Published in final edited form as:

Cancer Res. 2017 November 01; 77(21): e47–e50. doi:10.1158/0008-5472.CAN-17-0335.

P-MartCancer-Interactive Online Software to Enable Analysis of Shotgun Cancer Proteomic Datasets

Bobbie-Jo M. Webb-Robertson*, Lisa M. Bramer, Jeffrey L. Jensen, Markus A. Kobold, Kelly G. Stratton, Amanda M. White, and Karin D. Rodland

Pacific Northwest National Laboratory, Richland WA, 99352, USA

Abstract

P-MartCancer is an interactive web-based software environment that enables statistical analyses of peptide or protein data, quantitated from mass spectrometry-based global proteomics experiments, without requiring in-depth knowledge of statistical programming. P-MartCancer offers a series of statistical modules associated with quality assessment, peptide and protein statistics, protein quantification and exploratory data analyses driven by the user via customized workflows and interactive visualization. Currently, P-MartCancer offers access and the capability to analyze multiple cancer proteomic datasets generated through the Clinical Proteomics Tumor Analysis Consortium at the peptide, gene and protein levels. P-MartCancer is deployed as a web-service (<https://pmart.labworks.org/cptac.html>), alternatively available via Docker Hub (<https://hub.docker.com/r/pnnl/pmart-web/>).

Keywords

Proteomics; Statistical workflows; software

The use of mass spectrometry (MS)-based technologies for global protein profiling of cancer-related tissues and bodily fluids has become a major focus in research centered on biomarker discovery to better detect and treat cancer. Global proteomic technologies are of interest in the field of cancer research because they provide a vital source of information regarding biological functions at the protein level(1–4). Global MS-based proteomics often allow hundreds of thousands of peptides mapping to tens of thousands of proteins to be measured, offering scientists an unprecedented view into processes involved in cancer development and progression. However, as with many newer high-throughput molecular technologies, the data are complex with multiple sources of variability (1), consequently, the post peptide identification and quantification (i.e., downstream) analyses of these datasets require significant specialized expertise due to inherent data challenges, such as missing values and isoform quantification(5–7). Data processing is generally performed via custom and unpublished scripts, leading to continued issues in reproducibility of results (8). Thus, challenges in both access to the data and statistical methods in an easy to understand format for biomedical researchers has led to underutilization of publicly available databases, such as

*Bobbie-Jo Webb-Robertson, 902 Battelle Blvd, K7-20, Richland WA, 99352, bj@pnnl.gov.

The authors declare no potential conflicts of interest

those generated through the Clinical Proteomics Tumor Analysis Consortium (CPTAC) funded through the National Cancer Institute. The development of software that enables continued exploration and evaluation of existing data could increase the potential for new discovery from these comprehensive public datasets.

Existing software capabilities associated with CPTAC can broadly be placed in four categories for global MS-based data (<https://proteomics.cancer.gov/resources/softwaretools>); 1) spectral pre-processing, 2) peptide and protein identification, 3) specific methods for qualitative and quantitative differential statistics, and 4) network analyses. P-MartCancer broadly fits in the third category, however unlike these targeted analyses on data that has already been processed into a particular form, P-MartCancer offers a holistic approach to data analysis allowing all steps of analysis from quality control through pattern discovery to be performed in a workflow-based manner. In addition, the manner that P-Mart offers direct access to the data for statistical analysis at multiple levels (peptide, gene, and protein), with the clinical information aligned, is unique within the set of tools supporting CPTAC. P-MartCancer provides an open, web-based interactive platform for performing quality control processing, statistics, protein quantification and exploratory data analysis tasks in a manner that is reproducible (see Video 1).

Material and Methods

Software Development

P-MartCancer functions are developed in R or Rccp, Rserve (<https://www.rforge.net/Rserve/>) is used to communicate between R and the web-service, and the interface is developed in Java. A subset of the R functions are currently available via GitHub (<https://github.com/pmartR/>) and the web-service can also be installed via Docker Hub (<https://hub.docker.com/r/pnnl/pmart-web/>). Developed in this manner, adding functionality to P-MartCancer is straight-forward through a standardized pipeline.

Cancer Proteomics Data

P-MartCancer currently accesses multiple proteomic datasets generated through the CPTAC available on the Data Portal (<https://cptac-data-portal.georgetown.edu/cptacPublic/>). Data is available at the peptide, protein and/or gene levels where the protein and gene data is based on a defined Common Data Analysis Pipeline (CDAP)(9). P-MartCancer offers flexibility to the user to either perform statistical processing on the peptide data (before or after parsimony) with P-MartCancer functions for gene or protein quantitation, or to use the data as supplied by CPTAC via the CDAP. Currently various numbers of datasets are available for ovarian cancer and breast cancer(2), however; new datasets are being added as they become available. Each of these datasets contains meta-data about the experiment and the user selects the clinical variable of interest (e.g., vital status, tumor stage) for data processing allowing various hypotheses to be explored.

Results - P-MartCancer Analysis Modules

P-MartCancer is a modular workflow tool (Figure 1A) with four primary capabilities; 1) quality control processing, 2) gene or protein quantification, 3) statistics, and 4) exploratory

data analysis. These are further divided into six key modules under which new functionality can be easily added through the R/Rserve to Java framework described previously. The modules and functions available depend upon the type of data being evaluated; peptide, protein or gene. Each function called is displayed visually either via tables or figures in sequential order highlighted by the green text on top of the P-MartCancer screen where the current module is blue (top Figure 1B). Finally, the entire process is documented for the user at the end of the workflow, Figure 1C, and all datasets and statistical results are available for download as .CSV files.

Quality Control Processing

A challenge with proteomics data is pre-processing in a manner that does not ignore the different sources of variability that contribute to the complexity of these datasets. For example, peptides are not uniformly identified for all samples and thus large quantities of missing values are common, due to both random and non-random mechanisms. P-MartCancer offers a suite of pre-processing capabilities that handle issues of peptide and protein coverage, as well as the identification of potential outlier samples(7,10,11). At the peptide level researchers can choose to 1) remove proteins with inadequate coverage (Peptide Filter); 2) remove samples with outlier behavior (Sample Outlier Filter)(10); 3) remove peptides too sparse for statistical analysis (Peptide Coverage Filter); 4) remove peptides with extremely high variability in the context of a Coefficient of Variation (CV Filter) and 5) perform normalization. The quality control processing generates a high-quality dataset for continued statistical analysis. The gene- and protein-level datasets will allow functions 2–4 above, evaluating outlier behavior and removing genes or proteins that are too sparse or variable to add value to downstream statistics.

Differential Statistics

Statistical analysis of peptide, gene or protein-level data is currently focused on quantitative analysis of variance (ANOVA)-based methods and qualitative G-test methods(11). The ANOVA method allows the comparison of any number of groups, performing a multiple test correction when more than two groups are compared (e.g., clinical variable ‘Tumor Residual Disease’ is separated into four categories by size). A Tukey adjustment is performed when the user compares all groups to one another and a Dunnett adjustment is performed when the user compares back to a single control group. Data is not imputed, statistical results are generated based only on the observed data to assure that accurate estimates of variance are being utilized for these tests. To identify qualitative changes, a G-test is also performed for each biomolecule to evaluate whether the number of non-missing observations in one group is more than expected by chance. Multiple test adjustments for the G-test are performed using a Holm-Bonferroni correction. The total significance is given in the context of a bar graph, and to facilitate exploratory data analysis capabilities a p-value threshold (default of 0.05) is used to move forward only a subset of the peptides, genes, or proteins for further evaluation.

Protein Quantification

There are numerous approaches to quantify proteins from the measured peptide-level data(12). P-MartCancer currently offers a standard reference-based approach that scales all

peptides to the most abundant, or most reproducible, peptide and gives the median signal(13).

Exploratory Data Analysis

P-MartCancer offers two exploratory data analysis capabilities. The first is Probabilistic Principal Component Analysis (PPCA), which allows P-MartCancer to perform PCA without imputing missing values, demonstrated to be valuable in proteomics(7). The resulting scores are plotted using a standard scatter plot of the scores from the first two principal components that most cancer and biomedical researchers are accustomed to, allowing visual exploration of clustering across samples.

P-MartCancer also offers an interactive and customizable plotting capability called Trelliscope that allows sorting and querying across the peptides, genes or proteins, Figure 1B. Trelliscope uses the statistical results to plot each peptide, gene or protein via either a boxplot of differential abundance or a bar graph of the number of observations, to view quantitative and qualitative changes, respectively. The entire space of the biomolecules being explored can be reduced by selecting various thresholds, such as p-value or fold-change, or the user can search for specific genes or proteins of interest. For each plot, the gene and protein information can be selected and the associated information can be viewed in webpages, <http://www.genecards.org> and <http://www.uniprot.org>, respectively. In the example in Figure 1B the user has searched for BRAF and rapidly views the associated protein-level quantified information based on the clinical variable selected, 'Macroscopic Disease', and other information, such as that BRAF has a p-value of ~0.04 for the specific comparison selected.

Discussion

P-MartCancer offers a new online platform to access CPTAC datasets to enable new analyses. There is a wealth of capabilities that could be extremely useful to the proteomics community, many of which are under active development. For example, proteoform discovery, that is the identification of proteins with multiple forms, is also an important component of protein quantification (6). Additional future work is focused on adding new capabilities in statistical testing, machine learning and gene set enrichment analysis, as well as the development of a user-upload capability to enable all researchers with MS-based peak-intensity data to create reproducible statistical downstream processing pipelines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

P-MartCancer was developed at Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under contract DE-AC06-76RL01830.

Financial Support: B.J.M. Webb-Robertson, National Cancer Institute, U01-1CA184783

References

1. Gajadhar AS, Johnson H, Slebos RJ, Shaddox K, Wiles K, Washington MK, et al. Phosphotyrosine signaling analysis in human tumors is confounded by systemic ischemia-driven artifacts and intra-specimen heterogeneity. *Cancer Res.* 2015; 75:1495–503. [PubMed: 25670172]
2. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature.* 2016; 534:55–62. [PubMed: 27251275]
3. Slebos RJ, Wang X, Wang X, Zhang B, Tabb DL, Liebler DC. Proteomic analysis of colon and rectal carcinoma using standard and customized databases. *Sci Data.* 2015; 2:150022. [PubMed: 26110064]
4. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell.* 2016; 166:755–65. [PubMed: 27372738]
5. Choi M, Eren-Dogu ZF, Colangelo C, Cottrell J, Hoopmann MR, Kapp EA, et al. ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. *J Proteome Res.* 2017
6. Webb-Robertson BJ, Matzke MM, Datta S, Payne SH, Kang J, Bramer LM, et al. Bayesian proteoform modeling improves protein quantification of global proteomic measurements. *Mol Cell Proteomics.* 2014; 13:3639–46. [PubMed: 25433089]
7. Webb-Robertson BJ, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res.* 2015; 14:1993–2001. [PubMed: 25855118]
8. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010; 11:R86. [PubMed: 20738864]
9. Markey, SP., Rudnick, PA., Mirokhin, YI., Roth, J., Stein, SE. Common Data Analysis Pipeline (CDAP). 2014. <<https://cptac-data-portal.georgetown.edu/cptac/aboutData/show?scope=dataLevels>>.
10. Matzke MM, Waters KM, Metz TO, Jacobs JM, Sims AC, Baric RS, et al. Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics.* 2011; 27:2866–72. [PubMed: 21852304]
11. Webb-Robertson BJ, McCue LA, Waters KM, Matzke MM, Jacobs JM, Metz TO, et al. Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data. *J Proteome Res.* 2010; 9:5748–56. [PubMed: 20831241]
12. Matzke MM, Brown JN, Gritsenko MA, Metz TO, Pounds JG, Rodland KD, et al. A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics.* 2013; 13:493–503. [PubMed: 23019139]
13. Polpitiya AD, Qian WJ, Jaitly N, Petyuk VA, Adkins JN, Camp DG 2nd, et al. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics.* 2008; 24:1556–8. [PubMed: 18453552]

Figure 1. Screenshots from P-MartCancer. (A) The user selects the workflow to be implemented based on the data type selected. (B) The exploratory evaluation capability allows users to find proteins or genes of interest and evaluate the associated data visually. (C) The log of all of the steps in the analysis performed by the user for use in publication or to facilitate reproducible analyses.