Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy

Supporting Information

Anthony J. Cesnik, † Michael R. Shortreed, † Gloria M. Sheynkman, † Brian L. Frey, † and Lloyd M. Smith*,†,‡

†Department of Chemistry, University of Wisconsin-Madison, 1101 University Avenue, Madison, Wisconsin 53706, United States

‡Genome Center of Wisconsin, University of Wisconsin-Madison, 425G Henry Mall, Madison, Wisconsin 53706, United States

TABLE OF CONTENTS

Supporting Tables (Excel spreadsheet)

Datasets, Settings, and Reference Database

- **Table S-1.** Hyperlinked repository information for the RNA-Seq datasets used in this project.
- **Table S-2.** The MS/MS datasets for the 10 human cell lines used in this study have deep and consistent coverage.
- **Table S-3.** Tophat aligner settings.
- **Table S-4.** Accessions of proteins found in UniProt-XML reference proteome, downloaded on January 5, 2015, that were combined with sequence variant databases in each sample-specific databases.

Peptide Identification Summaries

- **Table S-5.** Results for searching sample-specific databases for ten cell lines with selected cutoffs.
- **Table S-6.** Compiled lists of all identified SAV, NSJ, PTM, and unmodified peptides at 1% global FDR across the 10 cell lines.
- **Table S-7.** Detailed information for unique SAV peptides at 1% global FDR across the 10 cell lines.
- **Table S-8.** Detailed information for unique NSJ peptides at 1% global FDR across the 10 cell lines.
- **Table S-9.** Compiled protein and variant peptide identifications at 1% global FDR across the 10 cell lines.
- **Table S-10.** PSM counts for each type of PTM peptide across the 10 cell lines.

Notes

The summaries in Tables S-6, S-7, S-8, S-9, and S-10 were constructed using the results of searching databases constructed for each cell line with the least stringent cutoffs noted in Table S-5.

Supporting Figures

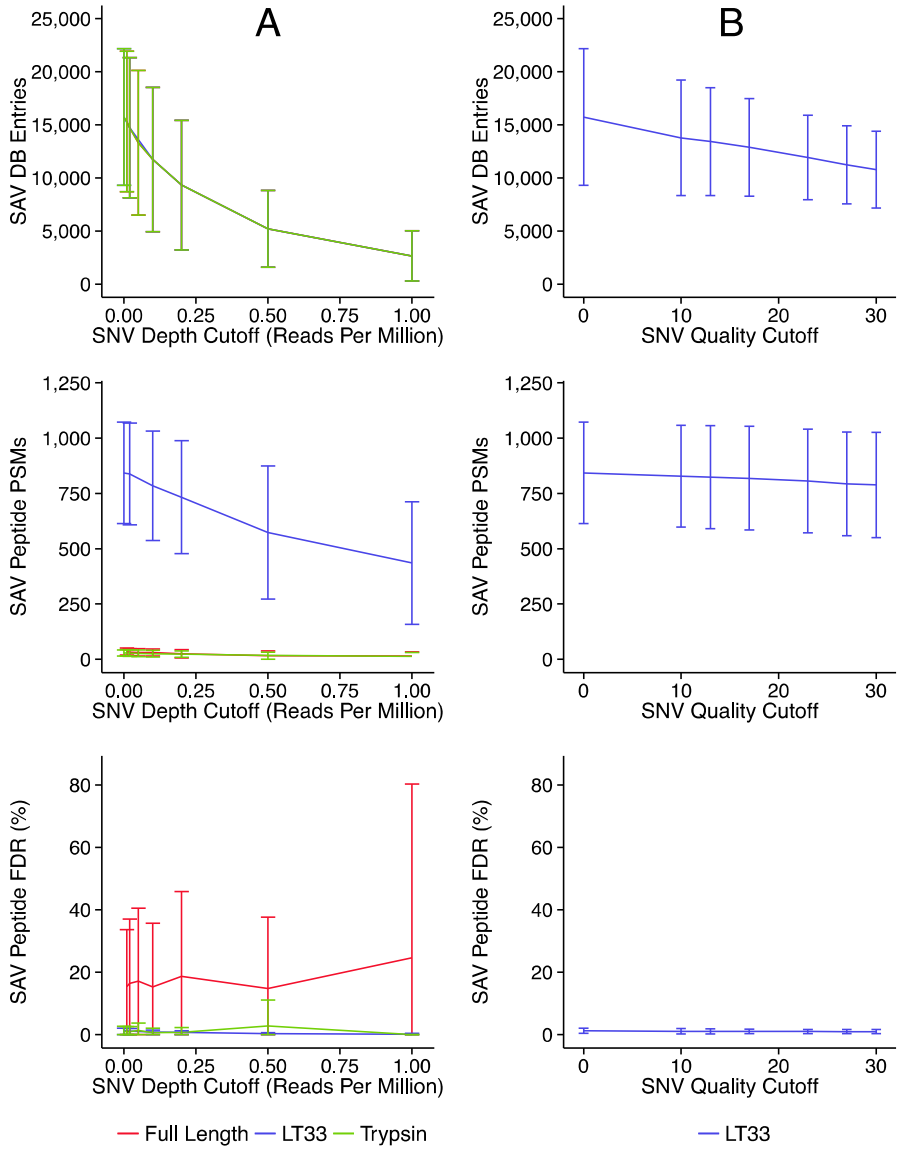

Figure S-1. Comparison of methods for constructing and filtering SAV peptide database entries.

First, two methods for filtering missense SNVs were used prior to constructing SAV databases: the number of reads crossing a variant nucleotide, also noted as the SNV Depth Cutoff (panel A), and the SNV quality score, Q = -10 log P, where Q is the quality score and P is the probability of an incorrect base call (panel B). Second, SAV peptide entries were constructed in three different ways before deciding on keeping 33 amino acids on either side of the variant residue, termed lead-trail 33 or LT33 (blue trace). The other two types of SAV peptide entries were full-length variant protein sequences (red trace) and tryptic peptide sequences containing the variant amino acid (green trace). Panel A contains a comparison of these types of SAV peptide entries, where sample-specific databases for all 10 cell lines were searched against the

corresponding MS/MS data to yield SAV peptide results, displayed as an average value with standard deviations as error bars. These error bars are wide, yet generally consistent, because these searches yielded greatly different numbers of SAV peptides that responded similarly to the filtering methods. Because the trends observed for these missense SNV filtering methods are consistent across the 10 cell lines, these data are useful for comparing the types of SAV entries and filtering methods. First, appending full length variant protein sequences or tryptic SAV peptide sequences led to very few SAV peptide PSMs, and furthermore full length entries led to high SAV peptide FDRs. Therefore, LT33 was the method of choice for appending SAV entries to the sample-specific databases. Second, for LT33 SAV peptide sequences, filtering using a SNV depth cutoff leads to improved SAV peptide FDRs, as described in the main text. Although the quality score is commonly used to filter SNVs, filtering with this score led to no appreciable improvement in the accuracy of peptide identifications.

We note that these searches were performed with a precursor mass tolerance of ± 2.1 Da (monoisotopic) instead of ± 10 ppm. This may have lead to a small number of misidentifications, such as of deamidated PTM peptides having a mass difference of ~1 Da from the unmodified peptide (*i.e.* lacking a SAV, NSJ, or PTM), but we do not expect these few misassignments to change the conclusions from these searches regarding SAV peptide database construction.

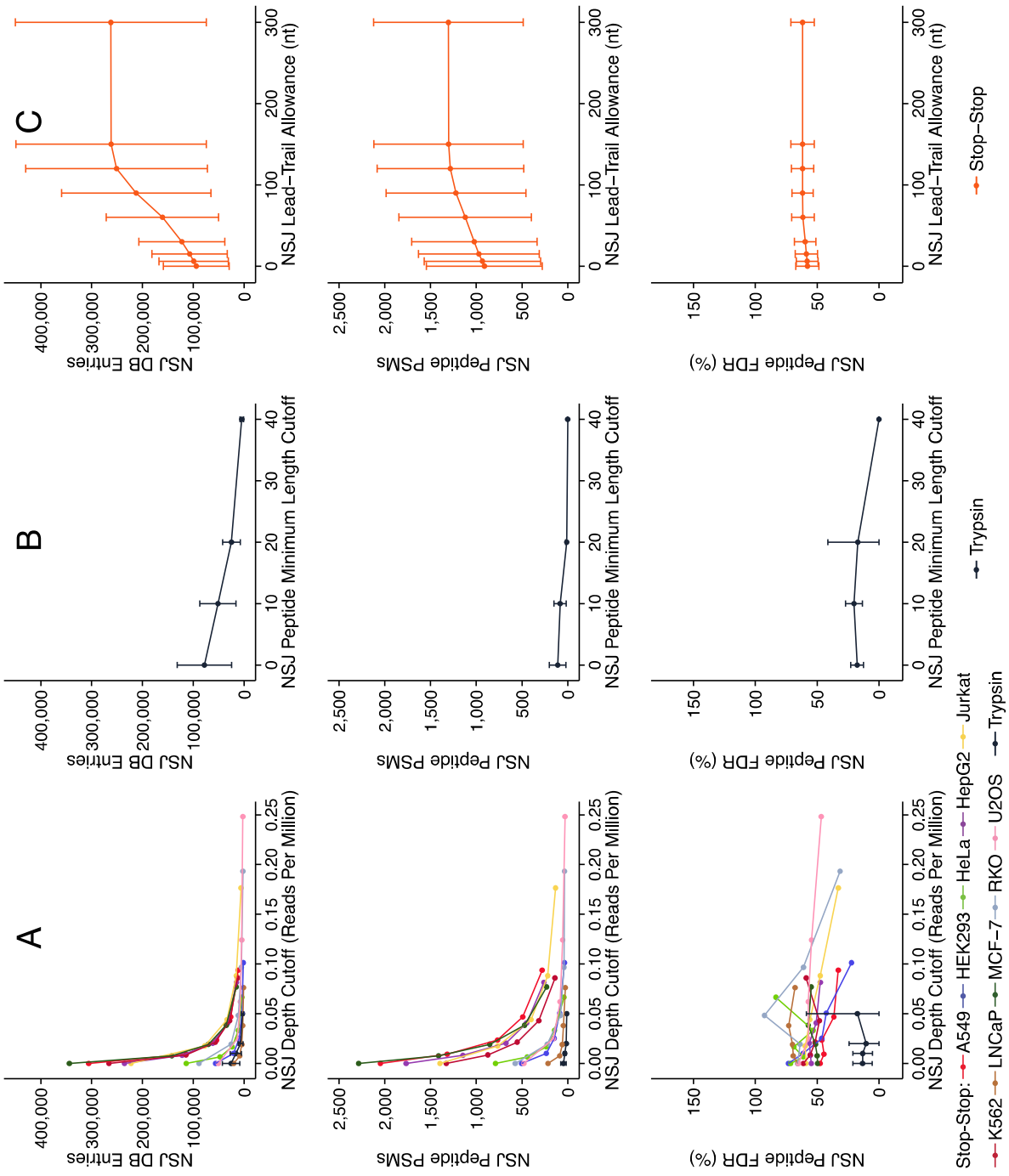

Figure S-2. Comparison of methods for constructing and filtering NSJ peptide database entries.

We evaluated three methods for filtering NSJs prior to constructing the NSJ databases. First, we filtered novel splice junctions with fewer than a certain number of RNA-Seq reads aligned crossing a junction and searched the resulting databases. This is noted as the “NSJ Depth Cutoff,” and the results are shown in panel A. Second, we filtered NSJ peptide sequences shorter than a certain length, noted as the “Minimum Length Cutoff”, and the search results of databases constructed with this filter are shown in panel B. Finally, rejecting all translation frames coding for stop codons within the first exon may exclude spliced nucleotide sequences containing 3’ untranslated regions, and

so we allowed NSJ peptide entries to be constructed from these translation frames. We allowed sequences with stop codons within a specified number of bases from the start of the first exon. We termed this last method the “Lead-Trail Allowance,” and the search results of databases constructed using this method are shown in panel C. The Lead-Trail Allowance was not helpful in creating more accurate databases, and it also led to exceptionally large NSJ peptide databases.

In addition, two methods were used to generate NSJ peptide entries. The most successful one involved keeping only tryptic peptides containing NSJs (“Tryptic” in the legend). The second method, “Stop-Stop,” involves keeping 66 nucleotides on both sides of the splice junction, performing a 3-frame translation to generate 3 amino acid sequences, and then discarding the amino acid sequences before the last stop codon in the first exon and after the first stop codon in the second exon. The resulting amino acid sequence contains the splice junction in all cases, except when the first codon in the second exon is a stop codon; only peptides containing the splice junction were considered NSJ peptides in the analysis. We searched these Stop-Stop databases with a precursor mass tolerance of ± 2.1 Da (monoisotopic) instead of ± 10 ppm. This may have lead to a small number of misidentifications, such as of deamidated PTM peptides having a mass difference of ~1 Da from the unmodified peptide (*i.e.* lacking a SAV, NSJ, or PTM), but we do not expect these small number of misassignments to change the conclusions from these searches regarding NSJ peptide database construction. A comparison of tryptic and Stop-Stop NSJ entries can be found in the panel A of this figure. Across NSJ depth cutoffs, tryptic peptides produced more accurate results, as illustrated by the NSJ peptide FDR falling mostly below 25%, where Stop-Stop entries led to values of above 50% NSJ peptide FDR.

There was a need for filtering with NSJ peptides, as compared to SAV peptides, due to the >50% FDR for these experiments. The NSJ depth cutoff described in the main text of this work and illustrated in panel A was the most successful filtering method, leading to a decrease in the NSJ peptide FDR for 7 of the 10 cell lines. However, it led to a dramatic decrease in the number of peptide identifications, so that tradeoff was considered before choosing final cutoffs (see main text). The other method noted above was not successful for filtering NSJ entries; applying the “minimum length cutoff” filter illustrated in panel B did not change the NSJ peptide FDR before eliminating almost all peptide identifications with a diminishing NSJ database size.

We attempted two other filtering methods that led to no improvement (results not shown). First, we attempted to keep only entries that contained genomic loci annotated as protein coding by comparing the exon loci recorded in the splice junction BED files to gene model (GTF file) annotations. This diminished the database size slightly, but it did not lead to a notable decrease in the high NSJ peptide FDR, indicating it did not improve the accuracy of the NSJ peptide database. Second, we filtered splice junctions based on the number of nucleotides translated before the end of tryptic peptides in the second exon. This did not lead to a decrease in the NSJ peptide FDR, indicating that peptides crossing exons are equally likely to be true positives, independent of the length of translation in the second exon.