ProteomeGenerator: A framework for comprehensive proteomics based on de novo transcriptome assembly and high-accuracy peptide mass spectral matching

Paolo Cifani¹, Avantika Dhabaria¹, Akihide Yoshimi², Omar Abdel-Wahab^{2,3}, John T. Poirier^{1,3†}, and Alex Kentsis^{1,4,5†}

¹ Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY.

² Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center New York, NY.

³ Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY.

⁴ Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY.

⁵ Departments of Pediatrics, Pharmacology, and Physiology & Biophysics, Weill Cornell Medical College, Cornell University, New York, NY.

† To whom correspondence should be addressed: Email: poirierj@mskcc.org and kentsisresearchgroup@gmail.com.

RUNNING TITLE

High-accuracy sample-specific proteomics using ProteomeGenerator

ABBREVIATIONS

BCA: bicinchoninic acid; FDR: false discovery rate; LC: liquid chromatography; MS: mass spectrometry; PGX: proteogenomic; PSM: peptide-spectral match; RNA-seq: RNA sequencing; RP: reversed phase; SCX: strong cation exchange.

SUMMARY

Modern mass spectrometry now permits genome-scale and quantitative measurements of biological proteomes. However, analyses of specific specimens are currently hindered by the incomplete representation of biological variability of protein sequences in canonical reference proteomes, and the technical demands for their construction. Here, we report ProteomeGenerator, a framework for *de novo* and reference-assisted proteogenomic database construction and analysis based on sample-specific transcriptome sequencing and high-resolution and high-accuracy mass spectrometry proteomics. This enables assembly of proteomes encoded by actively transcribed genes, including sample-specific protein isoforms resulting from non-canonical mRNA transcription, splicing, or editing. To improve the accuracy of protein isoform identification in non-canonical proteomes, ProteomeGenerator relies on statistical target-decoy database matching augmented with spectral-match calibrated sample-specific controls. We applied this method for the proteogenomic discovery of splicing factor *SRSF2*-mutant leukemia cells, demonstrating high-confidence identification of non-canonical protein isoforms arising from alternative transcriptional start sites, intron retention, and cryptic exon splicing, as well as improved accuracy of genome-scale proteome discovery. Additionally, we report proteogenomic performance metrics for the current state-of-the-art implementations of SEQUEST HT, Proteome Discoverer, MaxQuant, Byonic, and PEAKS mass spectral analysis algorithms. Finally, ProteomeGenerator is implemented as a Snakemake workflow, enabling open, scalable, and facile discovery of sample-specific, non-canonical and neomorphic biological proteomes (<https://github.com/jtpoirier/proteomegenerator>).

INTRODUCTION

Functional analysis of physiologic and pathologic cell activities requires accurate and complete identification and quantification of all involved effector molecules. Such global studies are principally based on the decoding and assembly of the human genome [1,2]. Recent advances in messenger RNA (mRNA) sequencing and bioinformatics now enable routine analysis of biological gene expression [3-5]. However, direct and proteome-wide studies of proteins and their biological variation remain confined to specialized approaches [6-8].

Modern mass spectrometry now permits genome-scale and quantitative measurements of biological proteomes [9-13]. This approach is based on mass spectrometric analysis of peptides, generated by proteolysis of biological proteomes, followed by matching their observed fragmentation spectra to those predicted from known or expected protein sequences [14,15]. Generally, this is accomplished using statistical peptide-spectrum matching techniques that leverage scoring functions to assess the similarity of observed and theoretical mass spectra [16], with the corresponding confidence of spectral identification expressed as a global false discovery rate (FDR), estimated using target-decoy approaches [17].

Peptide identification using peptide-spectrum matching (PSM) and target-decoy FDR estimation is based on the fundamental assumption that mass spectrometry search databases contain a complete and accurate list of all potential protein sequences. Consequently, sensitivity and specificity of peptide identification depend on the fidelity of the target and decoy databases. Advances in genome analysis and

assembly have led to the development of high-quality databases of consensus protein sequences such as UniProt and RefSeq [1,2,18,19]. However, germline and somatic genetic variants, messenger RNA (mRNA) splicing, and other biological processes can diversify polypeptide sequences, thereby generating sequence variants that are not catalogued in the consensus or canonical databases [20,21]. These sources of proteome variation are particularly prevalent in human cancers, which can be caused by structural aberrations in genes and dysregulation of their expression [22-24], ultimately hindering the discovery of cancer proteomes based on reference consensus databases.

In principle, proteogenomic approaches that integrate genome and transcriptome sequencing data with mass spectrometric protein analysis can overcome this limitation, by generating sample-specific target databases for proteomic analysis that more accurately reflect the expressed proteome [25-31]. Such an approach was first introduced to support gene annotation using proteomic data [32], and has since become a powerful tool for quantitative and integrative studies [30,33-38]. In addition, related approaches were recently developed for cancer biology [39-44], and immunology studies [45]. Specifically, sequences of expressed gene transcripts obtained by high-throughput sequencing of mRNA (RNA-seq) are a convenient source for proteogenomic sample-specific database construction for two main reasons: i) these measurements reflect sequence variability introduced by transcriptional and post-transcriptional processes, and ii) restriction of the mass spectral match search space to the specifically expressed proteins can improve its sensitivity and accuracy, particularly as compared to proteins predicted from translation of all possible reading frames [46].

Based on this rationale, several approaches have recently been developed to generate customized mass spectrometry search databases from RNA-seq data [28,44,47,49]. However, sensitivity and accuracy of proteogenomic detection of neomorphic and non-canonical proteins remain limited by the under-sampling of rare peptides and challenges in the generation of sample-specific databases from non-strand-specific short-read RNA-seq data [25,27]. Furthermore, while reference sequence databases such as UniProt are manually curated to improve accuracy [18], automated workflows are needed to enable facile sample-specific proteogenomic analyses at scale.

Here, we describe ProteomeGenerator, an open, modular, and scalable framework for *de novo* and referenced proteogenomic database construction and analysis written in the Snakemake workflow management system. To improve the accuracy of peptide-spectrum matching, we augmented established statistical target-decoy mass spectral matching with spectral-match calibrated sample-specific controls and parameterized this method using four current state-of-the-art mass spectrometry search algorithms. Lastly, we used ProteomeGenerator for genome-scale proteomic discovery of splicing factor-mutant leukemia cells, based on the integration of deep mRNA sequencing, and multidimensional, high-capacity nano-scale chromatography and high-accuracy mass spectrometry. This led to the high-confidence identification of non-canonical protein isoforms arising from alternative transcription start sites, intron retention, and cryptic exon splicing, as well as improved accuracy of genome-scale proteome discovery as compared to conventional approaches.

EXPERIMENTAL PROCEDURES

Reagents. Mass spectrometry grade (Optima LC/MS) water, acetonitrile (ACN), and methanol were from Fisher Scientific (Fair Lawn, NJ). Formic acid of >99% purity (FA) was obtained from Thermo Scientific. All other reagents at MS-grade purity were obtained from Sigma-Aldrich (St. Louis, MO).

Cell culture. Human K052 cells were obtained from the Japanese Collection of Research Bioresources Cell Bank, identity confirmed using STR genotyping (Genetica DNA Laboratories, Burlington, NC), and cultured as described [50]. Cells were collected while in exponential growth phase, washed twice in ice-cold PBS, snap frozen and stored at -80°C. Protein extraction and proteolysis was performed as previously described [51]. Briefly, frozen cell pellets were thawed on ice, resuspended in 6 M guanidinium hydrochloride, 100 mM ammonium bicarbonate at pH 7.6 (ABC), and lysed using the E210 adaptive focused sonicator (Covaris, Woburn, CA). The protein content in cell lysate was determined using the BCA assay, according to the manufacturer's instructions (Pierce, Rockford, IL). Upon reduction and alkylation, proteomes were digested using 1:100 w/w (protease:proteome) LysC endopeptidase (Wako Chemical, Richmond, VA) and 1:50 w/w MS sequencing-grade modified trypsin (Promega, Madison WI). Digestion was stopped by acidifying the reactions to pH 3 using formic acid (Thermo Scientific), and peptides were subsequently desalted using solid phase extraction using C18 Macro Spin columns (Nest Group, Southborough, MA).

mRNA sequencing. RNA was extracted using QIAGEN RNeasy columns (Qiagen, Valencia CA). Poly(A)-selected, unstranded Illumina libraries were prepared

with a modified TruSeq protocol, and 0.5x AMPure XP beads (Beckman Coulter, Indianapolis IA) were added to the sample library to select for fragments of <400 bp, followed by 1x beads to select for fragments of >100 bp. These fragments were then amplified with PCR (15 cycles) and sequenced on the Illumina HiSeq 2000 (100 million 2 x 49 bp reads/sample).

ProteomeGenerator. ProteomeGenerator is written in Snakemake, a scalable, Python-based workflow management system (Figure 2) [52]. The entire workflow is available for download at <https://github.com/jtpoirier/proteomegenerator>. ProteomeGenerator ingests RNAseq data, which is aligned to a reference genome by the STAR splice aware aligner (v. 2.5.2a) [53]. Aligned reads are subsequently filtered to exclude low quality and poorly mapping reads using samtools (v. 1.3) [54]. A sample-specific transcript model is then assembled either *de novo* or with assistance from reference transcript model if one is available using StringTie (v. 1.3.3b) [55], with simultaneous filtering for transcripts with coverage ≥ 2.5 , length ≥ 300 base pairs, and abundance of $\geq 1\%$ of expressed transcripts for a given gene. All resulting transcript models are then merged with StringTie --merge using an expression threshold of 1 transcript per million with permissive intron inclusion. The resulting merged transcript model is then used to generate corresponding cDNA sequences using gffread (v. 0.9.8). The longest uninterrupted open reading frame is detected within each cDNA using TransDecoder (v. 2.1, <https://github.com/TransDecoder>) [56]. Shorter open reading frames with low expect values when searched against the UniProt database using BLAST (v. 2.2.31) are retained [57]. The predicted longest open reading frames are

subsequently mapped back to genomic coordinates and translated to their respective unique peptide sequences.

Peptide and database characterization. Peptides assigned to mass spectra are mapped back to their genomic coordinates using ProteomeGenerator for visualization in the Integrative Genomics Viewer [58]. Unique tryptic peptide search space was calculated for each database using the EMBOSS [59] tool pepdigest and filtered to include all peptides of at least 6 amino acids in length having a molecular weight between 600 and 4000 Daltons.

Databases. Consensus protein sequences databases were downloaded from UniProt [18] as of January 2016 (*Homo sapiens*), September 2015 (*Archaeobacteria loki*) and June 2014 (*Escherichia coli*). Contaminant sequences were retrieved from cRAP [60] as of June 2014.

High-resolution peptide chromatography. Peptide chromatographic fractionation was performed using the Alliance e2695 high-performance liquid chromatograph (Waters, Milford MA). High-pH reverse phase separation was performed using the Xselect CSH 3.0 mm x 150 mm column (Waters, part no. 186006728) at constant flow-rate of 250 μ l/min. After an initial equilibration at 100% buffer A (50 mM ammonium hydroxide in water, pH 10) for 5 minutes, peptides were resolved by a 75 minutes 0-70% gradient of buffer B (80% ACN in water, pH 10), followed by 10 minutes at 100% buffer B. The eluate was collected in 0.5 ml aliquots using a fraction collector between 25 and 75 minutes, lyophilized to dryness in a vacuum centrifuge, and stored at -80°C

until analysis. Before LC-MS analysis, peptides were resuspended in 20 μ l 0.1% formic acid in water, and 2 μ l were analyzed.

Strong cation exchange chromatography was performed using the Xselect Hi Res SP, 7 μ m, 4.6 mm x 100 mm column (Waters, part no. 186004930) at constant flow-rate of 500 μ l/min. After an initial equilibration at 100% buffer A (0.1 % formic acid, 5% ACN) for 5 minutes, peptides were resolved by a 80 minutes 0-30% gradient of buffer B (1 M KCl, 5% ACN), followed by 5 minutes gradient 30-50% buffer B, and a final hold for 5 minutes at 100% buffer B. The eluate was collected in 1 ml aliquots using a fraction collector between 25 and 85 minutes, and lyophilized to dryness in a vacuum centrifuge. Pellets were desalted by solid phase extraction using C18 Macro Spin columns, and stored at -80°C until analysis. Before LC-MS analysis, peptides were resuspended in 20 μ l 0.1% formic acid in water, and 2 μ l were analyzed.

Nanoscale liquid chromatography. Nanoscale liquid chromatography experiments were performed using the Ekspert NanoLC 425 chromatograph (Eksigent, Redwood city, CA), equipped with an autosampler module, two 10-port and one 6-port rotary valves, and one isocratic and two binary pumps. Column and emitter fabrication were performed as previously described [61]. Briefly, samples were initially aspirated into a 10 μ l PEEK sample loop. Chromatographic columns were fabricated by pressure filling the stationary phase into silica capillaries fritted with K-silicate. Reversed phase columns were fabricated by packing Reprosil 1.9 μ m silica C18 particles (Dr. Meisch, Ammerbauch-Entrigen, Germany) into 75 μ m x 40 cm fritted capillaries. Trap columns were fabricated by packing Poros R2 10 μ m C18 particles (Life Technologies, Norwalk,

CT) into 150 μm x 4 cm fritted capillaries. Vented trap-elute architecture was used for chromatography [62]. Peptides were resolved by reversed phase chromatography hyphenated to the nano-electrospray ion source. Upon valve switch to connect the trap column in line with the analytical reversed phase column and ion emitter, the pressure was equilibrated at a flow of 250 nl/min for 5 minutes in 5% buffer B (ACN, 0.1% FA) in buffer A (water, 0.1% FA). Subsequently, a 120-minutes (high-pH reverse phase samples) or 180 minutes (SCX samples) linear gradient of 5-40% of buffer B was used to resolve peptides, followed by a 5 minutes 40-80% gradient prior to column wash at 80% buffer B for 30 minutes.

Nano-electrospray ionization and Orbitrap mass spectrometry. Electrospray emitters with terminal opening diameter of 10 μm were obtained from New Objective (Woburn, MA). The emitter was connected to the outlet of the reversed phase column using a metal union that also served as the electrospray current electrode. Electrospray ionization was achieved using constant 1700 V voltage. During column loading, the electrospray emitter was washed with 50% aqueous methanol using the DPV-565 PicoView ion source (New Objective).

For all measurements, we used the Orbitrap Fusion mass spectrometer (Thermo Scientific, San Jose, CA). Precursor scans in the 400-2000 Th were performed in the orbitrap detector at 120,000 resolution, with 100 ms maximum injection time, and automatic gain control set at 10^5 ions. Fragment spectra were recorded in the linear ion trap in rapid mode, with maximum injection time of 75 ms and target of 10^4 ions, and 1.2 Da quadrupolar precursor selection.

Data Analysis. Peptide-spectral matching calculations were performed using a custom-built computer server equipped with 4 Intel Xeon E5-4620 8-core CPUs operating at 2.2 GHz, and 512 GB physical memory (Exxact Corporation, Fremont, CA). Peptide-spectral matching was performed using SEQUEST HT and Percolator [16,63,64] as part of Proteome Discoverer 2.1.0.81 (Thermo Scientific), Byonic 2.7.84 [65,66], MaxQuant 1.5.4.1 [67,68], and PEAKS 8.0 [69]. For all searches, MS1 and HCD MS2 mass tolerances were set to 10 ppm and 0.6 Da, respectively. Cysteine carbamidomethylation was set as fixed, while methionine oxidation and glutamine and asparagine deamidation were set as variable modifications, with a maximum of 3 variable modifications per peptide. Only peptides containing 7-35 residues and up to 2 missed trypsin cleavages were considered. Tryptic peptides observable by mass spectrometry were predicted based on proteome composition using the generate-peptides utility of Crux [70].

Data availability: ProteomeGenerator is openly available at <https://github.com/jtpoirier/proteomegenerator>. RNA sequencing data are available from the Gene Expression Omnibus. Mass spectrometry raw files and search results are publicly available through the ProteomeXchange data repository with identifiers PXD008381 and PXD008463 [71].

RESULTS

ProteomeGenerator for deep genome-scale transcriptomic and proteomic integration

Consensus reference proteomes provide a useful set of canonical protein sequences to be used for statistical peptide-spectral matching in mass spectrometry proteomics. However, these databases do not fully recapitulate the natural and pathologic variation in protein sequences, therefore limiting the sensitivity and specificity of proteome discovery. To enable facile construction of sample-specific proteomic databases containing expressed non-canonical protein variants, we developed a proteogenomic framework, termed ProteomeGenerator, programmed in the Snakemake workflow management system for open and scalable analysis. This method i) generates a sample-specific protein sequence database based on de novo or referenced transcriptome assembly using high-coverage RNA-seq data, ii) deep mass spectrometric analysis of the matched proteome using high-resolution nanoscale chromatography and high-accuracy mass spectrometry, and iii) accurate protein isoform identification using statistical target-decoy database matching augmented with spectrally calibrated sample-specific controls (Figure 1).

Figure 1. ProteomeGenerator overview. Transcriptomes and proteomes from the same biologic sample are analyzed in parallel by high-coverage Illumina sequencing and high-resolution, high-accuracy mass spectrometry, respectively. ProteomeGenerator assembles fastq-formatted mRNA sequencing reads into predicted transcripts, identifies reading frames and isoforms, and produces fasta-formatted proteogenomic (PGX) databases containing canonical and non-canonical expressed protein isoforms for subsequent mass spectrometry searches.

We applied this method to analyze the human K052 leukemia cells harboring splicing factor *SRSF2* mutations, which were recently described to cause recurrent mRNA mis-splicing, and are therefore expected to express non-canonical and neomorphic protein isoforms [50]. Thus, we extracted mRNA and obtained high-coverage RNA-seq data of nearly 60 million reads by Illumina sequencing. We then used ProteomeGenerator to process the raw sequencing reads in the following steps: i) two-pass splice aware alignment to the user-supplied reference genome, in this case GRCh38; ii) assembly of possible transcript isoforms using StringTie either de novo or assisted by the user-supplied transcript model, in this case GENCODE v20; iii) prediction of the longest open reading frame from for each possible transcript; and iv) generation of fasta-formatted proteogenomic database comprising unique protein sequences for subsequent peptide sequence-mass spectral matching (Figure 2).

Figure 2. Schema for the ProteomeGenerator snakemake workflow. Sequencing reads are aligned using STAR, followed by their de novo and referenced assembly into transcriptomes using StringTie, and processing to identify reading frames and protein isoforms.

In the case of analyzed K052 cells, the constructed proteogenomic database, to which we refer as PGX, consisted of 17,348 protein entries, expected to produce 743,148 observable tryptic peptides with lengths between 7 and 35 residues, assuming

1 missed trypsin cleavage (Figure 3). For comparison, the canonical reference human proteome in the UniProt database (as of March 2016), contained 42,123 proteins corresponding to 1,460,257 mass spectrometry observable tryptic peptides. The PGX database contained 37,158 peptides with no counterparts in UniProt, presumably originating from novel predicted protein isoforms specific for splicing factor-mutant K052 cells and consequently not annotated in UniProt. In addition, the PGX database was approximately 51% smaller with respect to mass spectrometry observable peptides as compared to UniProt, presumably because it contained only expressed protein isoforms, in contrast to the canonical reference proteomes, which contain the protein complement of all canonical protein isoforms.

Figure 3. Comparison of the canonical and proteogenomic protein databases, displaying (A) number of protein entries, (B) number of amino acid residues, and (C) theoretical tryptic peptides amenable for mass spectrometry analysis specific for either UniProt, PGX, or common to both.

Genome-scale mass spectrometry proteomics

To generate mass spectrometry data of sufficient coverage to sample genome-scale proteomes, we used recently developed high-resolution, multi-dimensional

nanoscale chromatography and high-accuracy mass Orbitrap spectrometry. To increase chromatographic resolution, we leveraged the orthogonal properties of strong-cation exchange and alkaline reversed phase as compared to acidic reversed phase chromatography, thereby enabling the detection of low abundance and rare peptides and their isoforms [72]. Each peptide fraction was then resolved by high-resolution nano-scale reverse phase chromatography and analyzed by nanoelectrospray ionization mass spectrometry, and nearly 3 million high-resolution precursor and fragmentation spectra were recorded.

To assess sampling efficiency, we used statistical database matching against UniProt to identify unique peptides and proteins at 1% FDR. Using a sub-sampling analysis, we observed that this strategy indeed maximized the sensitivity of detection of canonical proteins, though peptide sampling was apparently incompletely saturated (Supplementary Figure 1A).

Supplementary Figure 1. Subsampling analysis of mass spectrometric (A) RNA sequencing (B) dataset. Plateauing of the number of sequences detected indicates reaching of the limit of detection for the specific analytical method used, as observed in transcriptomic (B) but not in proteomic (A) datasets.

Likewise, we found that high-coverage mRNA sequencing apparently saturated sampling of unique sequence reads obtained by transcriptomic analysis (Supplementary Figure 1B). Thus, high-resolution chromatography coupled with high-accuracy mass spectrometry and high-coverage transcriptome sequencing is suitable for genome-scale proteogenomics.

Accurate proteome discovery using statistical target-decoy matching with spectral-match calibration

Having obtained genome-scale mass spectrometry data and the transcriptome-assembled PGX database based on high-coverage mRNA sequencing, we next sought to identify algorithms for scoring peptide-spectral matches and estimating FDR confidence suitable for proteogenomic analysis. Such algorithms should ideally not only maximize sensitivity (i.e. the fraction of identified spectra), but also ensure high specificity, in particular when searching a non-curated proteogenomic target database potentially containing erroneous sequences. To calibrate the sensitivity and specificity of these algorithms, we introduced two negative controls [73,74]: i) a set of mass spectra from a non-human proteome (*E. coli*) recorded under identical experimental conditions as the experimental human proteome, and ii) a set of protein sequences from evolutionarily divergent non-human species with minimal identity to the experimental human proteome, in this case the recently published proteome of *A. loki* archaeobacteria [75], as confirmed by direct comparisons (Figure 4A-B).

For mass spectrometry search algorithms, we used four current state-of-the-art programs, chosen for their distinct methods for candidate sequence selection and FDR

estimation: SEQUEST HT with Percolator as part of Proteome Discoverer [16,63], Byonic [65,66], MaxQuant [67,68], and PEAKS [69]. For benchmarking purposes, we searched the experimental human K052 and negative control bacterial *E. coli* mass spectra against a concatenated database containing PGX and UniProt human databases, supplemented with negative control archaeobacterial *A. loki* proteins and the common contaminant cRAP sequences. All searches were performed using identical search parameters at the FDR of less than 0.01. We assessed sensitivity based on the number of *bona fide* correct identifications, mapping to either human or common contaminant databases (Figure 4C, Supplementary table 1).

Figure 4. Sensitivity and specificity of mass spectrometry search algorithms. (A, B) Comparison of unique theoretical peptides in the experimental PGX proteome, canonical UniProt, and bacterial proteomes used as negative controls. (C) Sensitivity of tested algorithms, expressed as the number of identified peptides. (D) Specificity of tested algorithms, evaluated from the fraction of peptide-spectrum matches mapped to the negative controls. The PSM fraction mapped to *A. loki* is reported both in absolute terms (black) and normalized to take into account the relative sizes of the human and archaeobacterial proteomes (grey).

We estimated specificity from the fraction of peptides identified from *E. coli* spectra matched to human protein sequences, and experimental human spectra matched to archaeobacterial *A. loki* sequences (Figure 4D). To empirically estimate FDR, we normalized the number of observed peptide-spectrum matches to account for the differences in size between human and archaeobacterial proteomes.

We found that algorithms that select candidate sequences based on *de novo* sequence tags, such as Byonic and PEAKS, have a lower probability to match spectra to incorrect sequences in the database, and are thus preferred for proteogenomic analyses. We observed that with the exception of Byonic, all algorithms had comparable sensitivity. However, PEAKS also exhibited the lowest rate of incorrectly matching *E. coli* spectra to human sequences, and thus had the highest overall specificity among the compared algorithms. This finding, combined with its high sensitivity, led us to select PEAKS for integration into ProteomeGenerator as the algorithm for peptide-spectrum matching and FDR control.

Identification of non-canonical protein isoforms using ProteomeGenerator

Compelled by the high quality of the ProteomeGenerator-constructed PGX database and genome-scale mass spectrometry, and high sensitivity and specificity of mass spectral identification by PEAKS, we used this approach to match the experimental spectra from K052 cells against either PGX or UniProt databases at the FDR of less than 0.01. Based on the same set of 2,736,597 fragmentation spectra, we observed 611,275 and 621,960 peptide-spectrum matches (22% PSM rate) when

searching PGX and UniProt databases, respectively (Figure 5A, Supplementary Tables 2-3).

Figure 5. Accurate proteome discovery using statistical target-decoy matching with spectral calibration. (A) Number of peptides identified (FDR<0.01) based on matching spectra from K052 proteome against proteogenomic (PGX, red) and canonical (UniProt, grey) databases. (B) Overlap between the peptides identified in PGX (red) and UniProt (grey) databases. (C) Comparison of PEAKS scores for peptides identified in both PGX and UniProt databases. (D) PEAKS score distribution for peptides identified exclusively in PGX (red) and UniProt (grey) databases. (E) For peptides exclusively identified against the PGX database, PEAKS score distributions for peptides not mapping in UniProt (red) or present in the canonical database (grey). Boxes delimit 25th and 75th percentiles, middle line corresponds to median, whiskers correspond to 5th and 95th percentiles. (F) PEAKS score distributions for peptides identified exclusively in PGX but also mapping in UniProt (grey), or exclusively mapping in the PGX database (red).

Most of these PSMs defined peptide sequences that were shared between the two databases with 97% and 94% of the total PGX and UniProt identifications, respectively (Figure 5B), and with nearly identical confidence of identification (Figure 5C). We observed peptides uniquely identified when searching against UniProt, due to either their incomplete representation in the PGX database because of limited mRNA sequencing sensitivity, as previously described in proteogenomic analysis of HeLa cells [13], or because the greater diversity of sequences in UniProt can result in incorrect matching of homeometric peptides [76]. Importantly, the reduced size of the sample-specific PGX database led an effectively lower PSM score threshold as compared to searches against Uniprot. As a result, at identical global FDR, searches of the ProteomeGenerator-constructed PGX database were effectively more sensitive (Figure 5D-F).

We next assessed the ability of ProteomeGenerator to identify non-canonical protein isoforms. To this end, we analyzed the subset of peptides with sequences not mapping in UniProt, as prioritized based on the apparent statistical confidence of their identifications, with PEAKS peptide-spectrum matching scores greater than 50. Analysis of these sequences using BLAST indicated that the majority of them (94%) mapped to isoforms annotated in the non-reviewed section of UniProt or RefSeq (Supplementary Table 4). Notably, we identified peptides corresponding to previously un-annotated isoforms of APH (*APEH*), YB-1 (*YBX1*), and MUNC13D (*UNC13D*). These isoforms were found to be consistent with alternative splicing and intron retention for APEH and MUNC13D (Figure 6AB, Supplementary Figure 2-3).

Figure 6. Identification of non-canonical protein isoforms using ProteomeGenerator. (A) Genome tracks of non-canonical APEH isoform generation via inclusion of an intronic sequence normally spliced in the canonical APEH isoform. (B) The K052-specific isoform of APEH contains a novel N-terminal sequence, with the splicing junction encompassed by peptide AGPDPGVSPAQVLLSEPEEAAALYR. Residues 35-276 of the protein sequences defined by ProteomeGenerator are identical to residues 4-245 of the canonical UniProt protein sequence. (C) Fragmentation spectrum of the peptide encompassing the novel splice junction, with diagnostic fragment ions and amino acid residues labeled. Italicized ion labels indicate ions with relative intensity above 25% of the maximum. A star (*) symbol denotes internal ions. (D) Peptide abundance as based on total fragment ion current for all identified APEH peptides (red: peptides from K052-specific sequence; grey: peptides from canonical sequence).

We confirmed these identifications by manual inspection of the mass fragmentation spectra (Figure 6C, Supplementary Figure 2).

In the case of APEH, the non-canonical N-terminal domain produced by intron retention was detected by mass spectrometry from two unique and independent peptides, including one spanning the non-canonical splice junction. Since no peptides corresponding to this region were detected from the canonical isoform, we used the total ion current of the fragmentation spectrum associated with this PSM to estimate the differential abundance of the novel and canonical APEH protein isoforms (Figure 6D). This analysis revealed that the identified non-canonical APEH isoform represents the majority of cellular APEH, identified specifically by ProteomeGenerator.

DISCUSSION

Here, we introduced an analytical framework for scalable *de novo* and reference-guided assembly of sample-specific proteomic databases based on mRNA sequencing, as combined with spectral calibration high-resolution, high-accuracy mass spectrometry for discovery of non-canonical biological proteomes. We designed ProteomeGenerator to produce sample-specific proteome search databases containing only proteins that are predicted to be expressed. As a result, PGX databases have markedly reduced search space as compared to canonical reference databases such as UniProt, and consequently, exhibit more accurate control of potential false discovery by statistical database matching techniques. Since currently no methods exist for *a priori* definition of complete transcriptomes and proteomes, we confirmed the accuracy of ProteomeGenerator based on three measures: i) the set of peptides identified using ProteomeGenerator largely overlaps with that obtained by conventional matching against consensus human proteomes; ii) the novel non-canonical peptides identified by ProteomeGenerator were confirmed by high-confidence observations both at proteomic and transcriptomic levels, or were annotated in other non-reviewed or provisional shotgun databases; and iii) the accuracy of peptide-spectral matching was confirmed by stringent benchmarking of scoring functions and their calibration using spectral and sequence negative controls.

ProteomeGenerator and related approaches are limited by the accuracy of transcriptome assembly, which is dependent on mRNA sequencing quality, depth, length, and strand specificity. We anticipate that ProteomeGenerator will benefit from increasing adoption of long-read and strand-specific RNA sequencing technologies. To

this end, it will also be important to further improve the sensitivity of proteome sampling by peptide mass spectrometry, such as optimizing proteome proteolysis, its chromatographic resolution, and methods for *de novo* mass spectral identification [77,78]. Lastly, ProteomeGenerator is implemented as a Snakemake workflow, enabling open, scalable, and facile discovery of sample-specific, non-canonical and neomorphic biological proteomes. This should facilitate the discovery of natural variation in cellular and tissue proteomes, which can contribute to normal tissue development and its dysregulation in human disease such as cancer.

Author Contributions: P.C., J.T. and A.K. designed the study, JT. wrote the ProteomeGenerator computer scripts, A.Y. and O.AW. performed mRNA sequencing, P.C. and A.D. performed proteomic analysis, P.C., J.T. and A.K. performed data analyses, P.C., J.T. and A.K. wrote the manuscript with contributions from other authors.

Acknowledgements: We thank John Phillips for constructive discussion, Nathaniel Kwok for assistance with computer scripting, Wilfred Tang for assistance with Byonic searches, and Robert Holmes for computer server support.

Funding: This work was supported by the NIH R21 CA188881, R01 CA204396, P30 CA008748, Burroughs Wellcome Fund, Josie Robertson Investigator Program, Rita

Allen Foundation, Alex's Lemonade Stand Foundation, American Society of Hematology, and Gabrielle's Angel Foundation (A.K.), and American Italian Cancer Foundation (P.C.). A.Y. is supported by grants from the Aplastic Anemia and MDS International Foundation and the Lauri Strauss Leukemia Foundation. A.K. is the Damon Runyon-Richard Lumsden Foundation Clinical Investigator.

This article contains Supplemental Material.

Competing Financial Interests: The authors have no competing financial interests.

REFERENCES

- [1] E. S. Lander, L. M. Linton, *et al.* International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome.," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.
- [2] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome.," *Nature*, vol. 431, no. 7011, pp. 931–945, Oct. 2004.
- [3] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq.," *Nat. Methods*, vol. 5, no. 7, pp. 621–628, Jul. 2008.
- [4] A. Shlien, K. Raine, F. Fuligni, R. Arnold, S. Nik-Zainal, S. Dronov, L. Mamanova, A. Rosic, Y. S. Ju, S. L. Cooke, M. Ramakrishna, E. Papaemmanuil, H. R. Davies, P. S. Tarpey, P. Van Loo, D. C. Wedge, D. R. Jones, S. Martin, J. Marshall, E. Anderson, C. Hardy, ICGC Breast Cancer Working Group, Oslo Breast Cancer Research Consortium, V. Barbashina, S. A. J. R. Aparicio, T. Sauer, Ø. Garred, A. Vincent-Salomon, O. Mariani, S. Boyault, A. Fatima, A. Langerød, Å. Borg, G. Thomas, A. L. Richardson, A.-L. Børresen-Dale, K. Polyak, M. R. Stratton, and P. J. Campbell, "Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer.," *Cell Rep*, vol. 16, no. 7, pp. 2032–2046, Aug. 2016.
- [5] P. G. Engström, T. Steijger, B. Sipos, G. R. Grant, A. Kahles, G. Rätsch, N. Goldman, T. J. Hubbard, J. Harrow, R. Guigó, P. Bertone, RGASP Consortium, "Systematic evaluation of spliced alignment programs for RNA-seq data.," *Nat. Methods*, vol. 10, no. 12, pp. 1185–1191, Dec. 2013.
- [6] A. M. Edwards, R. Isserlin, G. D. Bader, S. V. Frye, T. M. Willson, and F. H. Yu, "Too many roads not taken.," *Nature*, vol. 470, no. 7333, pp. 163–165, Feb. 2011.
- [7] A. I. Nesvizhskii, "Proteogenomics: concepts, applications and computational strategies.," *Nat. Methods*, vol. 11, no. 11, pp. 1114–1125, Nov. 2014.
- [8] K. V. Ruggles, K. Krug, X. Wang, K. R. Clauser, J. Wang, S. H. Payne, D. Fenyo, B. Zhang, and D. R. Mani, "Methods, Tools and Current Perspectives in Proteogenomics.," *Mol. Cell Proteomics*, vol. 16, no. 6, pp. 959–981, Jun. 2017.
- [9] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics.," *Nature*, vol. 422, no. 6928, pp. 198–207, Mar. 2003.
- [10] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster, "Mass-spectrometry-based draft of the human proteome.," *Nature*, vol. 509, no. 7502, pp. 582–587, May 2014.
- [11] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The one hour yeast proteome.," *Molecular & Cellular Proteomics*, vol. 13, no. 1, pp. 339–347, Jan. 2014.
- [12] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Fröhlich, T. C. Walther, and M. Mann, "Comprehensive mass-spectrometry-based

- proteome quantification of haploid versus diploid yeast.,” *Nature*, vol. 455, no. 7217, pp. 1251–1254, Oct. 2008.
- [13] N. Nagaraj, J. R. Wiśniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo, and M. Mann, “Deep proteome and transcriptome mapping of a human cancer cell line.,” *Mol. Syst. Biol.*, vol. 7, no. 1, p. 548, Nov. 2011.
- [14] M. Mann and M. Wilm, “Error-tolerant identification of peptides in sequence databases by peptide sequence tags.,” *Anal. Chem.*, vol. 66, no. 24, pp. 4390–4399, Dec. 1994.
- [15] A. Shevchenko, M. Wilm, O. Vorm, and M. Mann, “Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels.,” *Anal. Chem.*, vol. 68, no. 5, pp. 850–858, Mar. 1996.
- [16] J. K. Eng, A. L. McCormack, and J. R. Yates, “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.,” *J Am Soc Mass Spectrom*, vol. 5, no. 11, pp. 976–989, Nov. 1994.
- [17] J. E. Elias and S. P. Gygi, “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.,” *Nat. Methods*, vol. 4, no. 3, pp. 207–214, Mar. 2007.
- [18] The UniProt Consortium, “UniProt: the universal protein knowledgebase.,” *Nucleic Acids Res.*, vol. 45, no. 1, pp. D158–D169, Jan. 2017.
- [19] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M.-M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, “The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.,” *Genome Res.*, vol. 19, no. 7, pp. 1316–1323, Jul. 2009.
- [20] S. Samandi, A. V. Roy, V. Delcourt, J.-F. Lucier, J. Gagnon, M. C. Beaudoin, B. Vanderperre, M.-A. Breton, J. Motard, J.-F. Jacques, M. Brunelle, I. Gagnon-Arsenault, I. Fournier, A. Ouangraoua, D. J. Hunting, A. A. Cohen, C. R. Landry, M. S. Scott, and X. Roucou, “Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins.,” *Elife*, vol. 6, p. e27860, Oct. 2017.
- [21] J.-P. Couso and P. Patraquim, “Classification and function of small open reading frames.,” *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 9, pp. 575–589, Sep. 2017.
- [22] H. Dvinge and R. K. Bradley, “Widespread intron retention diversifies most cancer transcriptomes.,” *Genome Med*, vol. 7, no. 1, p. 45, 2015.
- [23] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis, “Genomic instability--an evolving hallmark of cancer.,” *Nat. Rev. Mol. Cell Biol.*, vol. 11, no. 3, pp. 220–228, Mar. 2010.
- [24] C. DeBoever, E. M. Ghia, P. J. Shepard, L. Rassenti, C. L. Barrett, K. Jepsen, C. H. M. Jamieson, D. Carson, T. J. Kipps, and K. A. Frazer, “Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers.,” *PLoS Comput. Biol.*, vol. 11, no. 3, p. e1004105, Mar.

- 2015.
- [25] K. Krug, S. Popic, A. Carpy, C. Taumer, and B. Macek, "Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants.," *Proteomics*, vol. 14, no. 23, pp. 2699–2708, Dec. 2014.
- [26] Y. Li, X. Wang, J.-H. Cho, T. I. Shaw, Z. Wu, B. Bai, H. Wang, S. Zhou, T. G. Beach, G. Wu, J. Zhang, and J. Peng, "JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells.," *J. Proteome Res.*, vol. 15, no. 7, pp. 2309–2320, Jul. 2016.
- [27] K. V. Ruggles, Z. Tang, X. Wang, H. Grover, M. Askenazi, J. Teubl, S. Cao, M. D. McLellan, K. R. Clauser, D. L. Tabb, P. Mertins, R. Slebos, P. Erdmann-Gilmore, S. Li, H. P. Gunawardena, L. Xie, T. Liu, J.-Y. Zhou, S. Sun, K. A. Hoadley, C. M. Perou, X. Chen, S. R. Davies, C. A. Maher, C. R. Kinsinger, K. D. Rodland, H. Zhang, Z. Zhang, L. Ding, R. R. Townsend, H. Rodriguez, D. Chan, R. D. Smith, D. C. Liebler, S. A. Carr, S. Payne, M. J. Ellis, and D. Fenyó, "An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer.," *Mol. Cell Proteomics*, vol. 15, no. 3, pp. 1060–1071, Mar. 2016.
- [28] X. Wang and B. Zhang, "customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search.," *Bioinformatics*, vol. 29, no. 24, pp. 3235–3237, Dec. 2013.
- [29] B. Wen, S. Xu, R. Zhou, B. Zhang, X. Wang, X. Liu, X. Xu, and S. Liu, "PGA: an R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq.," *BMC Bioinformatics*, vol. 17, no. 1, p. 244, Jun. 2016.
- [30] S. Woo, S. W. Cha, G. Merrihew, Y. He, N. Castellana, C. Guest, M. MacCoss, and V. Bafna, "Proteogenomic database construction driven from large scale RNA-seq data.," *J. Proteome Res.*, vol. 13, no. 1, pp. 21–28, Jan. 2014.
- [31] F. Zickmann and B. Y. Renard, "MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms.," *Bioinformatics*, vol. 31, no. 12, pp. i106–115, Jun. 2015.
- [32] J. D. Jaffe, H. C. Berg, and G. M. Church, "Proteogenomic mapping as a complementary method to perform genome annotation.," *Proteomics*, vol. 4, no. 1, pp. 59–77, Jan. 2004.
- [33] V. C. Evans, G. Barker, K. J. Heesom, J. Fan, C. Bessant, and D. A. Matthews, "De novo derivation of proteomes from transcriptomes for transcript and protein identification.," *Nat. Methods*, vol. 9, no. 12, pp. 1207–1211, Dec. 2012.
- [34] Y. Liu, M. González-Porta, S. Santos, A. Brazma, J. C. Marioni, R. Aebersold, A. R. Venkitaraman, and V. O. Wickramasinghe, "Impact of Alternative Splicing on the Human Proteome.," *Cell Rep*, vol. 20, no. 5, pp. 1229–1241, Aug. 2017.
- [35] N. E. Castellana, S. H. Payne, Z. Shen, M. Stanke, V. Bafna, and S. P. Briggs, "Discovery and revision of Arabidopsis genes by proteogenomics.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 52, pp. 21034–21038, Dec. 2008.
- [36] U. Omasits, M. Quebatte, D. J. Stekhoven, C. Fortes, B. Roschitzki, M. D. Robinson, C. Dehio, and C. H. Ahrens, "Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome.,"

- Genome Res.*, vol. 23, no. 11, pp. 1916–1927, Nov. 2013.
- [37] G. Menschaert, W. Van Criekeing, T. Notelaers, A. Koch, J. Crappé, K. Gevaert, and P. Van Damme, “Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events,” *Mol. Cell Proteomics*, vol. 12, no. 7, pp. 1780–1790, Jul. 2013.
 - [38] J. Khatun, Y. Yu, J. A. Wrobel, B. A. Risk, H. P. Gunawardena, A. Secrest, W. J. Spitzer, L. Xie, L. Wang, X. Chen, and M. C. Giddings, “Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions,” *BMC Genomics*, vol. 14, no. 1, p. 141, Feb. 2013.
 - [39] H. Zhang, T. Liu, Z. Zhang, S. H. Payne, B. Zhang, J. E. McDermott, J.-Y. Zhou, V. A. Petyuk, L. Chen, D. Ray, S. Sun, F. Yang, L. Chen, J. Wang, P. Shah, S. W. Cha, P. Aiyetan, S. Woo, Y. Tian, M. A. Gritsenko, T. R. Clauss, C. Choi, M. E. Monroe, S. Thomas, S. Nie, C. Wu, R. J. Moore, K.-H. Yu, D. L. Tabb, D. Fenyo, V. Bafna, Y. Wang, H. Rodriguez, E. S. Boja, T. Hiltke, R. C. Rivers, L. Sokoll, H. Zhu, I.-M. Shih, L. Cope, A. Pandey, B. Zhang, M. P. Snyder, D. A. Levine, R. D. Smith, D. W. Chan, K. D. Rodland, CPTAC Investigators, “Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer,” *Cell*, vol. 166, no. 3, pp. 755–765, Jul. 2016.
 - [40] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. C. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. C. Slebos, D. C. Liebler, NCI CPTAC, “Proteogenomic characterization of human colon and rectal cancer,” *Nature*, vol. 513, no. 7518, pp. 382–387, Sep. 2014.
 - [41] P. Mertins, D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, P. Wang, X. Wang, J. W. Qiao, S. Cao, F. Petralia, E. Kawaler, F. Mundt, K. Krug, Z. Tu, J. T. Lei, M. L. Gatz, M. Wilkerson, C. M. Perou, V. Yellapantula, K.-L. Huang, C. Lin, M. D. McLellan, P. Yan, S. R. Davies, R. R. Townsend, S. J. Skates, J. Wang, B. Zhang, C. R. Kinsinger, M. Mesri, H. Rodriguez, L. Ding, A. G. Paulovich, D. Fenyo, M. J. Ellis, S. A. Carr, NCI CPTAC, “Proteogenomics connects somatic mutations to signalling in breast cancer,” *Nature*, vol. 534, no. 7605, pp. 55–62, Jun. 2016.
 - [42] D. C. M. Rolland, V. Basrur, Y.-K. Jeon, C. McNeil-Schwalm, D. Fermin, K. P. Conlon, Y. Zhou, S. Y. Ng, C.-C. Tsou, N. A. Brown, D. G. Thomas, N. G. Bailey, G. S. Omenn, A. I. Nesvizhskii, D. E. Root, D. M. Weinstock, R. B. Faryabi, M. S. Lim, and K. S. J. Elenitoba-Johnson, “Functional proteogenomics reveals biomarkers and therapeutic targets in lymphomas,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, no. 25, pp. 6581–6586, Jun. 2017.
 - [43] R. M. M. Branca, L. M. Orre, H. J. Johansson, V. Granholm, M. Huss, Å. Pérez-Bercoff, J. Forshed, L. Käll, and J. Lehtiö, “HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics,” *Nat. Methods*, vol. 11, no. 1, pp. 59–62, Jan. 2014.
 - [44] M. A. Komor, T. Pham, A. C. Hiemstra, S. R. Piersma, A. S. Bolijn, T. Schelfhorst, P. M. Delis-van Diemen, M. Tijssen, R. P. Sebra, M. Ashby, G. A. Meijer, C. R. Jimenez, and R. J. A. Fijneman, “Identification of differentially

- expressed splice variants by the proteogenomic pipeline Splicify.,” *Mol. Cell Proteomics*, Jul. 2017.
- [45] C. M. Laumont, T. Daouda, J.-P. Laverdure, É. Bonneil, O. Caron-Lizotte, M.-P. Hardy, D. P. Granados, C. Durette, S. Lemieux, P. Thibault, and C. Perreault, “Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames.,” *Nat Commun*, vol. 7, p. 10238, Jan. 2016.
- [46] N. C. Soares, P. Spät, K. Krug, and B. Macek, “Global dynamics of the Escherichia coli proteome and phosphoproteome during growth in minimal medium.,” *J. Proteome Res.*, vol. 12, no. 6, pp. 2611–2621, Jun. 2013.
- [47] H. Park, J. Bae, H. Kim, S. Kim, H. Kim, D.-G. Mun, Y. Joh, W. Lee, S. Chae, S. Lee, H. K. Kim, D. Hwang, S.-W. Lee, and E. Paek, “Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses.,” *Proteomics*, vol. 14, no. 23, pp. 2742–2749, Dec. 2014.
- [48] X. Wang, R. J. C. Slebos, D. Wang, P. J. Halvey, D. L. Tabb, D. C. Liebler, and B. Zhang, “Protein identification using customized protein sequence databases derived from RNA-Seq data.,” *J. Proteome Res.*, vol. 11, no. 2, pp. 1009–1017, Feb. 2012.
- [49] A. J. Cesnik, M. R. Shortreed, G. M. Sheynkman, B. L. Frey, and L. M. Smith, “Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy.,” *J. Proteome Res.*, vol. 15, no. 3, pp. 800–808, Mar. 2016.
- [50] E. Kim, J. O. Ilagan, Y. Liang, G. M. Daubner, S. C.-W. Lee, A. Ramakrishnan, Y. Li, Y. R. Chung, J.-B. Micol, M. E. Murphy, H. Cho, M.-K. Kim, A. S. Zebari, S. Aumann, C. Y. Park, S. Buonomici, P. G. Smith, H. J. Deeg, C. Lobry, I. Aifantis, Y. Modis, F. H.-T. Allain, S. Halene, R. K. Bradley, and O. Abdel-Wahab, “SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition.,” *Cancer Cell*, vol. 27, no. 5, pp. 617–630, May 2015.
- [51] A. Dhabaria, P. Cifani, C. Reed, H. Steen, and A. Kentsis, “A High-Efficiency Cellular Extraction System for Biological Proteomics.,” *J. Proteome Res.*, vol. 14, no. 8, pp. 3403–3408, Aug. 2015.
- [52] J. Köster and S. Rahmann, “Snakemake--a scalable bioinformatics workflow engine.,” *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, Oct. 2012.
- [53] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, “STAR: ultrafast universal RNA-seq aligner.,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.
- [54] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, “The Sequence Alignment/Map format and SAMtools.,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.
- [55] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.,” *Nat. Biotechnol.*, vol. 33, no. 3, pp. 290–295, Mar. 2015.
- [56] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden,

- M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, and A. Regev, "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.," *Nat Protoc*, vol. 8, no. 8, pp. 1494–1512, Aug. 2013.
- [57] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
- [58] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer.," *Nat. Biotechnol.*, vol. 29, no. 1, pp. 24–26, Jan. 2011.
- [59] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite.," *Trends Genet.*, vol. 16, no. 6, pp. 276–277, Jun. 2000.
- [60] D. Mellacheruvu, Z. Wright, A. L. Couzens, J.-P. Lambert, N. A. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardi, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. Al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z.-Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. R. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A.-C. Gingras, and A. I. Nesvizhskii, "The CRAPome: a contaminant repository for affinity purification-mass spectrometry data.," *Nat. Methods*, vol. 10, no. 8, pp. 730–736, Aug. 2013.
- [61] P. Cifani and A. Kentsis, "High sensitivity quantitative proteomics using automated multidimensional nano-flow chromatography and accumulated ion monitoring on quadrupole-Orbitrap-linear ion trap mass spectrometer.," *Mol. Cell Proteomics*, vol. 16, no. 11, pp. 2006–2016, Aug. 2017.
- [62] S. B. Ficarro, Y. Zhang, M. J. Carrasco-Alfonso, B. Garg, G. Adelmant, J. T. Webber, C. J. Luckey, and J. A. Marto, "Online nanoflow multidimensional fractionation for high efficiency phosphopeptide analysis.," *Mol. Cell Proteomics*, vol. 10, no. 11, p. O111.011064, Nov. 2011.
- [63] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss, "Semi-supervised learning for peptide identification from shotgun proteomics datasets.," *Nat. Methods*, vol. 4, no. 11, pp. 923–925, Nov. 2007.
- [64] M. Spivak, J. Weston, L. Bottou, L. Käll, and W. S. Noble, "Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets.," *J. Proteome Res.*, vol. 8, no. 7, pp. 3737–3745, Jul. 2009.
- [65] M. Bern, Y. Cai, and D. Goldberg, "Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry.," *Anal. Chem.*, vol. 79, no. 4, pp. 1393–1400, Feb. 2007.
- [66] M. W. Bern and Y. J. Kil, "Two-dimensional target decoy strategy for shotgun proteomics.," *J. Proteome Res.*, vol. 10, no. 12, pp. 5296–5301, Dec. 2011.
- [67] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.," *Nat. Biotechnol.*, vol. 26, no. 12, pp. 1367–1372, Dec. 2008.
- [68] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: a peptide search engine integrated into the MaxQuant

- environment.," *J. Proteome Res.*, vol. 10, no. 4, pp. 1794–1805, Apr. 2011.
- [69] J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. A. Lajoie, and B. Ma, "PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.," *Mol. Cell Proteomics*, vol. 11, no. 4, p. M111.010587, Apr. 2012.
- [70] S. McIlwain, K. Tamura, A. Kertesz-Farkas, C. E. Grant, B. Diamant, B. Frewen, J. J. Howbert, M. R. Hoopmann, L. Käll, J. K. Eng, M. J. MacCoss, and W. S. Noble, "Crux: rapid open source protein tandem mass spectrometry analysis.," *J. Proteome Res.*, vol. 13, no. 10, pp. 4488–4491, Oct. 2014.
- [71] J. A. Vizcaíno, R. G. Côté, A. Csordas, J. A. Dienes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O'Kelly, A. Schoenegger, D. Ovelheiro, Y. Pérez-Riverol, F. Reisinger, D. Ríos, R. Wang, and H. Hermjakob, "The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D1063–9, Jan. 2013.
- [72] M. Gilar, P. Olivova, A. E. Daly, and J. C. Gebler, "Orthogonality of separation in two-dimensional liquid chromatography.," *Anal. Chem.*, vol. 77, no. 19, pp. 6426–6434, Oct. 2005.
- [73] V. Granholm, W. S. Noble, and L. Käll, "On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics.," *J. Proteome Res.*, vol. 10, no. 5, pp. 2671–2678, May 2011.
- [74] U. Keich and W. S. Noble, "On the importance of well-calibrated scores for identifying shotgun proteomics spectra.," *J. Proteome Res.*, vol. 14, no. 2, pp. 1147–1160, Feb. 2015.
- [75] A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, and T. J. G. Ettema, "Complex archaea that bridge the gap between prokaryotes and eukaryotes.," *Nature*, vol. 521, no. 7551, pp. 173–179, May 2015.
- [76] A. M. Frank, M. M. Savitski, M. L. Nielsen, R. A. Zubarev, and P. A. Pevzner, "De novo peptide sequencing and identification with precision mass spectrometry.," *J. Proteome Res.*, vol. 6, no. 1, pp. 114–123, Jan. 2007.
- [77] X. Wang, S. G. Codreanu, B. Wen, K. Li, M. Chambers, D. C. Liebler, and B. Zhang, "Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity.," *Mol. Cell Proteomics*, p. mcp.RA117.000155, Dec. 2017.
- [78] F. Zhou, Y. Lu, S. B. Ficarro, G. Adelmant, W. Jiang, C. J. Luckey, and J. A. Marto, "Genome-scale proteome quantification by DEEP SEQ mass spectrometry.," *Nat Commun*, vol. 4, p. 2171, 2013.

FIGURE LEGENDS

Figure 1. ProteomeGenerator overview. Transcriptomes and proteomes from the same biologic sample are analyzed in parallel by high-coverage Illumina sequencing and high-resolution, high-accuracy mass spectrometry, respectively. ProteomeGenerator assembles fastq-formatted mRNA sequencing reads into predicted transcripts, identifies reading frames and isoforms, and produces fasta-formatted proteogenomic (PGX) databases containing canonical and non-canonical expressed protein isoforms for subsequent mass spectrometry searches.

Figure 2. Schema for the ProteomeGenerator snakemake workflow. Sequencing reads are aligned using STAR, followed by their de novo and referenced assembly into transcriptomes using StringTie, and processing to identify reading frames and protein isoforms.

Figure 3. Comparison of the canonical and proteogenomic protein databases, displaying (A) number of protein entries, (B) number of amino acid residues, and (C) theoretical tryptic peptides amenable for mass spectrometry analysis specific for either UniProt, PGX, or common to both.

Figure 4. Sensitivity and specificity of mass spectrometry search algorithms. (A, B) Comparison of unique theoretical peptides in the experimental PGX proteome, canonical UniProt, and bacterial proteomes used as negative controls. (C) Sensitivity of tested algorithms, expressed as the number of identified peptides. (D) Specificity of tested algorithms, evaluated from the fraction of peptide-spectrum matches mapped to the negative controls. The PSM fraction mapped to *A. loki* is reported both in absolute

terms (black) and normalized to take into account the relative sizes of the human and archaeobacterial proteomes (grey).

Figure 5. Accurate proteome discovery using statistical target-decoy matching with spectral calibration. (A) Number of peptides identified (FDR<0.01) based on matching spectra from K052 proteome against proteogenomic (PGX, red) and canonical (UniProt, grey) databases. (B) Overlap between the peptides identified in PGX (red) and UniProt (grey) databases. (C) Comparison of PEAKS scores for peptides identified in both PGX and UniProt databases. (D) PEAKS score distribution for peptides identified exclusively in PGX (red) and UniProt (grey) databases. (E) For peptides exclusively identified against the PGX database, PEAKS score distributions for peptides not mapping in UniProt (red) or present in the canonical database (grey). Boxes delimit 25th and 75th percentiles, middle line corresponds to median, whiskers correspond to 5th and 95th percentiles. (F) PEAKS score distributions for peptides identified exclusively in PGX but also mapping in UniProt (grey), or exclusively mapping in the PGX database (red).

Figure 6. Identification of non-canonical protein isoforms using ProteomeGenerator. (A) Genome tracks of non-canonical APEH isoform generation via inclusion of an intronic sequence normally spliced in the canonical APEH isoform. (B) The K052-specific isoform of APEH contains a novel N-terminal sequence, with the splicing junction encompassed by peptide AGPDPGVSPAQVLLSEPEEAAALYR. Residues 35-276 of the protein sequences defined by ProteomeGenerator are identical to residues 4-245 of the canonical UniProt protein sequence. (C) Fragmentation spectrum of the peptide encompassing the novel splice junction, with diagnostic fragment ions and amino acid residues labeled. Italicized ion labels indicate ions with

relative intensity above 25% of the maximum. A star (*) symbol denotes internal ions.
(D) Peptide abundance as based on total fragment ion current for all identified APEH peptides (red: peptides from K052-specific sequence; grey: peptides from canonical sequence).

Supplementary Figure 1. Subsampling analysis of mass spectrometric (A) RNA sequencing (B) dataset. Plateauing of the number of sequences detected indicates reaching of the limit of detection for the specific analytical method used, as observed in transcriptomic (B) but not in proteomic (A) datasets.

Supplementary Figure 2. Raw spectra for peptide AGPDPGVSPAQVLLSEPEEAA-ALYR showing (A) precursor ion (in red: monoisotopic peak) within 5 ppm from expected m/z (846.4343). (B) Original fragmentation spectrum as per Figure 6C, without intensity axis resizing.

TABLES LEGENDS

Supplementary Table 1. Calibration of peptide-spectral matching using negative controls.

Supplementary Table 2. Peptides matched to the PGX database at $FDR < 0.01$, using PEAKS.

Supplementary Table 3. Peptides matched to the UniProt database at FDR<0.01, using PEAKS.

Supplementary Table 4. Identified peptides with PEAKS score higher than 50 and not mapping the reference UniProt database.

This manuscript contains 38 pages, including this one.