# Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq*[S]

**Gloria M. Sheynkman‡, Michael R. Shortreed‡, Brian L. Frey‡, and Lloyd M. Smith‡§¶**

**Human proteomic databases required for MS peptide identification are frequently updated and carefully curated, yet are still incomplete because it has been challenging to acquire every protein sequence from the diverse assemblage of proteoforms expressed in every tissue and cell type. In particular, alternative splicing has been shown to be a major source of this cell-specific proteomic variation. Many new alternative splice forms have been detected at the transcript level using next generation sequencing methods, especially RNA-Seq, but it is not known how many of these transcripts are being translated. Leveraging the unprecedented capabilities of next generation sequencing methods, we collected RNA-Seq and proteomics data from the same cell population (Jurkat cells) and created a bioinformatics pipeline that builds customized databases for the discovery of novel splice-junction peptides. Eighty million paired-end Illumina reads and ~500,000 tandem mass spectra were used to identify 12,873 transcripts (19,320 including isoforms) and 6810 proteins. We developed a bioinformatics workflow to retrieve high-confidence, novel splice junction sequences from the RNA data, translate these sequences into the analogous polypeptide sequence, and create a customized splice junction database for MS searching. Based on the RefSeq gene models, we detected 136,123 annotated and 144,818 unannotated transcript junctions. Of those, 24,834 unannotated junctions passed various quality filters (*e.g.* minimum read depth) and these entries were translated into 33,589 polypeptide sequences and used for database searching. We discovered 57 splice junction peptides not present in the Uniprot-Trembl proteomic database comprising an array of different splicing events, including skipped exons, alternative donors and acceptors, and noncanonical transcriptional start sites. To our knowledge this is the first example of using sample-specific RNA-Seq data to create a splice-junction database and discover new peptides resulting from alternative splicing.** ***Molecular & Cellular Proteomics 12: 10.1074/mcp.O113.028142, 2341–2353, 2013.***

From the ‡Department of Chemistry, §Genome Center of Wisconsin, University of Wisconsin-Madison, 1101 University Ave., Madison, Wisconsin 53706

Received February 5, 2013, and in revised form, April 1, 2013
Published, MCP Papers in Press, April, 29, 2013, DOI 10.1074/mcp.O113.028142

Mass spectrometry-based proteomics relies on accurate databases to identify and quantify proteins, including those derived from splice variants, indels, and single nucleotide variants (SNVs)[1] (1). Most computational search algorithms detect peptides by scoring the degree of similarity between *in silico* derived and experimental peptide spectra, and thus can only identify peptides that are present in the proteomic database. If the polypeptide sequence is not present in the database used for searching, even if the peptide is present in the sample, it will fail to be detected.

Human proteomic databases used for mass spectrometric peptide identification are frequently updated and carefully curated, yet are still incomplete. Despite efforts to comprehensively annotate every gene product, there are still many undiscovered proteoforms (2) because the complete human proteome—the aggregate of all protein products expressed in every tissue, cell, and cellular state—turns out to be vastly more complex than was predicted (3–5). Furthermore, each cell or tissue-type may express a unique subset of all possible proteoforms, many of which may not be represented in existing proteomic databases. These databases are assembled from multiple datasets originating from an assortment of different human tissue and cell samples (6–11).

In recent years, alternative splicing has been shown to be a major source of cell-specific proteomic variation in humans (3, 4, 12). Human genes are comprised of introns and protein-coding exons; a protein machine, the spliceosome, removes introns from pre-mRNAs, joining exons to form a mature transcript ready for translation. Since exons can be joined in various configurations, one gene typically produces a "canonical" protein (defined as the most abundant form of the protein) as well as one or more alternatively spliced protein products, which are often thought to have modulated or altered biological function (13–16). Many alternative splice variants

[1] The abbreviations used are: SNV, single nucleotide variant; cDNA, complementary DNA; FASP, filter aided sample preparation; GENCODE, component of the ENCODE project that aims to build accurate human reference annotations; GTF, gene annotation file; ppm, parts per million (difference between theoretical and experimental peptide precursor *m/z*); PSI, percentage spliced in; RNA-Seq, RNA Sequencing; RSEM, RNA-Seq Expectation Maximization; SDT, Buffer used in FASP protocol containing SDS and dithiothreitol; TPM, transcripts per million; XCorr, SEQUEST cross-correlation score.

have been detected at the transcript level using next generation sequencing methods, especially RNA-Seq. However, it is not known exactly how many of these newly discovered alternatively spliced transcripts are being translated and if these translated products are functional.

Several approaches have been employed in the last decade to expand detection of alternatively spliced proteins using mass spectrometry. Initial approaches searched proteomic data against databases containing splice variant sequences and then confirmed the translation of a spliced sequence by detecting a peptide unique to that form (17–26). Other approaches expanded the number of alternatively spliced sequences beyond entries present in databases by constructing exon-exon databases. In this approach, exon coordinates are first determined by obtaining exon sequences from databases such as Ensembl or by using *ab initio* computational algorithms to predict the location of putative exon boundaries. Next, these exon sequences are assembled into all theoretical exon-exon (and exon-intron) combinations, and then the sequences are translated into polypeptide sequences and used for MS-based searching to discover novel splice variant peptides (27–30). To extend this approach, several research groups have restricted their exon-exon database to include only those sequences corroborated with transcript expression data (31–33), thereby eliminating spurious sequences. Two other approaches developed include a method that directly translates RNA sequence from expressed sequence tag (EST) contigs (34–37) and a proteogenomics strategy that uses the genome as a template for peptide sequence alignment (38, 39).

Several of the above methods expand proteomic databases to include entries for putative or experimentally confirmed splice variants; however, unbounded addition of more and more splice variants compiled from thousands of human cell-types is not the preferred solution. MS searching with inordinately large databases containing many more proteins than actually present in the sample causes decreased peptide identification sensitivity (as the probability of spurious spectral matches to *in silico* peptide spectra is greater) (40, 41), complications in protein parsimony (from including many redundant sequences) (42, 43), and longer analysis and search times.

Given the unprecedented advances of next generation sequencing and the maturation of RNA-Seq—longer read length, improved accuracy, increased affordability, better software—the whole transcriptome of a single sample can now be sequenced in a matter of days. As a result, all of the alternative splice junctions expressed in a single cell-type can be determined empirically. Many of the aforementioned splice detection methods rely on gene prediction programs, where reliable detection of splice forms is a challenge, or the use of data from public repositories, an amalgamation of data from multiple samples that may not reflect the splicing patterns in a given cell-type. Because RNA-Seq methods are increasingly accessible to proteomicists and these methods can empirically determine the full spectrum of alternative splicing in a sample, there is a need for bioinformatic methods that provide sample-specific, splice junction proteomic sequences from RNA-Seq data for mass spectrometry database searching.

Though the focus of this paper is on the study of alternative splice junctions, other bioinformatics strategies to extract information from RNA-Seq data have been employed to create customized mass spectrometry databases. These include reducing a database to only include sequences with transcript expression evidence (40), including fusion or chimeric sequences (44), incorporating nonsynonymous single nucleotide polymorphism (SNP) or SNV sequences (40), and, for non-model systems, building a proteomic database from *de novo* assembled transcripts (45, 46). The advent of next generation proteomics will most certainly arrive when all these sources of transcriptomic variation can be seamlessly incorporated into sample-specific proteomic databases.

We have developed a method to create a sample-specific splice junction database from RNA-Seq data and used it to discover novel splice junction peptides. We collected both RNA-Seq and proteomic data from the same cell population (Jurkat cells) and identified 12,873 transcripts and 6810 proteins. We developed a bioinformatics pipeline to retrieve high-confidence, novel splice junction sequences, translate these sequences into the analogous polypeptide sequences, and then create customized splice-sequence databases that allow for novel splice junction discovery. We discovered 57 splice junction peptides not present in the Uniprot-Trembl proteomic database using appropriately stringent MS search parameters and post-processing steps, including the use of a conservative 1% local false discovery rate and manual validation of junction peptide $MS^2$ spectra. To our knowledge this is the first example of using sample-specific RNA-Seq data to discover new peptides resulting from alternative splicing.

## EXPERIMENTAL PROCEDURES

*Cell Culture*—The Jurkat cell line (TIB-152) was obtained from the American Type Culture Collection (ATCC, Manassas, VA). Jurkat cell culture was grown in 10% Fetal Bovine Serum and 90% RPMI 1640 buffer (ATCC, Manassas, VA) at 37 °C. Cell concentration was measured using the TC10 Automated Cell Counter system (BioRad, Hercules, CA), which was validated via hemocytometer counting. Before harvesting, cells were grown to $\sim 1.3 \times 10^6$ cells/ml and had 95%+ viability as measured with the trypan blue assay.

*Proteomic Sample Preparation and Analysis*—Approximately 25 ml of Jurkat cell suspension was centrifuged at 180×g at 4 °C for 10 min. After removal of the supernatant, cells were resuspended in an equivalent volume of ice-cold PBS buffer (Invitrogen, Grand Island, NY) and centrifuged again. This step was repeated twice and the final pellet was stored at −80 °C. For cell lysis, pellets were thawed on ice and a volume of SDT lysis buffer equaling 2/3 the volume of the cell pellet was added. The pellet was pipeted up and down to assist in its solubilization, followed by incubation of the solution at 95 °C for 5 min. The SDT lysis buffer consisted of 4% SDS, 500 mM Tris-HCl (pH 7.4), and 180 mM dithiotreitol (DTT) (all reagents from Sigma-Aldrich, St. Louis, MO). The resulting lysate was sonicated (power level be-

tween 2 and 3) on ice—alternating between 30 s on and 30 s off—for 3–5 min until the viscous chromatin was solubilized and lysate had an aqueous consistency for improved sample pipetting during later steps (Misonix Sonicator XL2015, Misonix microtip PN/418, Farmingdale, NY). Protein content was measured using the 660 nm Protein Assay and the Ionic Detergent Compatibility Reagent (Pierce, Rockford, IL) to allow for accurate protein quantification in the presence of SDS.

Detergents and salts in the sample were removed and the protein was subjected to tryptic digestion by following the Filter-Aided Sample Preparation or FASP protocol developed by Wisniewski *et al.* (47). Five aliquots of lysate containing ~150 $\mu$g of protein were each added to a 100K MW Amicon Ultra filter (Millipore, Billerica, MA). After multiple FASP wash steps, reduction, and alkylation, trypsin was added directly to the filters (50:1 protein:trypsin w/w) and digested overnight at 37 °C. The next morning, filters were centrifuged at 14,000 × g for 15 min and the amount of peptide recovered was assessed via the Nanodrop UV-Vis spectrometer (Thermo Fisher Scientific, Wilmington, DE).

Approximately 500 $\mu$g of tryptic peptide digest was fractionated using high pH reverse-phase chromatography on a Shimadzu HPLC system (LC-10AD, SCL-10A VP, SPD-10A VP, Shimadzu, Columbia, MD) and a Phenomenex C18 Gemini 3$\mu$, 110Å, 3.0 × 150 mm column (Phenomenex, Torrance, CA). The high pH method was adopted from Gillar *et al.* (48). Mobile phase A (MPA) was 20 mM ammonium formate, pH 10, and B (MPB) was 20 mM ammonium formate, pH 10, in 70% acetonitrile. The HPLC flow was 0.5 ml/min and the gradient is as follows: 0% MPB isocratic for 15 min (trapping step), linear ramp to 100% MPB over 60 min, hold at 100% MPB for 5 min, to 0% MPB over 2 min, and equilibration at 0% MPB for 20 min. Fractions were collected every minute using a Gilson 203 fraction collector (Gilson, Middleton, WI) for a total of 28 fractions collected within the range of peptide elution as discernable from the UV-Vis trace. Fractions were dried down using vacuum centrifugal concentration (Savant SpeedVac, Thermo, Pittsburgh, PA) and stored at −80 °C.

Each of the lyophilized fractions generated from the high pH LC separation was reconstituted in sample solution consisting of 2% acetonitrile and 0.2% formic acid in water and then chromatographically separated on a nanoAquity LC system (Waters, Milford, MA) using a 20 cm reverse phase capillary column (100 $\mu$m i.d.) packed with 3 $\mu$m MAGIC aqC18 beads (Bruker-Michrom, Auburn, CA). Mobile phase A was 0.2% formic acid in water and B was 0.2% formic acid in acetonitrile. The full HPLC method was 180 min long and included a 90 min gradient. The mass spectrometric analysis was conducted on a Velos-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) operating in data-dependent mode. A full scan (300–1500 *m/z*) was collected at a resolution of 30,000 followed by fragmentation of the top ten precursor peptides, with +2 charge or higher, in higher-energy collision dissociation (HCD) mode (collision energy = 40) and analysis of the tandem mass spectra in the Orbitrap at a resolution of 7500. Precursor fragmentation repeat count was set to two and the dynamic exclusion was set to 60 s. XCalibur software version #2.1.0 was used for data collection.

*RNA-Seq Analysis*—RNA was extracted from Jurkat cells using Trizol Reagent (Invitrogen, Grand Island, NY). Two milliliters of Jurkat culture (~2.6 × $10^6$ cells) was centrifuged at 110 × *g* and 4°C for 5 min. After removal of the supernatant, 1 ml of Trizol reagent was added to the pellet and the solution was incubated for 15 min at room temperature. The subsequent steps are described in the Trizol Reagent RNA isolation procedure. The final total RNA pellet was solubilized in 20 $\mu$l water. The amount of RNA extracted was quantified using the Nanodrop UV-Vis spectrometer (Thermo, Rockford, IL) and mRNA integrity (RNA Integrity Number ≈10) was assessed using a 2100 Agilent Bioanalyzer (Agilent, Santa Clara, CA).

RNA-Seq paired end libraries were prepared using the Illumina TruSeq RNA Sample Prep Rev. A (kit lot #6849988, Illumina, San Diego, CA). First, mRNA was purified from total RNA using poly dT bead isolation and fragmented by heating in the presence of a divalent cation. The fragmented RNA was then converted to cDNA with reverse transcriptase using random hexamer priming and the resultant double stranded cDNA was purified. cDNA ends were repaired, adenylated at the 3′ ends, and then ligated to Illumina adapter sequences. Primers matching the adapter sequences were then used to PCR amplify the cDNA sequences. These sequences were run on an Invitrogen 2% Size Select Gel (Lot# R19090–01) and a band corresponding to ~350 base pairs was excised and used for paired end (2 × 200 bp) sequencing on an Illumina HiSeq 2000. Raw cluster station data was post-processed and a total of 80 million RNA-Seq reads in fastq format were used for splice junction discovery.

All fastq files used in this study can be accessed at NCBI's Gene Expression Omnibus (GEO) repository (49) by using the following link: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE45428.

*Construction of the Splice Junction Database*

*Splice Junction Discovery with Bowtie-Tophat Software*—Annotated and unannotated junctions were detected using the Bowtie (v0.12.7) and Tophat (v1.4.0) splice-junction discovery programs (50, 51). All default Bowtie parameters were used. In Tophat, the mate inner distance was set to 150. Two rounds of Bowtie-Tophat processing were conducted with a supplied set of RefSeq gene model annotations in GTF format (7): the first round detected junctions only matching the gene annotation file (option -no-novel-junctions) and the second round detected all junctions, both aligning to the GTF file and novel (option -G). All data processing was conducted on the Phoenix cluster at the University of Wisconsin-Madison Chemistry department. The set of novel junctions not matching the RefSeq gene annotation was extracted from sets of output .bed files by in-house Perl scripts.

*Translation of the Junction Nucleotide Sequences*—The set of unannotated splice junction coordinates containing six or more supporting RNA-Seq reads were translated into putative peptide splice junctions. The exon coordinates were extended by 66 nucleotides on both of the flanking ends of the junction. Junctions frequently overlapped with known genes, therefore the transcriptional strand (*i.e.* forward or reverse) of the junction was inferred from this association. The sequences resulting from a three frame translation, either on the forward or reverse strand, were extracted from the reference genome (hg19), translated into amino acid sequence, and trimmed to the first arginine or lysine (MS data was from a tryptic digest). Sequences less than 5 amino acids or containing a stop codon near the splice site were removed. All splice junction sequences were appended to the canonical Uniprot proteomic (release-2012_10; 20,225 entries) and Global Proteome Machine CrAP database (version 2012.01.01, 115 sequences). Two additional customized databases were built by appending junction sequences to the Uniprot/Trembl (release-2012_10; 86,881 entries) and to the Ensembl (release GRCh37.70.pep.all) protein databases and searches were conducted as described below.

*Mass Spectrometry Junction Database Searching*—Raw mass spectrometry files were searched against the customized UniProt+CrAP+Junction (53,476 entries total) database using the Percolator search node within Proteome Discoverer (v1.3.0.339, Thermo Fisher Scientific, San Jose, CA). Percolator is a machine-learning supplement to the SEQUEST search algorithm that increases the sensitivity and specificity of peptide identifications (52). Default peaklist-generating parameters were used. Precursor *m/z* tolerance was set to 10 ppm and product *m/z* tolerance was set to 0.05 Da. Peptides with up to two missed cleavages (trypsin) were permitted. Variable methionine oxidation and static carbamidomethylation were used. Using reversed sequences as a decoy database, peptides passing both 1

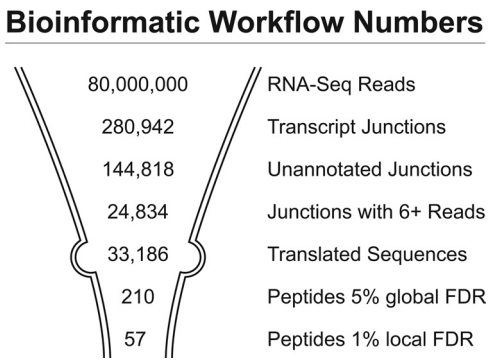


FIG. 1. **Results overview for the bioinformatic pipeline.**

and 5% global false discovery rate (FDR) and 1 and 5% local FDR (splice junction hit group) were used for downstream analysis. Validation was based on q-values generated by Percolator. For identification of a protein using Proteome Discoverer, protein grouping and strict parsimony principle was enabled, leucine and isoleucine were considered equal, and only peptides passing 1% FDR and having a delta Cn higher than 0.15 were used. A minimum of two peptides per protein was required for identification.

All mass spectrometric raw files associated with this study may be downloaded via FTP from the PeptideAtlas data repository (53) by accessing the following link: http://www.peptideatlas.org/PASS/PASS00215.

## RESULTS

*Overview*—RNA-Seq and MS-based proteomics data was collected; 19,873 transcripts, which map to 12,873 genes, and 6810 proteins were identified. RNA-Seq reads were used to discover 144,818 unannotated splice junctions using Bowtie and Tophat software. A total of 24,834 Tophat junctions passing an expression cutoff were translated into polypeptide sequences. Either three frames or the one frame inferred from comparison to gene models was translated, resulting in 33,136 polypeptide entries. The splice junction sequences were appended to the Uniprot canonical database (~20,000 entries) and searched against the mass spectrometric data. A total of 210 splice junction peptides that were absent in the complete Uniprot/Trembl database (~87,000 entries) but present in RNA-Seq derived junctions passed 5% global FDR. A local FDR was applied to the splice junction peptides and 72 (5% local FDR) and 57 (1% local FDR) peptides were identified. An overview of these results are depicted in Fig. 1.

*Terminology Employed to Define the Types of Peptides Identified in this Study*—"Uniprot peptides" are all peptides identified by searching proteomics data against the full Uniprot/Trembl database that includes isoforms (86,766 entries). "Splice junction peptides" are all the peptides identified from RNA-Seq data (translated splice junctions) in this study that were not present in the full Uniprot/Trembl database.

*mRNA and Protein Data Collection*—The transcriptomic and proteomic data collection workflow was designed to allow for accurate splice peptide detection (Fig. 2). The protein and mRNA samples were extracted from the same Jurkat cell population to build a sample-specific junction database, one

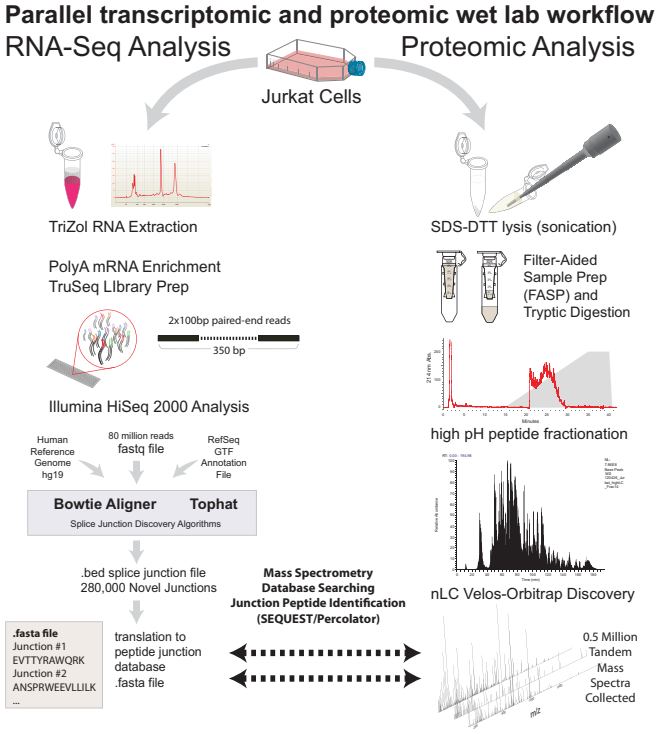


FIG. 2. **Transcriptomic and proteomic data collection.** Both total RNA and protein were extracted from the same Jurkat cell population. mRNA was purified from Trizol extracted total RNA, then Illumina TruSeq paired-end libraries were prepared. RNA-Seq data was processed using Bowtie/Tophat to discover new junctions that were then converted into polypeptide sequences. Protein was extracted using an SDT-based buffer, sonicated, and desalted and tryptically digested using the FASP protocol (Mann *et al.*). Peptides were chromatographically fractionated on a high pH LC system, and the resultant fractions were analyzed on a nanoLC-Velos Orbitrap operating in data-dependent mode. A customized junction database derived from the RNA-Seq data was used for MS database searching.

with minimal intra and inter-laboratory variation. Protein was extracted from cells using an SDS and DTT-based buffer (SDT) and the FASP protocol (47). This protocol allows unbiased extraction and digestion of all protein groups (including hard-to-solubilize transmembrane proteins), an important factor when seeking to identify a proteoform (2). Wisniewski *et al.* demonstrated that the composition of proteins identified using FASP corresponded to the expected abundances of Gene Ontology groups, with all protein groups evenly represented (47). In addition, when we compared SDT-based and urea-based extractions, we found that ~20% more protein (Bicinchoninic acid assay) was extracted with SDT and that membrane proteins were more represented (results not shown). Total RNA was extracted from cells using a standard Trizol protocol. To provide a comprehensive RNA-Seq data set for the sensitive discovery of alternative splice forms, 80 million reads of the longest RNA-Seq read type available on the Illumina platform were analyzed: libraries were derived from 350 bp cDNA sequences and 100 bp paired-ends were sequenced. In summary, the RNA and protein wet laboratory

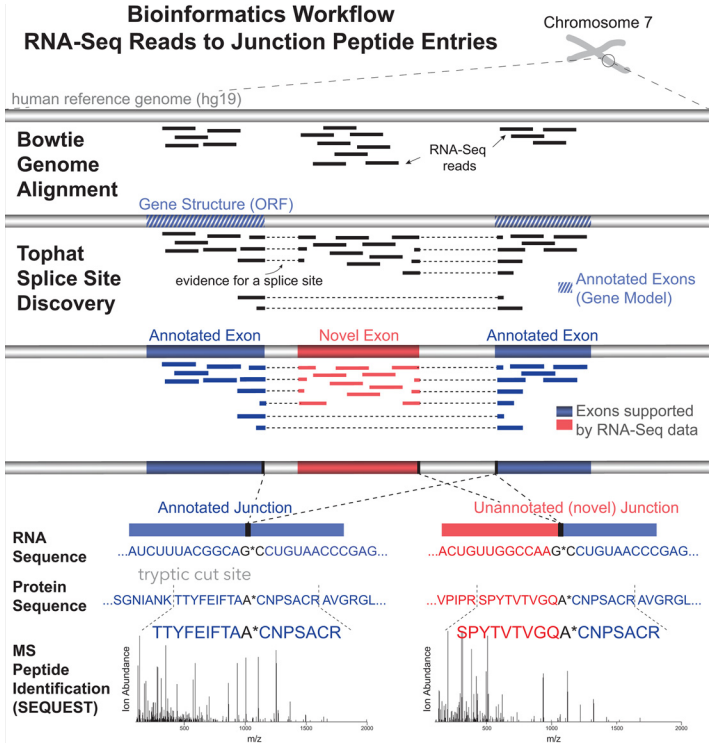


FIG. 3. **Bioinformatics workflow to convert raw RNA-Seq reads into junction peptide sequences.** Bowtie and Tophat are used to align reads to the genome and annotated gene structure sequences (RefSeq). During Tophat splice junction alignment, reads are segmented and can align across exon-exon boundaries. When many reads support the presence of a novel splicing event, the junction is reported in .bed format. The list of unannotated junctions are converted to peptide sequence and searched against tandem mass spectra. Here, we show an example of a canonical and an alternative splice site identification from the detection of two tryptic peptides, where A* represents the amino acid residing, alanine (A) in this case, at the junction.

experiments were designed so that transcript-level junctions are sensitively detected and included in a comprehensive splice-junction database and the maximum number of discoverable splice-junction peptides using bottom-up proteomics are detected.

We measured the number of transcripts and proteins detected from both the RNA-Seq and peptide MS data, respectively, to compare the transcriptomic and proteomic data sets. RNA-Seq reads were processed by RSEM (RNA-Seq by Expectation-Maximization) to estimate transcript abundances (54). Reads were aligned to a synthetic transcriptome and the number of reads associated with a given transcript was used to estimate that transcript's abundance in TPM (transcripts per million). RSEM processing of 80 million RNA-Seq reads resulted in 19,320 transcripts that mapped to 12,873 genes (TPM>1). Tandem mass spectra were processed by Proteome Discoverer (SEQUEST + Percolator algorithm) to infer protein identities. Experimental peptide MS spectra were processed with SEQUEST, followed by rounds of semi-supervised machine learning with Percolator, a target-decoy search using a 1% FDR, and grouping of proteins using maximum parsimony. Proteome Discoverer processing of 488,149 $MS^2$ higher energy collision dissociation spectra resulted in 77,733 Uniprot peptides and 6810 proteins. Full results are in the supplemental table.

We also searched the mass spectrometric data against the UniProt/Trembl database (~87,000 entries) to measure the number of isoforms. We were able to detect two or more protein isoforms for 86 genes, where each isoform required at least one unique peptide that passed a 1% FDR cutoff. However, this number is likely to be artificially low because the detection of isoforms using bottom-up proteomics requires a tryptic peptide unique to each isoform and protein sequence coverage is typically low (<25% coverage). The actual number of genes expressing more than one protein isoform is believed to be much higher (9).

*Discovering Alternative Splice-Junctions from RNA-Seq Data*—Bowtie and Tophat software were used to discover splice junctions from 80 million RNA-Seq reads, and from these junctions, a peptide junction database was created for use in mass spectrometric data searching. Part of the procedure described in this section is illustrated in Fig. 3.

Bowtie software efficiently aligns short RNA-Seq reads to a reference sequence (human reference genome, synthetic transcriptome, etc.) and Tophat discovers junctions not represented in the gene models. Both methods work together to discover novel junctions. Tophat discovers novel junctions primarily by finding RNA-Seq reads that span an exon-exon boundary, the most direct evidence of transcript splicing. It does this by segmenting the reads into subsequences and aligning the subsequences to the genome. When a read is "split"—one half of the read aligns upstream of an intron and the other half of the read aligns downstream of the intron—this is evidence for a novel splice junction. Tophat uses Bowtie for the alignment process and because both programs efficiently process RNA-Seq reads, the software can be run on desktop computers or local computer clusters accessible to most labs.

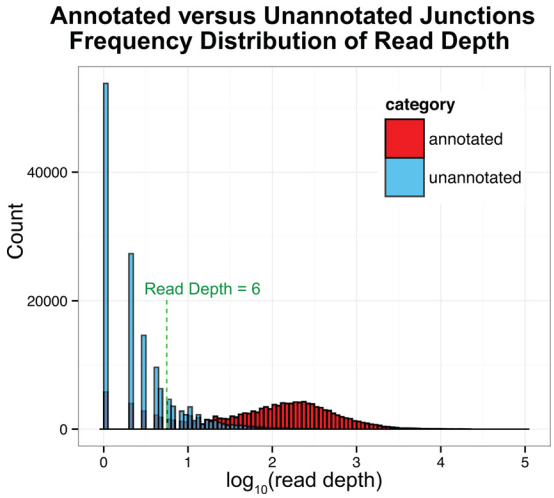


FIG. 4. **Read depth frequency distribution for Tophat-detected annotated and unannotated junctions.** A majority of the reads aligned to RefSeq annotated junctions, as shown in the red histogram, whereas a lower number of reads, on average, aligned to unannotated junctions. Unannotated junctions with fewer than 6 supporting reads were removed prior to downstream analysis (green dotted line).

Processing of 80 million paired-end reads by Tophat/Bowtie resulted in a total of 280,942 junctions before filtering: 136,123 junctions present in RefSeq annotations (NM accession entries, representing RefSeq mRNA sequences) and 144,818 unannotated junctions. The list of annotated splice junctions were derived from NCBI RefSeq gene annotations because RefSeq has high quality, conservative annotations with minimum redundancy. Of the 144,818 unannotated junctions, 19,942, 1185, and 22 junctions had over 10×, 100×, and 1000 × read coverage (depth), respectively.

*Selection of Minimum RNA-Seq Read Depth for Junction Inclusion in Database*—The RNA-Seq read depth, the number of RNA-Seq reads supporting the existence of a novel junction, was examined: a majority of the unannotated junctions were lower in abundance (read depth), and a significant number of junctions had just one supporting RNA-Seq read (Fig. 4). We hypothesize that junctions containing a small number of supporting reads are either expressed at low levels and represent stochastic transcription (55, 56) or are the result of errors in the sequencing reads or Bowtie alignment step (57, 58). We reasoned that many of these low coverage junction sequences are unlikely to result in a peptide identification, either because they are false positives or expressed at an extremely low-level (below 1 copy/cell).

We elected to use junctions with six supporting reads or higher in the customized database to strike a balance between inclusion of novel junction sequences to promote peptide discovery and exclusion of junction sequences to minimize false positives. Two observations support using this cutoff. First, the transcript expression levels (RSEM output) were plotted against the protein expression levels (spectral counting), and the minimum transcriptional abundance required to detect a protein corresponded to ~6 RNA-Seq reads per junction. Second, multiple proteomic searches were performed, each differing by the minimum RNA-Seq read depth required for a junction sequence to be included in the database. For example, one search was against a database that included junctions having 1× RNA-Seq read depth or higher whereas another search was against a database that included junctions having 10× RNA-Seq read depth or higher. Uniprot peptide and splice junction peptide score distributions (see previous nomenclature section) were compared to determine the incidence of false positives in the group of splice junction peptide identifications. After taking into account the above observations, a lower read depth cutoff of six was selected for database construction.

*Construction of a Customized Junction Database from RNA-Seq Data*—A pipeline was developed to convert unannotated junction sequences into putative polypeptide entries for mass spectrometry searching. A total of 24,834 junction sequences with six or more reads were translated into 33,186 amino acid sequences. To accomplish this, junction ends were extended, translation frame was inferred (when possible), and improbable sequences were trimmed or removed.

Transcript sequences were extended upstream and downstream of the Tophat junction. Each Tophat junction is represented by four coordinates: the start and end nucleotides of both the upstream and downstream Tophat "exon" (coordinates 1, 2, 3, and 4 in Fig. 5). In humans, the average exon size is 148 nucleotides in length (7), but the reported Tophat "exons" ranged from 8 to 100 nucleotides and are an average of 64 nucleotides (Fig. 6). This exon size distribution results from the Tophat Software. Tophat reports only the stretch of sequence—upstream and downstream of the splice site—that has evidence: aligned 100 bp RNA-Seq reads that overlap the junction. To increase the probability of detecting peptides that extend past Tophat junction ends (coordinates 1 and 4 in Fig. 5a), additional sequence was appended to both sides of the junction. Before translation, the sequence coordinates of each Tophat junction were thus lengthened at flanking exon ends (5′ end of upstream exon, 3′ end of downstream exon) by 66 nucleotides.

The frame translation was inferred for a subset of the Tophat junction sequences. In the case that a novel junction's left splice site (coordinate 2 in Fig. 5*B*) corresponded to the left splice site of a known gene structure, the frame translation was inferred. This is reasonable because the upstream exon is part of a known gene model and will most likely be translated in the same frame as the canonical splice form. For all other junctions where the frame could not be inferred, such as when there were novel left and right (coordinate 2 and 3 in Fig. 5*B*) splice sites or a novel left splice site (coordinate 2), all three frames were translated.

Improbable sequences were either removed or trimmed. First, short peptides (<5 amino acids), peptides with an abun-

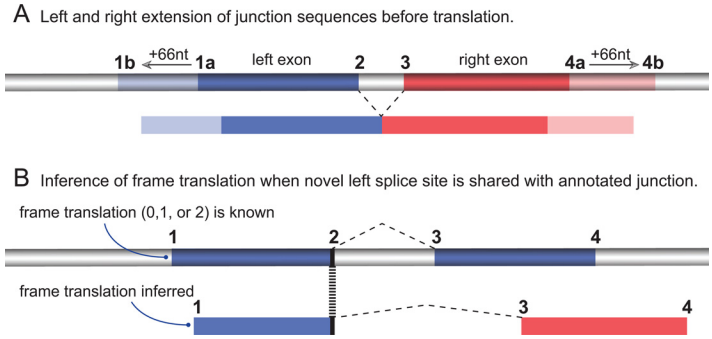


FIG. 5. **Junction sequence processing before translation into peptide sequences.** *A*, Each Tophat junction consists of four coordinates: *1a*, *2*, *3*, and *4a*. Sequences were extended by 66 nucleotides before translation to increase the probability of detecting peptides that may partially protrude past *1a* and *4a*. *B*, The frame translation was inferred for the subset of cases in which the left splice site, *2*, of the unannotated junction corresponded exactly to the left splice site of an annotated junction.

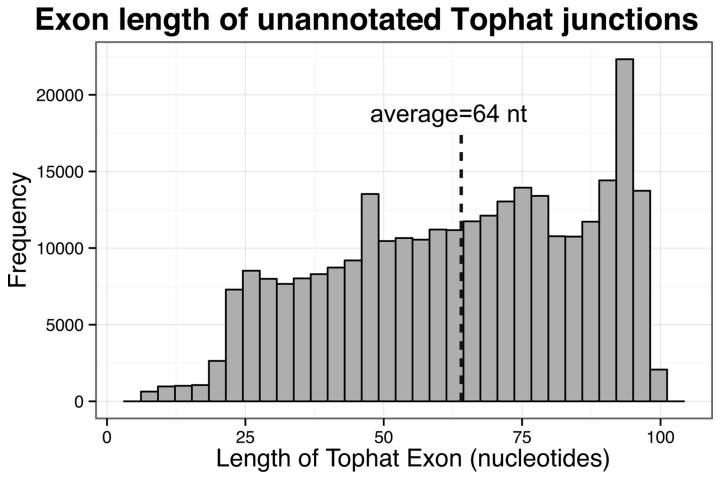


FIG. 6. **Distribution of the Tophat exon lengths for unannotated junctions.** The Tophat exon lengths are a consequence of the *de novo* splice discovery program, where only RNA-Seq reads spanning a splice site can be used as direct evidence for the existence of a junction.

dance of stop codons, and peptides that did not include the splice site amino acid were removed. Second, polypeptide start and end sites each were trimmed to the first occurrence of a lysine (K) or arginine (R), preventing the inclusion of nontryptic fragment sequences. Sequences were trimmed because the proteomics data for this study was based on detection of tryptic peptides, all of which begin after the C terminus of a lysine (K) or arginine (R) and likewise end with a K or R.

After subjecting the 33,186 unannotated transcript-level junctions to the aforementioned processing steps, 24,834 remained and these sequences were translated into 33,589 junction peptide entries (the higher number of peptide entries resulted from requisite 3-frame translations) and were integrated into a customized junction database (see supplemental table for full list). Most of the 8352 junctions filtered out were because of high frequency stop codons, or possibly to out-of-frame translation. To create customized junction databases, the junction sequence entries were appended to the following protein databases: canonical UniProt reference (20,225 entries), UniProt/Trembl (86,881 entries), or Ensembl (104,785 entries, version 70). The addition of junction peptide entries increased the size of the Uniprot, UniProt/Trembl, and Ensembl databases by 13.1% (1,474,776 aa were added to 11,291,209 aa), 4.1% (1,474,776 aa were added to 36,164,128 aa), and 3.7% (1,474,776 aa were added to 39,786,499 aa), respectively. The raw MS files were searched against each of these three combination databases to identify the subset of splice junction peptides. The lists of splice junction peptides among the three searches were very similar (see supplemental Table S1); we have chosen here to focus on MS results from the UniProt reference + junction sequence database. Junction peptide sequences identified from the Uniprot reference + junction sequence database were BLAST searched against the full UniProt/Trembl database (~87,000 entries) and peptides not present in UniProt/Trembl (hence new splice junction peptides) were retrieved.

*Balancing Splice Peptide Discovery and False Positives*—It has previously been demonstrated in multiple settings that when expanding a proteomic database to include possible proteoform sequences or when searching MS data against six-frame translated reference genomes, the false positive rate increases and the sensitivity of peptide identification decreases (41, 59, 60). Therefore, proper statistical methods and scoring thresholds must be employed for accurate identification of new variants.

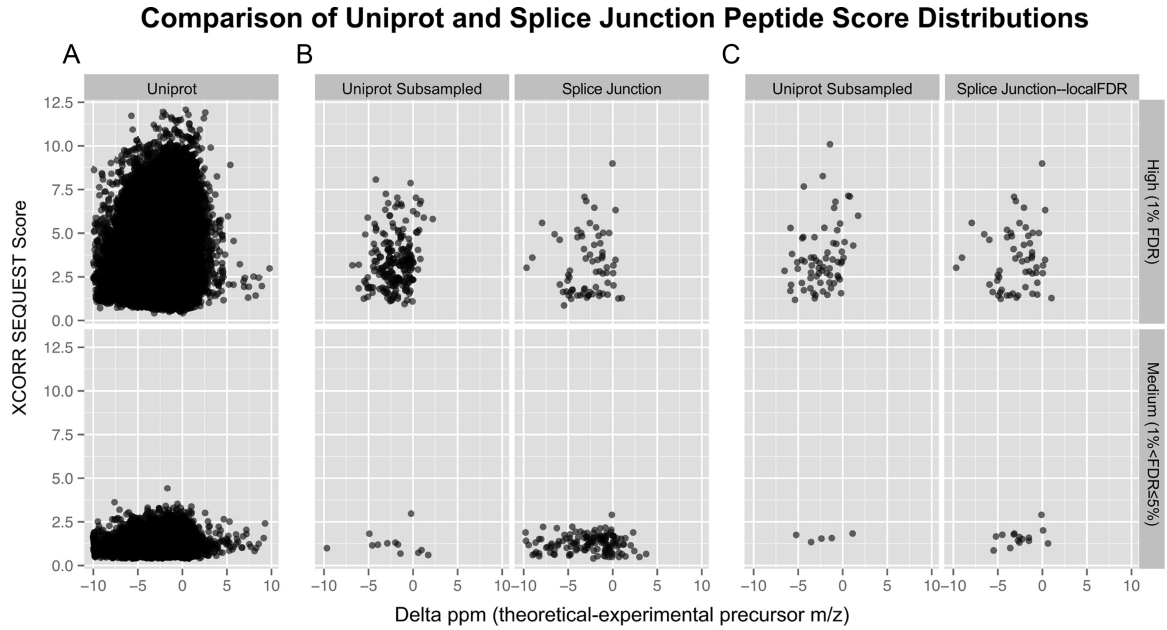


FIG. 7. **Comparison of peptide score distributions for canonical and junction peptides.** For the comparison of peptide scores, delta ppm (difference, in parts per million, between measured and experimental precursor *m/z*) *versus* XCorr (cross-correlation value from MS search) SEQUEST score was plotted. *A*, Score distributions for peptides matching the UniProt/Trembl human proteomic database. *B*, Junction peptide score distribution ($n = 210$, 5% FDR) compared with the score distribution of 210 peptides randomly subsampled from *A*. *C*, Local FDR junction peptide score distribution ($n = 72$, 5% local FDR) compared with score distributions for 72 peptides randomly subsampled from *A*. The plotted points in *B* illustrate that junction peptides tend to have lower scores than the canonical ones when employing the global FDR thresholds. This mismatch indicates the 210 junction peptides will have greater than 5% false positives. *C*, shows the remedy to this situation, namely calculation of a more strict local FDR (based on the Percolator posterior expectation probability score), which then makes the canonical and junction distributions quite similar.

A conservative local FDR based on Posterior Error Probability (PEP) values was used for identified splice junction peptides to reduce the number of false positives (61). MS data was processed using Percolator, a machine-learning adaptation to SEQUEST (52), and a reverse target-decoy database. The search yields were 77,733 and 83,385 identified Uniprot peptides at a 1 and 5% false discovery rate, respectively. To ascertain any peptide scoring biases for the Uniprot and splice junction peptides, the delta precursor parts per million (ppm) and XCorr SEQUEST scores were plotted for different subsets of peptides (Fig. 7). Fig. 7*B* shows a comparison of splice junction peptide score distributions to the score distributions from the same number of subsampled Uniprot peptides ($n = 210$). The population of splice junction peptides contained a disproportionately higher number of lower scoring peptides (passing 5% FDR, but not 1% FDR). This is probably because of the large number of specious sequences resulting from the three-frame translated or even noncoding junctions. To resolve this issue, we elected to apply a 1% local FDR threshold to the splice junction peptides based on the Posterior Error Probability, or PEP, values. The Posterior Error Probability for a peptide identification is the probability that the experimental spectra actually originated from the sequence reported. The local FDR for a subgroup of peptides is calculated by dividing the expected number of false positives (the sum of Posterior Error Probability values for all peptides within the group) by the total number of peptides (62). Fig. 7*C* shows that applying a local FDR threshold to the splice junction peptides allows these peptides to achieve a similar score distribution to the sub-sampled Uniprot peptides ($n = 57$, 1% FDR; $n = 72$, 5% FDR). Thus, 57 novel junction peptides have been discovered at the local false discovery rate of 1%.

*Presence of Splice Junction Peptides in Various Databases*—The UniProt reference protein set is a popular database used in proteomics; however, many other databases and sequence repositories are also available for researchers to use. Therefore, we checked to see how many of the 72 splice junction peptide sequences that were not in the Uniprot/Trembl database were present in other publicly available databases. We did this by BLAST searching each of the 72 splice junction peptide sequences against the human datasets found within NCBI's dbEST, INSDC (the International Nucleotide Sequence Database Collection, which includes Genbank and the DNA Data Bank of Japan), Ensembl, Genscan, and the NIST peptide mass spectral library. We determined how many sequences were already present in these databases and found 22, 22, 7, 12, and 5 peptides, respectively. A table showing each peptide sequence and the database(s) it was found in is available in the supplementary information. All in all, 39 of the sequences cor-

TABLE I
*Events represented by 57 discovered splice-junction peptides. Frequency of splicing events represented by the 57 junction peptides passing 1% local FDR. A variety of different splicing events were detected from RNA-Seq specific splice junction entries*

| Splicing event | Frequency |
|---|---|
| Alternative acceptor | |
| +3nt | 13 |
| -3nt | 2 |
| 35nt | 1 |
| 77nt | 1 |
| Skipped exon | |
| 1 Exon | 9 |
| 2 Exon | 2 |
| 3 Exon | 1 |
| Novel exon | |
| Left | 3 |
| Right | 4 |
| Completely unannotated | 7 |
| Alternative donor (-21, −12, +12, +23, +24, +58) | 5 |
| Alternative Transcriptional Start Site (TSS) | 2 |
| Within intron | 2 |
| Cross gene | 1 |

responding to the 72 splice junction peptides were found in one or more of the nucleotide sequence or proteomic repositories; however, most of these had limited or no evidence of protein expression. It may be noted that although BLAST analysis easily determines if a particular sequence is present in a database, that does not mean that the splice junction peptide would be identified with statistical significance in a mass spectral search against that same database.

*Discovered Alternative Splice Junction Peptides*—We designed a bioinformatic workflow that leverages RNA-Seq data to create a customized splice-junction database. Despite the comprehensiveness of the UniProt/Trembl human proteomic database—86,766 discrete protein entries ranging from manually validated to computationally predicted sequences (entries without evidence for the expression of the protein)—we still discovered 57 novel splice junction peptide sequences that were absent in the UniProt/Trembl database. The RNA-Seq customized splice junction database provided a promising mechanism for discovery of these peptides.

The discovered peptides represented many different types of splicing including exon skipping events, alternative donors and acceptors, novel exons, alternative transcriptional start sites and novel exon-exon junctions (Table I). A full table of each splice junction peptide that includes information such as the observed canonical peptide, a description of the splicing event (*e.g.* exon skipping), and transcript level alternative/canonical splicing frequencies, may be found in the supplemental table. The most frequent splicing types exhibited by the splice-junction peptides were alternative acceptor and donor sites and skipped exons.

The most common splicing events were small insertions and deletions (indels) occurring at the 3′ acceptor exons, frequently characterized by the NAGNAG motifs where two AG dinucleotide splice site acceptors sit in close proximity to each other: this agrees with recent gene validation efforts of the GENCODE gene annotation project in which mass spectrometry data retrieved from the Global Proteome Machine (GPM) and PeptideAtlas were aligned to GENCODE gene models to assess the number of translated products (17). NAGNAG tandem splicing may cause subtle changes in the protein sequences, just the insertion or deletion of one amino acid, yet there is evidence that these alternative forms are not merely the result of stochastic noise from splicing machinery. Recently, evidence has been mounting that NAGNAG splicing plays a functional role. These splicing sequences have been shown to be evolutionarily conserved across species and the ratio of canonical to alternative splicing has been shown to be tissue-specific—facts that suggest NAGNAG splicing is important to protein function (63–65). The PSI ($\Psi$) or "Percentage Spliced In" (66) was calculated for all fifteen peptides exhibiting alternative acceptor splicing. "Percentage Spliced In" (PSI) is the fraction of minor and major isoforms, expressed as a percentage. The PSI ranged from a low of 0.2% to a high of 27.1% and the average was 5.6%.

## DISCUSSION

A peptide or protein sequence must be listed in the database to be identified by mass spectrometry; hence, proteomics relies on databases to discover new proteoforms. Despite the large strides that groups curating databases such as Swiss-Prot/Trembl and GENCODE have made in completing gene models, including improving pipelines to better discriminate between putative and actual protein sequences by incorporating the latest high-throughput MS data, not all proteins are listed. The diversity of human proteoforms is immense and proteoforms expressed in thousands of human cell types have yet to be cataloged. Furthermore, the list of protein entries in the human reference proteome is consolidated from the human cells studied to date and may not reflect variants present in any particular sample. One of the major sources of cell-type specific proteomic variation is alternative splicing, where the protein coding exons of a gene are stitched together in various combinations to create multiple splice forms. While there have been efforts to create expanded databases that capture all alternative splicing variants, we suggest that the solution should not be unbounded expansion of a central database, but rather the customization of databases for specific cell-types. Due to recent unprecedented advances in next generation sequencing and RNA-Seq, this proteomics strategy is now within reach.

We describe here a novel strategy to use a sample-specific RNA-Seq dataset to characterize new cell-type specific splicing events not yet captured in proteomic databases. We collected RNA-Seq and proteomic data from a single cell population (Jurkat cells), constructed an empirically derived splice-junction database from RNA-Seq data, searched the

accompanying mass spectrometry data against the customized splice-junction database, and discovered new splice-junction peptides that were absent from the UniProt/Trembl proteomic database, which includes all putative gene annotations predicted from the Ensembl pipeline. To our knowledge, this is the first report of using RNA-Seq data to discover mRNA splice junctions *de novo* from direct alignment of RNA-Seq reads with the reference genome (exon boundaries not supplied) and construction of a customized splice junction database from the splicing events that were detected.

We found that an important element in creating such customized databases is achieving a balance between the inclusion of all putative proteoform sequences (for which there is transcript-level evidence) to maximize discovery of new forms, and the reduction of database size to control for sequence redundancy and false positives. Unbounded expansion of databases by including additional protein sequences, such as those derived from proteogenomics (six frame translation), *ab initio* gene predictors, and transcriptomics data (three or six frame translation), is problematic because it increases false positives, redundancy, and MS search times. The false positive rate is increased when many spurious protein sequences, corresponding to proteins not expressed in the sample, are added to the database, because the presence of these sequences increases the probability that an experimental spectrum matches that sequence by random chance (59). Note that some of the junction peptide sequences described in this article were found in expanded databases (*e.g.* GenBank), but mass spectrometric searching against these large, all-inclusive databases is problematic for the reasons stated above. Redundancy is also increased by adding many closely related proteoforms, and this confounds protein parsimony, the inference of protein from peptides (42, 43). Conversely, in the case of our experimentally determined splice-junctions, strict reduction of the database to include only those sequences with the highest expression levels (>30 TPM) was inappropriate: there are plenty of examples of low transcript abundance but high protein abundance and *vice versa* (67, 68). Therefore, to strike a balance between discovering novel alternative splice junctions and minimizing the number of spurious sequences, we included junction sequences with six or more supporting RNA-Seq reads and used a local 1% FDR for splice junction peptides.

Another important issue in the discovery of alternative splice forms at the protein level is the low number of splice-specific peptides actually identified, an issue that has been revealed by work reported in the literature (17, 19, 22, 24, 25, 30, 31). Part of the reason for the low number of alternative splice variants detected are the technical differences between RNA-Seq and bottom-up proteomics, namely sequence coverage and detection sensitivity. RNA-Seq reads are obtained by, first, randomly fragmenting mRNA molecules with a divalent cation and heat, and second, reverse transcribing these RNA fragments into cDNA and using PCR to amplify this initial cDNA library. These steps allow for the detection of reads spanning the whole transcript (100% coverage) and corresponding to transcripts expressed at a low-level (69). Peptide spectra, on the other hand, are obtained by, first, employing a proteolytic enzyme to cleave the protein at prescribed sites, and second, directly electrospraying the peptide into a mass spectrometer and collecting spectral scans. These steps allow the detection of only those peptides amenable to LC-MS/MS (~5–25% coverage) and corresponding to proteins expressed at a high enough level for detection (attomoles-femtomoles). The consequence of these RNA and protein measurement differences is that it is much more difficult to detect alternative splice variants at the protein level than the RNA level. Transcripts can be sensitively (<1 transcript/cell) and completely (100% sequence coverage) characterized, but for proteins, only moderately or highly expressed (>1 protein molecules/cell) proteins are usually detected and amino acid sequence coverage is typically low (~5–25%). Alternatively spliced proteins are difficult to detect because 1) they have lower cellular abundances than the canonical forms, 2) require at least one splice form-specific peptide for unambiguous detection, likely one spanning a junction or residing in a splice form-specific exon (24), and, 3) the alternative splice variant sequence is sometimes not yet in the database.

The number of alternative splice forms expected to be detected in a bottom-up proteomics experiment has been estimated using computational approaches (19, 22, 24). Some authors reported that they identified the expected number of splice-specific peptides while other authors identified far fewer peptides than predicted. These discrepancies were attributed to the underlying assumptions of their statistical models. In any case, this paper shows that new splice junction peptides can be detected directly from customized databases built from RNA-Seq data. It is likely that these peptides represent the tip of the iceberg, and that there are many more splice-specific peptides that are currently undetected. Extensions of the strategy employed in this paper may be employed to increase the ability to detect splice junction peptides. For example, utilizing multiple proteolytic enzymes (LysC, GluC, etc.) will increase the odds of creating a splice-specific peptide detectable by LC-MS/MS, or targeted proteomics strategies such as selected reaction monitoring (SRM) analysis could be employed to decrease detection limits for splice junction peptides of interest that have low abundances.

RNA-Seq has developed rapidly and its cost has decreased greatly making it accessible to most research organizations. Because of this technological revolution, there is a great opportunity for next generation proteomics to use sample-specific, customized databases built from RNA-Seq data. The present work on discovery of novel splice junctions is one important aspect of proteomic variation, but there are many other variations (*e.g.* SNVs, RNA fusion products) that may also be captured in custom databases. As RNA-seq technologies continue to become increasingly affordable, accessible,

and sensitive, the power and utility of this new strategy for the discovery of proteomic variation will continue to expand.

*Acknowledgments*—We would like to thank Dr. Mark Scalf for assistance with the mass spectrometric data collection. We would like to thank Dr. Victor Ruotti and Dr. Colin Dewey for helpful discussions regarding the transcriptomics pipeline. RNA-Sequencing work was performed at the University of Wisconsin-Madison Biotechnology Center.

* This work was supported by NIH grants 1P01GM081629 and 1P50HG004952. GMS was supported by the NIH Genomic Sciences Training Program 5T32HG002760. The Phoenix Computing Cluster at the University of Wisconsin-Madison Chemistry Department is supported by the National Science Foundation Grant CHE-0840494.

S This article contains supplemental Table S1.

¶ To whom correspondence should be addressed: Department of Chemistry, Genome Center of Wisconsin, University of Wisconsin-Madison, Madison, WI 53706. E-mail: smith@chem.wisc.edu.

REFERENCES

1. Ning, K., Fermin, D., and Nesvizhskii, A. I. (2010) Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* **10,** 2712–2718
2. Smith, L. M., and Kelleher, N. L. (2013) Proteoform: a single term describing protein complexity. *Nat. Meth*. **10,** 186–187
3. Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–1415
4. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476
5. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H. M., Yu, J., Wang, J., Huang, G. Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S. Z., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H. Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W. H., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J. R., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., and Int Human Genome Sequencing, C. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921
6. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res* **12,** 996–1006
7. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** D61–D65
8. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2005) The universal protein resource (UniProt). *Nucleic Acids Res.* **33,** D154–D159
9. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774
10. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: An integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988
11. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002) The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41
12. Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T. J., Sladek, R., and Majewski, J. (2008) Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40,** 225–231
13. Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., and Soreq, H. (2005) Function of alternative splicing. *Gene* **344,** 1–20
14. Blencowe, B. J. (2006) Alternative splicing: New insights from global analyses. *Cell* **126,** 37–47
15. Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z., and Dunker, A. K. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 8390–8395
16. Wang, G. S., and Cooper, T. A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **8,** 749–761
17. Ezkurdia, I., del Pozo, A., Frankish, A., Rodriguez, J. M., Harrow, J., Ashman, K., Valencia, A., and Tress, M. L. (2012) Comparative Proteomics Reveals a Significant Bias Toward Alternative Protein Isoforms with Conserved Structure and Function. *Mol. Biol. Evol.* **29,** 2265–2283
18. Menon, R., Zhang, Q., Zhang, Y., Fermin, D., Bardeesy, N., DePinho, R. A., Lu, C., Hanash, S. M., Omenn, G. S., and States, D. J. (2009) Identification of Novel Alternative Splice Isoforms of Circulating Proteins in a Mouse Model of Human Pancreatic Cancer. *Cancer Res.* **69,** 300–309
19. Tress, M. L., Bodenmiller, B., Aebersold, R., and Valencia, A. (2008) Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol*. **9,** R162
20. Menon, R., and Omenn, G. S. (2010) Proteomic Characterization of Novel Alternative Splice Variant Proteins in Human Epidermal Growth Factor Receptor 2/neu-Induced Breast Cancers. *Cancer Res.* **70,** 3440–3449
21. Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. A., Wesselink, J. J., Yeats, C., Olason, P. I., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo,

D., Lagarde, J., Laskowski, R. A., López, G., Sadowski, M. I., Watson, J. D., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Størling, Z., Orsini, M., Assenov, Y., Blankenburgh, H., Huthmacher, C., Ramirez, F., Schlicker, A., Denoued, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S. E., Reymond, A., Birney, E., Brunak, S., Casadio, R., Guigo, R., Harrow, J., Hermjakob, H., Jones, D. T., Lengauer, T., Orengo, C. A., Patthy, L., Thornton, J. M., Tramontano, A., and Valencia, A. (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 5495–5500

22. Severing, E.I., van Dijk, A. D., and van Ham, R. C. (2011) Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. *BMC Plant Biol.* **11,** 82
23. Leoni, G. L. G., Le Pera, L., Ferre, F., Raimondo, D., and Tramontano, A. (2011) Coding potential of the products of alternative splicing in human. *Genome Biol*. **12,** R9
24. Blakeley, P., Siepen, J. A., Lawless, C., and Hubbard, S. J. (2010) Investigating protein isoforms via proteomics: A feasibility study. *Proteomics* **10,** 1127–1140
25. Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., and Hubbard, T. (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res*. **21,** 756–767
26. Bitton, D. A., Smith, D. L., Connolly, Y., Scutt, P. J., and Miller, C. J. (2010) An Integrated Mass-Spectrometry Pipeline Identifies Novel Protein Coding-Regions in the Human Genome. *Plos One* **5,** e8949
27. Mo, F., Hong, X., Gao, F., Du, L., Wang, J., Omenn, G. S., and Lin, B. Y. (2008) A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. *BMC Bioinformatics* **9,** 537
28. Xing, X. B., Li, Q. R., Sun, H., Fu, X., Zhan, F., Huang, X., Li, J., Chen, C. L., Shyr, Y., Zeng, R., Li, Y. X., and Xie, L. (2011) The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. *Genomics* **98,** 343–351
29. Zhou, A., Zhang, F., and Chen, J. Y. (2010) PEPPI: a peptidomic database of human protein isoforms for proteomics experiments. *BMC Bioinformatics* **11,** S7
30. Chang, K. Y., Georgianna, D. R., Heber, S., Payne, G. A., and Muddiman, D. C. (2010) Detection of Alternative Splice Variants at the Proteome Level in Aspergillus flavus. *J. Proteome Res.* **9,** 1209–1217
31. Ning, K., and Nesvizhskii, A. I. (2010) 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,** S14
32. Lopez-Casado, G., Covey, P. A., Bedinger, P. A., Mueller, L. A., Thannhauser, T. W., Zhang, S., Fei, Z., Giovannoni, J. J., and Rose, J. K. C. (2012) Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case. *Proteomics* **12,** 761–774
33. Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S. P., and Bafna, V. (2007) Improving gene annotation using peptide mass spectrometry. *Genome Res*. **17,** 231–239
34. Chen, J. H., Shi, J. S., Tian, D. G., Yang, L. M., Luo, Y. M., Yin, D. H., and Hu, X. Y. (2011) Improved protein identification using a species- specific protein/peptide database derived from expressed sequence tags. *Plant Omics* **4,** 257–263
35. Power, K. A., McRedmond, J. P., de Stefani, A., Gallagher, W. M., and Gaora, P. O. (2009) High-Throughput Proteomics Detection of Novel Splice Isoforms in Human Platelets. *Plos One* **4,**
36. Edwards, N. J. (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol.* **3,**
37. Yates, J. R., Eng, J. K., and McCormack, A. L. (1995) Mining Genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases. *Anal. Chem.* **67,** 3202–3210
38. Castellana, N. E., Payne, S. H., Shen, Z. X., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 21034–21038
39. Castellana, N. E., Pham, V., Arnott, D., Lill, J. R., and Bafna, V. (2010) Template Proteogenomics: Sequencing Whole Proteins Using an Imperfect Database. *Mol. Cell Proteomics* **9,** 1260–1270
40. Wang, X., Slebos, R. J. C., Wang, D., Halvey, P. J., Tabb, D. L., Liebler, D. C., and Zhang, B. (2011) Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *J. Proteome Res.* 11(2), 1009–1017
41. Castellana, N., and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: A computational perspective. *J. Proteomics* **73,** 2124–2135
42. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data - The protein inference problem. *Mol. Cell Proteomics* **4,** 1419–1440
43. Meyer-Arendt, K., Old, W. M., Houel, S., Renganathan, K., Eichelberger, B., Resing, K. A., and Ahn, N. G. (2011) IsoformResolver: A Peptide-Centric Algorithm for Protein Inference. *J. Proteome Res*. **10,** 3060–3075
44. Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., del Pozo, A., Tress, M., Johnson, R., Guigo, R., and Valencia, A. (2012) Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res.* **22,** 1231–1242
45. Adamidi, C., Wang, Y., Gruen, D., Mastrobuoni, G., You, X., Tolle, D., Dodt, M., Mackowiak, S. D., Gogol-Doering, A., Oenal, P., Rybak, A., Ross, E., Sanchez, Alvarado, A., Kempa, S., Dieterich, C., Rajewsky, N., and Chen, W. (2011) De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Res.* **21,** 1193–1200
46. Evans, V. C., Barker, G., Heesom, K. J., Fan, J., Bessant, C., and Matthews, D. A. (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Meth.* advance online publication
47. Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362
48. Gilar, M., Olivova, P., Daly, A. E., and Gebler, J. C. (2005) Orthogonality of Separation in Two-Dimensional Liquid Chromatography. *Anal. Chem.* **77,** 6426–6434
49. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A., and Soboleva, A. (2011) NCBI GEO: archive for functional genomics data sets-10 years on. *Nucleic Acids Res.* **39,** D1005–D1010
50. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25
51. Trapnell, C., Pachter, L., and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111
52. Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Meth.* **4,** 923–925
53. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) The PeptideAtlas project. *Nucleic Acids Res.* **34,** D655–D658
54. Li, B., and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323
55. Melamud, E., and Moult, J. (2009) Stochastic noise in splicing machinery. *Nucleic Acids Res.* **37,** 4873–4886
56. Hebenstreit, D., Fang, M. Q., Gu, M. X., Charoensawan, V., van Oudenaarden, A., and Teichmann, S. A. (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **7,**
57. Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Meth.* **8,** 469–477
58. Ozsolak, F., and Milos, P. M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12,** 87–98
59. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73,** 2092–2123
60. Blakeley, P., Overton, I. M., and Hubbard, S. J. (2012) Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *J. Proteome Res.* **11,** 5221–5234
61. Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *J. Proteome Res.* **7,** 40–44
62. Choi, H., Ghosh, D., and Nesvizhskii, A. I. (2008) Statistical validation of

peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **7,** 286–292
63. Bradley, R. K., Merkin, J., Lambert, N. J., and Burge, C. B. (2012) Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution. *Plos Biol.* **10,**
64. Hiller, M., and Platzer, M. (2008) Widespread and subtle: alternative splicing at short-distance tandem sites. *Trends Genet.* **24,** 246–255
65. Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.* **36,** 1255–1257
66. Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015
67. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011) The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549
68. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol*. **7,**
69. Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21,** 1543–1551