Identifying genetically variant peptides in cancer samples by the NIST hybrid spectral library search

Guanghui Wang, Meghan C. Burke, Sergey L. Sheetlin, Yuri A. Mirokhin, Dmitrii V. Tchekhovskoi, Stephen E. Stein
Mass Spectrometry Data Center, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899 USA

ASMS 2020

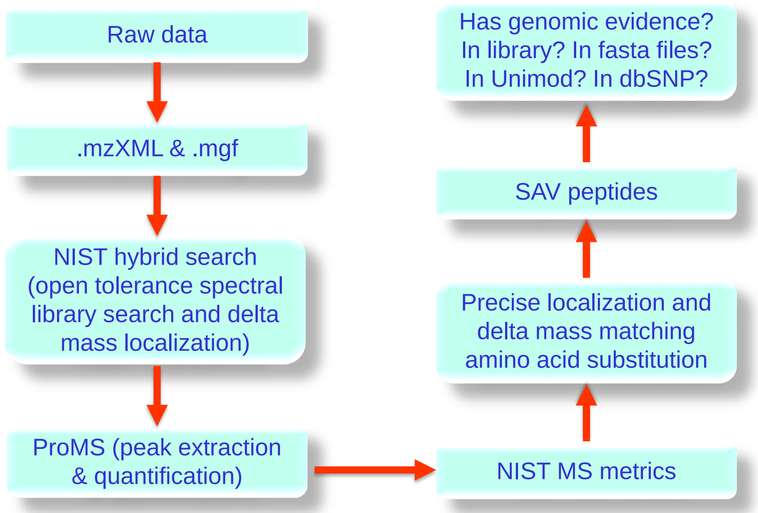
peptide.nist.gov

Introduction

Genetically variant peptides (GVPs) arising from nucleotide polymorphism play important roles in health and disease. Identifying them, however, has been a challenging task in proteomics, as conventional approaches typically employ a reference protein sequence database or a library of reference spectra where GVPs are unlikely to be found. The development of open mass tolerance search and localization tools such as the NIST hybrid spectral library search allows digging deeper into the mass spectrometry dark matter (i.e., uninterpreted spectra) and uncovering unexpected modifications and amino acid substitutions. Here we report a workflow utilizing the hybrid search algorithm to identify single amino acid variant (SAV) peptides and potential rare GVPs from LCMS data of patient samples.

Methods

LCMS data for iTRAQ 4plex-labeled breast cancer samples were downloaded from the CPTAC data portal and analyzed by the NIST mass spectrometry data processing pipeline. The pipeline consists of raw data conversion, hybrid spectral library search, peak extraction and quantification, and QC metrics generation. Custom scripts were developed in R to read in the pipeline output and identify SAV peptides based on a precise localization of modification and a delta mass corresponding to the modification site mutating to a different amino acid. These SAV peptides were filtered to exclude likely experimental artifacts and retain potential GVPs. Spectral libraries and the hybrid search program employed in this study are freely available at peptide.nist.gov.



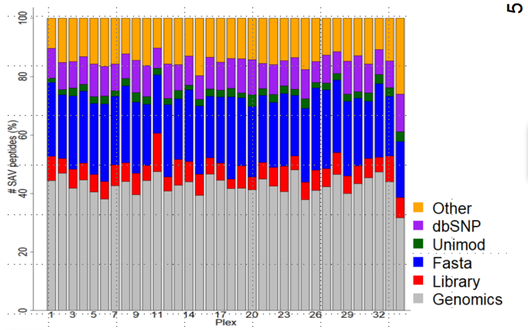
Acknowledgment: This work was supported by the Interagency Agreement NIST-NCI IAA ACO15005-0005/6 between the NIST MS Data Center and the NIH CPTAC Program.

Disclaimer: Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure identified adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

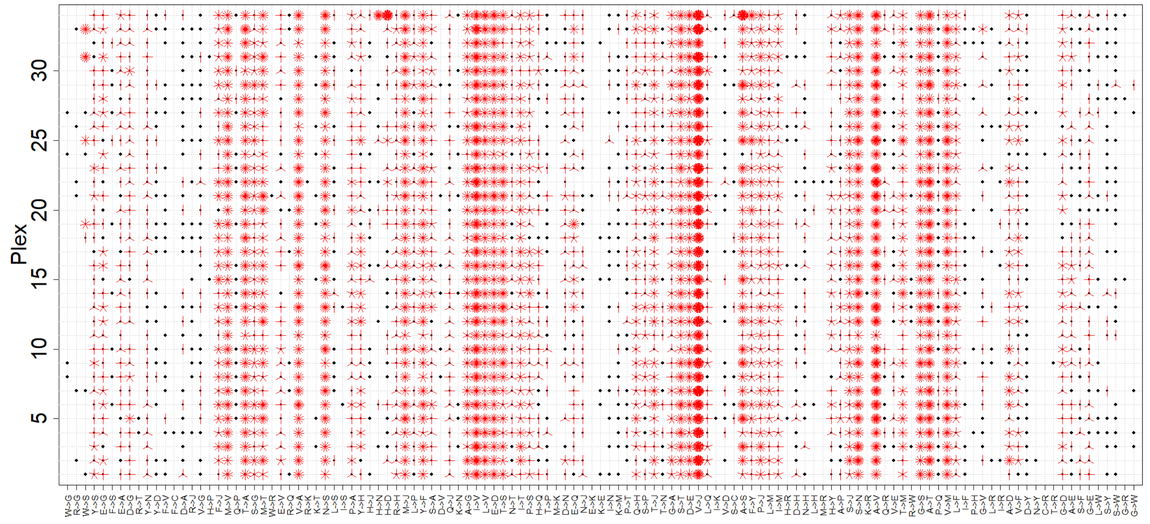
Results

Source	Count	Fraction	Count (cumulative)	Fraction (cumulative)
Genomic	129	40.44%	129	40.44%
Library	27	8.46%	156	48.90%
RefSeq	13	4.08%	169	52.98%
UniProt	64	20.06%	233	73.04%
Unimod	3	0.94%	236	73.98%
dbSNP	39	12.23%	275	86.21%
Other	44	13.79%	319	100.00%

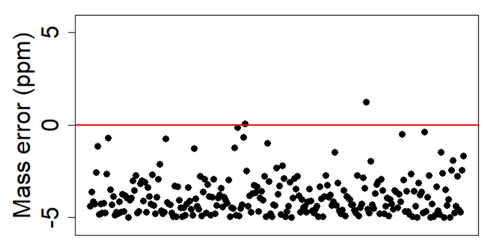
1 SAV peptides identified from a 24-fraction iTRAQ data set and their sources. More than 40% were confirmed as GVPs by genomic sequencing data. About 33% were present in the spectral library and fasta database (RefSeq and UniProt), representing common peptide variants. Another 12% were in dbSNP and were likely GVPs.

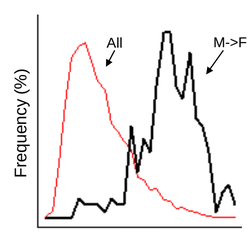


2 Similar distributions of identified SAV peptides in various sources across 34 data sets (or plexes; each plex consisted of 24 fractions).

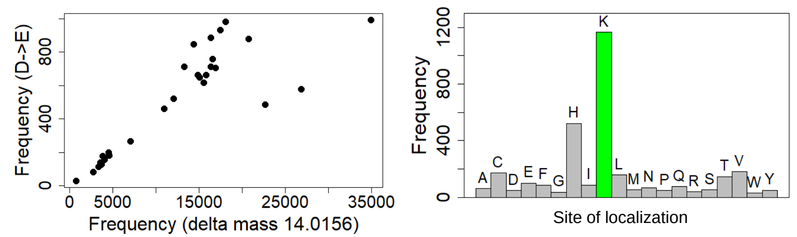


3 Sunflower plots showing the distribution of various SAV identifications among 34 data sets (plexes). Some SAVs (e.g., V->J [V->I/L]) were much more frequently observed than others (e.g., W->G). Some SAVs (e.g., V->J) had similar occurrences across data sets while others displayed more variability (e.g., H->D). The degree of variation probably reflects the nature of genetically variant peptides (rarer ones are less frequently seen and are more variable). A dramatic increase of certain SAVs in a data set could be caused by experimental artefacts. For example, the H->D in plex 34 seemed to be related to a high level of oxidation observed for this plex.





4 False-positive SAVs could result from misassigning experimental artefacts as amino acid substitutions. We propose three rules to distinguish them. (A) Asymmetrical mass error distribution. Mass measurement errors should center around zero. Shown is the lopsided error distribution for M->F SAV peptides (left panel; p<3e-71). These peptides also had significantly larger M.W. (right panel; mostly >3000 Da). In addition, samples with increased M->F had higher level of oxidation (+15.9949). All of these suggest the misassignment of M oxidation as M->F (+16.02792).



5 (B) High correlation with experimental artefacts. Left panel is an example of D->E SAVs highly correlated with elevated levels of methylation (+14.0156) at r=0.84 (p<2e-8), suggesting methylated D was mistaken as D->E in these samples. (C) Nonspecific localization. The right panel gives an example showing the delta mass for K->R was also localized to other sites that don't correspond to amino acid substitutions. Examination of K->R SAV peptides also indicates an asymmetrical mass error distribution, suggesting at least some K->R SAVs likely reflected artefacts.

Summary & Conclusions

- We present a workflow for the identification of a diverse array of SAV peptides (potential GVPs) across large-scale data sets of cancer samples.
- Variations in SAVs across samples very likely reflect the nature of GVPs, but a large increase in specific samples are often caused by experimental artefacts.
- We also propose rules for identifying false-positive SAVs resulting from artefacts.