse-phase C18 capillary column were used for sample injection. Data were acquired using DIA and positive polarity modes, consisting on a full-scan MS spectrum (350–1100 *m/z*) performed in the Orbitrap at 120,000 resolution, followed by HCD fragmentation (stepped energy) of precursor ions ≥+2 charge state within preselected overlapping 20 *m/z* isolation windows (Supplemental Table S3). MS/MS scans detected on the Orbitrap with 30,000 resolution. Xcalibur software was used for data collection.

Database search

DDA

Data were analyzed using the Sequest HT search engine in Proteome Discoverer 2.2 (Thermo Fisher Scientific), Percolator, and three databases: a full human proteome database (SWISS-PROT canonical plus isoforms, release 2018_05; 42,348 sequence elements), the MAJIQ LSV custom database, and the RefSeq assembled transcripts translated in three forward reading frames (both described above). We programmed a bioinformatics workflow to identify MS/MS spectra matching with protein sequences from the SWISS-PROT database for a global proteomics analysis (1% FDR). Our second database search used a MAJIQ custom database based on highly used exon junctions of cassette exons (Fig. 1A,B) with 1% FDR. Last, the third database search was performed with the RefSeq 3-frame transcript translation with 1% FDR. The parameters used for each database search are detailed in the Supplemental Table S2. The similarity between −pH and +pH peptide fractionation samples was determined by comparing the peptides and proteins identified by each of the two databases (Supplemental Tables S4, S5).

DIA

We imported the result file from PD 2.2 into Spectronaut (Biognosys) to create a custom spectral library for the subsequent DIA data analysis (Bruderer et al. 2015). After obtaining the spectral library, we searched the DIA raw files against the SWISS-PROT and MAJIQ junction usage databases. We used Spectronaut protein and peptide quantification values to determine changes in protein and peptide abundance upon T-cell stimulation.

For both DDA and DIA data, we performed normalization by dividing a given quantification value (peptide or protein) by the average of quantification values for that particular sample after $\log_2$ transformation. Additionally, we filtered out proteins/peptides that were identified in only one replicate and/or not quantified. To assess proteomic changes between our conditions, we imputed missing values at the protein quantification level using Excel. Significant changes were determined by performing a two-tailed homoscedastic *t*-test between unstimulated and stimulated conditions; our significance cutoff was *P*-value <0.05 (or 4.32 when $-\log_2$ transformed). Last, we calculated the adjusted *P*-value using the Benjamini and Hochberg (BH) correction in RStudio (RStudio Team 2015) to control for FDRs.

GSEA

GSEA was performed using the javaGSEA desktop application interface (Subramanian et al. 2005). Input data set files contained normalized $\log_2$ protein abundance values (from MS-DIA analysis) for each replicate/condition. We used the GO "c5.all.v6.2.symbols.gmt" gene set database with 1000 phenotype permutations. Enrichment statistic was set to "weighted" and metric for ranking genes "Signal2Noise."

Data access

All MS data generated in this study have been submitted to the EMBL PRIDE archive (https://www.ebi.ac.uk/pride/archive/) under accession number PXD012556.

Acknowledgments

We thank Caroline Davis and Donita Brady for PKM1 and PKM2 antibodies. This work was supported by a National Science Foundation predoctoral fellowship to L.M.A., National Institute of General Medical Sciences grant R35 GM118048 to K.W.L., and NIH grants R01 GM110174 (National Institute of General Medical Sciences) and R01 AI118891 (National Institute of Allergy and Infectious Diseases) to B.A.G.

Author contributions: L.M.A. cultured and stimulated JSL1 cells, designed and performed the proteomic sample processing workflow, and performed RT-PCR and western blotting to validate mRNA and proteoforms. M.R.G. extracted nucleotide sequences to build the MAJIQ custom databases. C.M.R. wrote the in-house R script to generate three-frame translation databases. J.B. and S.S. generated the MS data acquisition methods. L.M.A. analyzed proteomic data with the help of J.B., S.S., and C.M.R. L.M.A., K.W.L., and B.A.G. designed the experiments and wrote the manuscript.

References

Ajith S, Gazzara MR, Cole BS, Shankarling G, Martinez NM, Mallory MJ, Lynch KW. 2016. Position-dependent activity of CELF2 in the regulation of splicing and implications for signal-responsive regulation in T cells. *RNA Biol* **13:** 569–581. doi:10.1080/15476286.2016.1176663

Arce L, Yokoyama NN, Waterman ML. 2006. Diversity of LEF/TCF action in development and disease. *Oncogene* **25:** 7492–7504. doi:10.1038/sj.onc.1210056

Archbold HC, Yang YX, Chen L, Cadigan KM. 2012. How do they do Wnt they do?: regulation of transcription by the Wnt/β-catenin pathway. *Acta Physiol* **204:** 74–109. doi:10.1111/j.1748-1716.2011.02293.x

Brosch M, Yu L, Hubbard T, Choudhary J. 2009. Accurate and sensitive peptide identification with Mascot Percolator. *J Proteome Res* **8:** 3176–3181. doi:10.1021/pr800982s

Bruderer R, Bernhardt OM, Gandhi T, Miladinovic SM, Cheng L-Y, Messner S, Ehrenberger T, Zanotelli V, Butscheid Y, Escher C, et al. 2015. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen treated 3D liver microtissues. *Mol Cell Proteomics* **14:** 1400–1410. doi:10.1074/mcp.M114.044305

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution: molecules, networks, populations. Biological and medical physics, biomedical engineering* (ed. Bastolla U), pp. 207–232. Springer, Berlin. https://doi.org/10.1007/978-3-540-35306-5_10 [accessed September 11, 2019].

Chen M, David CJ, Manley JL. 2012. Concentration-dependent control of pyruvate kinase M mutually exclusive splicing by hnRNP proteins. *Nat Struct Mol Biol* **19:** 346–354. doi:10.1038/nsmb.2219

Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC. 2008. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452:** 230–233. doi:10.1038/nature06734

Cibrián D, Sánchez-Madrid F. 2017. CD69: from activation marker to metabolic gatekeeper. *Eur J Immunol* **47:** 946–953. doi:10.1002/eji.201646837

Cieply B, Carstens RP. 2015. Functional roles of alternative splicing factors in human disease. *Wiley Interdiscip Rev RNA* **6:** 311–326. doi:10.1002/wrna.1276

Clower CV, Chatterjee D, Wang Z, Cantley LC, Vander Heiden MG, Krainer AR. 2010. The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism. *Proc Natl Acad Sci* **107:** 1894–1899. doi:10.1073/pnas.0914845107

David CJ, Chen M, Assanah M, Canoll P, Manley JL. 2010. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature* **463:** 364–368. doi:10.1038/nature08697

Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. 2015. Most highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res* **14:** 1880–1887. doi:10.1021/pr501286b

Fiszbein A, Giono LE, Quaglino A, Berardino BG, Sigaut L, von Bilderling C, Schor IE, Steinberg JHE, Rossi M, Pietrasanta LI, et al. 2016. Alternative splicing of *G9a* regulates neuronal differentiation. *Cell Rep* **14:** 2797–2808. doi:10.1016/j.celrep.2016.02.063

Floor SN, Doudna JA. 2016. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5:** e10921. doi:10.7554/eLife.10921

Gazzara MR, Mallory MJ, Roytenberg R, Lindberg J, Jha A, Lynch KW, Barash Y. 2017. Ancient antagonism between CELF and RBFOX families tunes mRNA splicing outcomes. *Genome Res* **27:** 1360–1370. doi:10.1101/gr.220517.117

HaileMariam M, Eguez RV, Singh H, Bekele S, Ameni G, Pieper R, Yu Y. 2018. S-Trap, an ultrafast sample-preparation approach for shotgun proteomics. *J Proteome Res* **17:** 2917–2924. doi:10.1021/acs.jproteome.8b00505

Hess K, Yang Y, Golech S, Sharov A, Becker KG, Weng N-P. 2004. Kinetic assessment of general gene expression changes during human naive CD4$^+$ T cell activation. *Int Immunol* **16:** 1711–1721. doi:10.1093/intimm/dxh172

Ip JY, Tong A, Pan Q, Topp JD, Blencowe BJ, Lynch KW. 2007. Global analysis of alternative splicing during T-cell activation. *RNA* **13:** 563–572. doi:10.1261/rna.457207

Jeong S-K, Kim C-Y, Paik Y-K. 2018. ASV-ID, a proteogenomic workflow to predict candidate protein isoforms based on transcript evidence. *J Proteome Res* **17:** 4235–4242. doi:10.1021/acs.jproteome.8b00548.

Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4:** 923–925. doi:10.1038/nmeth1113

Lieberman J. 2010. Granzyme A activates another way to die. *Immunol Rev* **235:** 93–104. doi:10.1111/j.0105-2896.2010.00902.x

Lynch KW, Weiss A. 2000. A model system for activation-induced alternative splicing of CD45 pre-mRNA in T cells implicates protein kinase C and Ras. *Mol Cell Biol* **20:** 70–80. doi:10.1128/MCB.20.1.70-80.2000

Ma J, Saghatelian A, Shokhirev MN. 2018. The influence of transcript assembly on the proteogenomics discovery of microproteins. *PLoS One* **13:** e0194518. doi:10.1371/journal.pone.0194518

Mallory MJ, Jackson J, Weber B, Chi A, Heyd F, Lynch KW. 2011. Signal- and development-dependent alternative splicing of LEF1 in T cells is controlled by CELF2. *Mol Cell Biol* **31:** 2184–2195. doi:10.1128/MCB.05170-11

Martinez NM, Pan Q, Cole BS, Yarosh CA, Babcock GA, Heyd F, Zhu W, Ajith S, Blencowe BJ, Lynch KW. 2012. Alternative splicing networks regulated by signaling in human T cells. *RNA* **18:** 1029–1040. doi:10.1261/rna.032243.112

Martinez NM, Agosto L, Qiu J, Mallory MJ, Gazzara MR, Barash Y, Fu X, Lynch KW. 2015. Widespread JNK-dependent alternative splicing induces a positive feedback loop through CELF2-mediated regulation of MKK7 during T-cell activation. *Genes Dev* **29:** 2054–2066. doi:10.1101/gad.267245.115

Melton AA, Jackson J, Wang J, Lynch KW. 2007. Combinatorial control of signal-induced exon repression by hnRNP L and PSF. *Mol Cell Biol* **27:** 6972–6984. doi:10.1128/MCB.00419-07

Nagalski A, Irimia M, Szewczyk L, Ferran JL, Misztal K, Kuznicki J, Wisniewska MB. 2013. Postnatal isoform switch and protein localization of LEF1 and TCF7L2 transcription factors in cortical, thalamic, and mesencephalic regions of the adult mouse brain. *Brain Struct Funct* **218:** 1531–1549. doi:10.1007/s00429-012-0474-6

Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11:** 1114–1125. doi:10.1038/nmeth.3144

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415. doi:10.1038/ng.259

Rothrock C, Cannon B, Hahm B, Lynch KW. 2003. A conserved signal-responsive sequence mediates activation-induced alternative splicing of CD45. *Mol Cell* **12:** 1317–1324. doi:10.1016/S1097-2765(03)00434-9

RStudio Team. 2015. *RStudio: integrated development for R*. RStudio, Inc., Boston. http://www.rstudio.com/.

Sheynkman GM, Shortreed MR, Frey BL, Smith LM. 2013. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol Cell Proteomics* **12:** 2341–2353. doi:10.1074/mcp.O113.028142

Singleton KL, Gosh M, Dandekar RD, Au-Yeung BB, Ksionda O, Tybulewicz VLJ, Altman A, Fowell DJ, Wülfing C. 2011. Itk controls the spatiotemporal organization of T cell activation. *Sci Signal* **4:** ra66. doi:10.1126/scisignal.2001821

Smith SA, Lynch KW. 2014. Cell-based splicing of minigenes. In *Spliceosomal pre-mRNA splicing: methods in molecular biology* (ed. Hertel KJ), pp. 243–255. Humana Press, Totowa, NJ. http://dx.doi.org/10.1007/978-1-62703-980-2_18 [accessed March 22, 2017].

Sotillo E, Barrett DM, Black KL, Bagashev A, Oldridge D, Wu G, Sussman R, Lanauze C, Ruella M, Gazzara MR, et al. 2015. Convergence of acquired mutations and alternative splicing of *CD19* enables resistance to CART-19 immunotherapy. *Cancer Discov* **5:** 1282–1295. doi:10.1158/2159-8290.CD-15-1020

Sterne-Weiler T, Martinez-Nunez RT, Howard JM, Cvitovik I, Katzman S, Tariq MA, Pourmand N, Sanford JR. 2013. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res* **23:** 1615–1623. doi:10.1101/gr.148585.112

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550. doi:10.1073/pnas.0506580102

Takenaka M, Noguchi T, Sadahiro S, Hirai H, Yamada K, Matsuda T, Imai E, Tanaka T. 1991. Isolation and characterization of the human pyruvate kinase M gene. *Eur J Biochem* **198:** 101–106. doi:10.1111/j.1432-1033.1991.tb15991.x

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515. doi:10.1038/nbt.1621

Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5:** e11752. doi:10.7554/eLife.11752

Wang Q, Rio DC. 2018. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proc Natl Acad Sci* **115:** E8181–E8190. doi:10.1073/pnas.1806018115

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476. doi:10.1038/nature07509

Wang X, Codreanu SG, Wen B, Li K, Chambers MC, Liebler DC, Zhang B. 2018. Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Mol Cell Proteomics* **17:** 422–430. doi:10.1074/mcp.RA117.000155

Waterman ML, Fischer WH, Jones KA. 1991. A thymus-specific member of the HMG protein family regulates the human T cell receptor Cα enhancer. *Genes Dev* **5:** 656–669. doi:10.1101/gad.5.4.656

Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. 2016. The ribosome-engaged landscape of alternative splicing. *Nat Struct Mol Biol* **23:** 1117–1123. doi:10.1038/nsmb.3317

Zougman A, Selby PJ, Banks RE. 2014. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14:** 1006–1010. doi:10.1002/pmic.201300553

Received January 15, 2019; accepted in revised form September 16, 2019.

Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing

Laura M. Agosto, Matthew R. Gazzara, Caleb M. Radens, et al.

Genome Res. published online November 14, 2019
Access the most recent version at doi:10.1101/gr.248435.119