Diagnostic Ion Data Analysis Reduction (DIDAR) allows rapid quality control analysis and filtering of multiplexed single cell proteomics data

Conor Jenkins¹ and Benjamin C. Orsburn^{2*}

¹The University of Maryland, College Park, MD, USA

²The Johns Hopkins University Medical School, Baltimore, MD, USA

*Corresponding; borsbur1@jhmi.edu

Abstract Graphic

Abstract

Recent advances in the sensitivity and speed of mass spectrometers utilized for proteomics and metabolomics workflows has led to a dramatic increase in data file size and density. For a field already challenged by data complexity due to a dependence on desktop PC architecture and the Windows operating systems, further compromises appear inevitable as data density scales. As one method to reduce data complexity, we present herein a light-weight python script that can rapidly filter and provide analysis metrics from tandem mass spectra based on the presence and number of diagnostic fragment ions determined by the end user. Diagnostic Ion Data Analysis Reduction (DIDAR) can be applied to any mass spectrometry dataset to create smaller output files containing only spectra likely to contain post-translational modifications or chemical labels of interest. In this study we describe the application DIDAR within the context of multiplexed single cell proteomics workflows. When applied in this manner using reporter fragment ions as diagnostic signatures, DIDAR can provide quality control metrics based on the presence of reporter ions derived from single human cells and simplified output files for search engine analysis. The simple output metric text files can be used to rapidly flag entire LCMS runs with technical issues and remove them from downstream analysis based on end user minimum requirements. Acquisition files that pass these criteria are further improved through the automatic removal of spectra

where insufficient signal from single cells is observed. We describe the application of DIDAR to two recently described multiplexed single cell proteomics datasets.

Introduction

Liquid chromatography coupled to high resolution tandem mass spectrometry (LCMS) continues to remain the dominant technology in proteomics², metabolomics³⁻⁵ and lipidomics⁶ research. Nearly all proteomics and metabolomics data processing today is performed using desktop personal computer (PC) architecture.¹ While many open software packages are amenable to installation on high performance cluster environments (HPCs)⁷⁻⁹ these have yet to be adopted by the LCMS community. Likewise, software designed for use on the Cloud have been demonstrated but have been slow to be adopted despite the promise of improved speed and increased biological relevance processing on the Cloud can allow.^{10,11} One method to circumvent the size of the generated LCMS files is to apply pre-filters to remove tandem mass spectra (MS2) that are of interest. We have recently described Reporter Ion Data Analysis Reduction (RIDAR) which allows us to set minimum quality filters and relative quantification values in isobarically tagged proteomic datasets. Spectra lacking reporter ions of sufficient signal to noise ratio across a user defined minimum number of reporters can be discarded. In addition, spectra that are not quantitatively interesting to the end user, such as peptides that display no change in relative quantities can likewise be filtered prior to data analysis.¹²

Diagnostic oxonium ions produced during the fragmentation of glycopeptides are commonly used in both real time data acquisition and in data processing. Orbitrap systems can identify ions producing oxonium ion fragments for further analysis, with the most impressive results obtained by further analysis with electron transfer dissociation techniques.^{13,14} Software such as SugarQB begin by separating out MS2 spectra with apparent oxonium ions for specialized searches considering the glycan moieties that could lead to the production of those ions.¹⁵ Although not as frequently used, many other commonly studied post-translational modification also produce unique diagnostic fragment ions that could be utilized in a similar manner. A systematic analysis of modified peptides synthesized for the ProteomeTools project has identified fragment characteristics and diagnostic ions for many PTMs.¹⁶ The utilization of diagnostic ions can help prevent the misidentification of modified peptides such as arginine citrullination, which has been the topic of multiple controversial studies and recent retractions.¹⁷⁻¹⁹

Single cell proteomics by mass spectrometry is a new field of research largely made possible innovative applications of existing technology. The revolutionary work of Budnik *et al.*, first described the SCoPE-MS workflow whereby the additive signal of multiplexed peptide samples could be leveraged to effectively amplify peptide signal.²⁰ To achieve these effects SCoPE-MS employs a carrier channel of isobarically labeled peptides of higher relative concentration to resuspend labeled peptides from single cells. Data dependent LCMS analysis of these multiplexed samples leads to the fragmentation of and identification of peptides from the carrier channel. The isobaric reporter ion fragments corresponding to labels applied to single cells, if within the intrascan linear dynamic range of the instrument, can be used for both peptide identification and quantification.²¹⁻²⁴ Although other single cell technologies such as single cell RNAseq have been sorting and analyzing single cells for over a decade with increasingly sophisticated technology, relatively little of this can be applied to proteomics approaches.^{25,26} While new technologies such as nanoPOTS²⁷, nPOP²³ and ProteoChip²⁴ are improving these aspects, technical challenges in sample preparation may be the biggest limiter to entry into single cell proteomics. Single human cells and the

volumes of reagents used in their preparation and analysis can be affected by factors as ubiquitous as relative humidity and static electricity. As a relatively large number of single cells must be analyzed for sufficient statistics to capture both technical and biological variability, failed runs may not be detected until data analysis is completed which may take days or weeks on desktop PC architecture. Quality control tools that could be leveraged prior to complete data analysis could simplify and improve nearly all downstream processes.

Methods and Materials

Script development

DIDAR was written in Python 3.8 using the Spyder 4.1.4. IDE.²⁸ DIDAR is contained in less than 100 total lines of code inclusive of comments and line spacing. In addition, utilizing the built-in os python package as the only dependency adds to the lightweight nature of the program.

Single cell proteomics by mass spectrometry (SCoPE-MS) data quality filtering

For the development of DIDAR as a quality control analysis pipeline for SCoPE-MS we used publicly available files from two recent SCoPE-MS studies. The first utilized the SCoPE2 methodology using both TMT11-plex and TMTPro16-plex tags to study macrophage heterogeneity using a quadrupole Orbitrap MS2 based approach and 95 minute acquisition times. The second used the same sample preparation method but utilized the 9 unit resolution mass tags from the TMTPro16 reagents to study single human KRASG12C model cells (NCI-H-358) using a third-generation TIMSTOF mass analyzer and 30 minute acquisition methods. The vendor format instrument files from each study were converted to MGF using MSConvert Version: 3.0.20310-0d96039e2 (developer build) using the default conversions for each instrument. For applying DIDAR to these SCoPE-MS files, the DIDAR.py and diagnostic ion text was copied into the folder containing the MGF files from each respective study. The diagnostic ion text file referenced by DIDAR only contained the exact masses of the reporter ions as reported by the manufacturer. The DIDAR script was executed identically for each file, with the exception of the ion mass tolerance which was set at 0.002 Da for the Orbitrap files and 0.005 for the TIMSTOF files to compensate for the relative differences in mass accuracy between these analyzers.

Data Availability Statement

All original files used in this study are publicly available on the MASSIVE repository (massive.ucsd.edu). The SCoPE2 quadrupole Orbitrap files are available as accessions MSV000083945 and MSV000084660. The TIMSTOF single cell files were accessed as accessions MSV000088144 using the reviewer password provided in the study preprint (pasefSCOPE700N). All DIDAR filtered MGFs and output metrics for this study have been deposited at: <ftp://massive.ucsd.edu/MSV000088887/>.

Code availability statement

The DIDAR script and diagnostic ion text files for TMT reporter ions and video tutorials are available at: <https://github.com/orsburn/DIDARSCPQC>.

Results and Discussions

Reanalysis of two SCoPE-MS studies

Each file processed through DIDAR is assigned two output files. The first is a new filtered MGF containing only spectra meeting the end user's criteria. For example, by changing the minimum number of reporter ions to 5, we will create an output file that only contains spectra where at least 5 of the TMT reporter ions are observed. In this study we left the default number of ions at 1 to capture the widest variability in overall metrics. The second output file is a short diagnostic text file that provides a simple count of the number of spectra where each diagnostic ion is counted. **Figure 1C** is a visualization from the 25 pasefRIQ files analyzed. As shown, both files 775 and 782 have a number of 128n reporter ions that appear significantly below the average of the other files analyzed within this batch. In addition, we can observe a general increase in the number of spectra with reporter ions at later points in this batch. General trends of increasing or decreasing numbers of reporter ion signal may indicate systematic errors in sample handling or other hardware. When evaluating the original diagnostic counts output data (**Supplemental Data**) we see that in run 782 the failure was comprehensive for the entire experiment as no cells returned more than 1,500 spectra with observed reporter ion signal. In contrast, run 775 appears to have a near average number of reporter ions in all measured channels except 128n, suggesting a technical issue occurred in the handling or processing of that single cell. DIDAR output can also be combined to evaluate the reproducibility of data within and between batches. As shown **Figure 1D**, these data can be easily examined for simple analyses of each individual cell and file across batches of hundreds of individual cells.

Conclusions

We have described the application of a rapid lightweight Python tool for the quality assessment and filtering of single cell proteomics data. When applied in this manner, DIDAR can flag both entire LCMS runs that experienced technical issues for further analysis or for removal from analysis. In addition, DIDAR can generate new input files that only contain data meeting user required minimum cutoffs.

References

- (1) Prakash, A.; Ahmad, S.; Majumder, S.; Jenkins, C.; Orsburn, B. Bolt: A New Age Peptide Search Engine for Comprehensive MS/MS Sequencing Through Vast Protein Databases in Minutes. *J. Am. Soc. Mass Spectrom.* **2019**. <https://doi.org/10.1007/s13361-019-02306-3>.
- (2) Yates, J. R.; Ruse, C. I.; Nakorchevsky, A. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annual Review of Biomedical Engineering*. 2009. <https://doi.org/10.1146/annurev-bioeng-061008-124934>.
- (3) Martin, J. C.; Maillot, M.; Mazerolles, G.; Verdu, A.; Lyan, B.; Migné, C.; Defoort, C.; Canlet, C.; Junot, C.; Guillou, C.; Manach, C.; Jacob, D.; Bouveresse, D. J. R.; Paris, E.; Pujos-Guillot, E.; Jourdan, F.; Giacomoni, F.; Courant, F.; Favé, G.; Le Gall, G.; Chassaigne, H.; Tabet, J. C.; Martin, J. F.; Antignac, J. P.; Shintu, L.; Defernez, M.; Philo, M.; Alexandre-Gouaubau, M. C.; Amiot-Carlin, M. J.; Bossis, M.; Triba, M. N.; Stojilkovic, N.; Banzet, N.; Molinié, R.; Bott, R.; Goulitquer, S.; Caldarelli, S.; Rutledge, D. N. Can We Trust Untargeted Metabolomics? Results of the Metabo-Ring Initiative, a Large-Scale, Multi-Instrument Inter-Laboratory Study. *Metabolomics* **2015**. <https://doi.org/10.1007/s11306-014-0740-0>.
- (4) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O. Identification of Small Molecules Using Accurate Mass MS/MS Search. *Mass Spectrometry Reviews*. 2018. <https://doi.org/10.1002/mas.21535>.
- (5) Ribbenstedt, A.; Ziarrusta, H.; Benskin, J. P. Development, Characterization and Comparisons of

- Targeted and Non-Targeted Metabolomics Methods. *PLoS One* **2018**.
<https://doi.org/10.1371/journal.pone.0207082>.
- (6) Bowden, J. A.; Ulmer, C. Z.; Jones, C. M.; Koelmel, J. P.; Yost, R. A. NIST Lipidomics Workflow Questionnaire: An Assessment of Community-Wide Methodologies and Perspectives. *Metabolomics* **2018**. <https://doi.org/10.1007/s11306-018-1340-1>.
 - (7) Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; Brauer, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; Malmström, L. An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. *J. Proteome Res.* **2013**. <https://doi.org/10.1021/pr300992u>.
 - (8) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods* **2017**. <https://doi.org/10.1038/nmeth.4256>.
 - (9) Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **2018**. <https://doi.org/10.1021/acs.jproteome.8b00175>.
 - (10) Prakash, A.; Taylor, L.; Varkey, M.; Hoxie, N.; Mohammed, Y.; Goo, Y. A.; Peterman, S.; Moghekar, A.; Yuan, Y.; Glaros, T.; Steele, J. R.; Faridi, P.; Parihari, S.; Srivastava, S.; Otto, J. J.; Nyalwidhe, J. O.; Semmes, O. J.; Moran, M. F.; Madugundu, A.; Mun, D. G.; Pandey, A.; Mahoney, K. E.; Shabanowitz, J.; Saxena, S.; Orsburn, B. C. Reinspection of a Clinical Proteomics Tumor Analysis Consortium (CPTAC) Dataset with Cloud Computing Reveals Abundant Post-Translational Modifications and Protein Sequence Variants. *Cancers* . 2021.
<https://doi.org/10.3390/cancers13205034>.
 - (11) Prakash, A.; Mahoney, K. E.; Orsburn, B. C. Cloud Computing Based Immunopeptidomics Utilizing Community Curated Variant Libraries Simplifies and Improves Neo-Antigen Discovery in Metastatic Melanoma. *Cancers* . 2021. <https://doi.org/10.3390/cancers13153754>.
 - (12) Jenkins, C.; Norris, A.; O'Neill, M.; Das, S.; Andresson, T.; Orsburn, B. Reporter Ion Data Analysis Reduction (R.I.D.A.R) for Isobaric Proteomics Quantification Studies. *bioRxiv* **2018**.
<https://doi.org/10.1101/437210>.
 - (13) Singh, C.; Zampronio, C. G.; Creese, A. J.; Cooper, H. J. Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of N-Linked Glycoproteins. *J. Proteome Res.* **2012**. <https://doi.org/10.1021/pr300257c>.
 - (14) Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J. R.; Pevzner, P. A. The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search. *Mol. Cell. Proteomics* **2010**.
<https://doi.org/10.1074/mcp.M110.003731>.
 - (15) Stadlmann, J.; Hoi, D. M.; Taubenschmid, J.; Mechtler, K.; Penninger, J. M. Analysis of PNGase F-Resistant N-Glycopeptides Using SugarQb for Proteome Discoverer 2.1 Reveals Cryptic Substrate Specificities. *Proteomics* **2018**. <https://doi.org/10.1002/pmic.201700436>.
 - (16) Zolg, D. P.; Wilhelm, M.; Schmidt, T.; Médard, G.; Zerweck, J.; Knaute, T.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. ProteomeTools: Systematic Characterization of 21 Post-Translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides. *Mol. Cell. Proteomics* **2018**.
<https://doi.org/10.1074/mcp.tir118.000783>.

- (17) Lee, C. Y.; Wang, D.; Wilhelm, M.; Zolg, D. P.; Schmidt, T.; Schnatbaum, K.; Reimer, U.; Pontén, F.; Uhlén, M.; Hahne, H.; Kuster, B. Mining the Human Tissue Proteome for Protein Citrullination. *Mol. Cell. Proteomics* **2018**. <https://doi.org/10.1074/mcp.RA118.000696>.
- (18) Steckel, A.; Schlosser, G. Citrulline Effect Is a Characteristic Feature of Deiminated Peptides in Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2019**. <https://doi.org/10.1007/s13361-019-02271-x>.
- (19) Salinger, A. J.; Dubuke, M. L.; Carmona-Rivera, C.; Maurais, A. J.; Shaffer, S. A.; Weerapana, E.; Thompson, P. R.; Kaplan, M. J. Technical Comment on “Synovial Fibroblast-Neutrophil Interactions Promote Pathogenic Adaptive Immunity in Rheumatoid Arthritis.” *Science immunology*. 2020. <https://doi.org/10.1126/sciimmunol.aax5672>.
- (20) Budnik, B.; Levy, E.; Harmange, G.; Slavov, N. SCoPE-MS: Mass Spectrometry of Single Mammalian Cells Quantifies Proteome Heterogeneity during Cell Differentiation. *Genome Biol.* **2018**. <https://doi.org/10.1186/s13059-018-1547-5>.
- (21) Specht, H.; Emmott, E.; Petelski, A. A.; Huffman, R. G.; Perlman, D. H.; Serra, M.; Kharchenko, P.; Koller, A.; Slavov, N. Single-Cell Proteomic and Transcriptomic Analysis of Macrophage Heterogeneity Using SCoPE2. *Genome Biol.* **2021**, 22 (1), 50. <https://doi.org/10.1186/s13059-021-02267-5>.
- (22) Specht, H.; Harmange, G.; Perlman, D. H.; Emmott, E.; Niziolek, Z.; Budnik, B.; Slavov, N. Automated Sample Preparation for High-Throughput Single-Cell Proteomics. *bioRxiv* **2018**. <https://doi.org/10.1101/399774>.
- (23) Leduc, A.; Huffman, R. G.; Slavov, N. Droplet Sample Preparation for Single-Cell Proteomics Applied to the Cell Cycle. *bioRxiv* **2021**, 2021.04.24.441211. <https://doi.org/10.1101/2021.04.24.441211>.
- (24) Hartlmayr, D.; Ctortekca, C.; Seth, A.; Mendjan, S.; Tourniaire, G.; Mechtler, K. An Automated Workflow for Label-Free and Multiplexed Single Cell Proteomics Sample Preparation at Unprecedented Sensitivity. *bioRxiv* **2021**, 2021.04.14.439828. <https://doi.org/10.1101/2021.04.14.439828>.
- (25) Hicks, S. C.; Townes, F. W.; Teng, M.; Irizarry, R. A. Missing Data and Technical Variability in Single-Cell RNA-Sequencing Experiments. *Biostatistics* **2018**, 19 (4), 562–578. <https://doi.org/10.1093/biostatistics/kxx053>.
- (26) Linderman, G. C.; Zhao, J.; Roulis, M.; Bielecki, P.; Flavell, R. A.; Nadler, B.; Kluger, Y. Zero-Preserving Imputation of Single-Cell RNA-Seq Data. *Nat. Commun.* **2022**, 13 (1), 192. <https://doi.org/10.1038/s41467-021-27729-z>.
- (27) Kelly, R.; Zhu, Y.; Liang, Y.; Cong, Y.; Piehowski, P.; Dou, M.; Zhao, R.; Qian, W.-J.; Burnum-Johnson, K.; Ansong, C. Single Cell Proteome Mapping of Tissue Heterogeneity Using Microfluidic Nanodroplet Sample Processing and Ultrasensitive LC-MS. *J. Biomol. Tech.* **2019**, 30 (Suppl), S61–S61.
- (28) Kadiyala, A.; Kumar, A. Applications of Python to Evaluate Environmental Data Science Problems. *Environ. Prog. Sustain. Energy* **2017**, 36 (6), 1580–1586.

D.I.D.A.R.
Q.C.

Database search