MagLab FAIR Data Empowers 'Data Users'

Wenrong Chen[1], Xiaowen Liu[1,2]

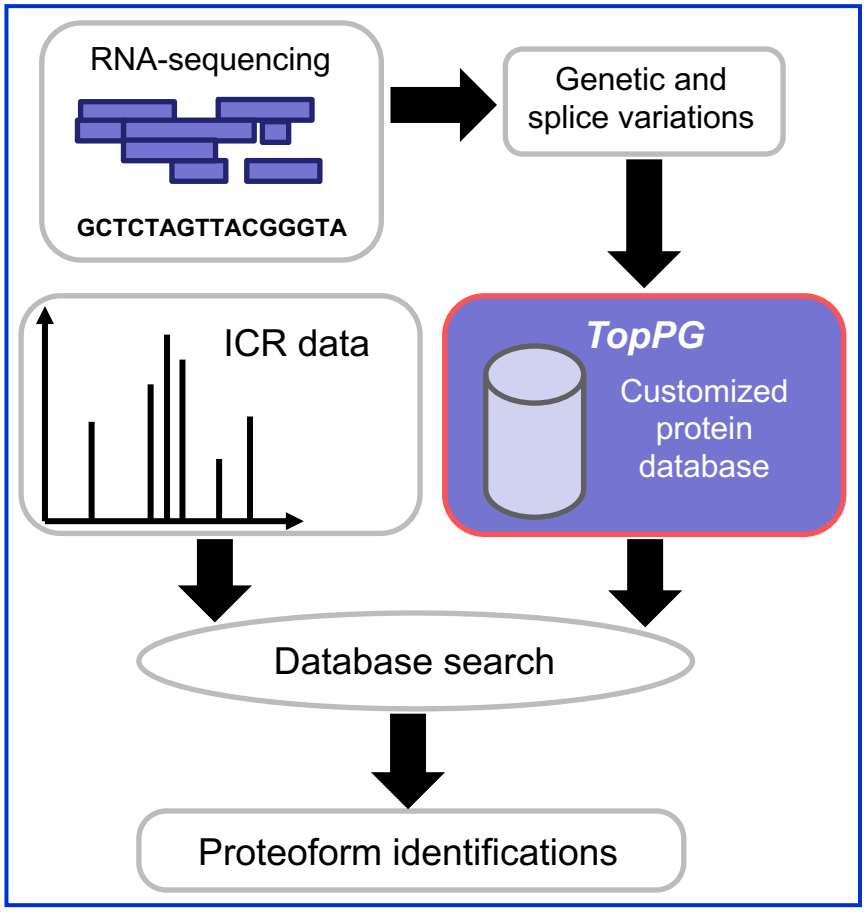
1. *Department of BioHealth Informatics, Indiana University—Purdue University Indianapolis*
2. *Center for Computational Biology and Bioinformatics, Indiana University School of Medicine*

Funding Grants: (Xiaowen Liu, PI) NIH R01GM118470; R01GM125991; R01AI14625; MagLab (G.S. Boebinger, NSF DMR-1644779)

*Recently, a new type of MagLab users - **'Data Users'** - accessed MagLab-generated top-down protein mass spectrometry data on DLD-1 colorectal cancer cells from the MassIVE data repository*, along with complementary RNA-sequencing data derived from the same cell line to perform a proteogenomics study. *Proteogenomics combines protein data with DNA and RNA data in order to improve identification of proteins with sample-specific sequence alterations, such as those caused by mutations and alternative splicing.*

The data had been originally published in 2017 as part of a benchmark study on 21T FT-ICR performance that produced unparalleled results on the number of intact proteins identified per experiment. Few proteogenomic studies have utilized top-down protein data, in which the exact protein forms (called *proteoforms*) are measured intact, as opposed to bottom-up protein data, which detects and identifies digested peptides and attempts to infer the original proteoform structures. Top-down data makes it possible to simultaneously identify all sequence variants and post-translational modifications for each observed protein, which elucidates combinatorial effects that cannot be captured by bottom-up proteogenomics.

The data were recently re-analyzed by a different set of users with a new proteogenomics software tool, *TopPG* (see figure), which revealed 112 proteoforms covering 43 single nucleotide variant events, and 128 proteoforms covering 131 splicing variations, including 13 novel events.



These 'Data Users' demonstrated that databases generated by TopPG facilitated identification of novel, sample-specific proteoforms, a discovery that will improve our understanding of biology, health, and disease.

Facility used: Ion Cyclotron Resonance (21 T FT-ICR)
Original Citation: *Journal of Proteome Research* **2017** 16 (2), 1087-1096 DOI: 10.1021/acs.jproteome.6b00696
Dataset: Mass Spectrometry Interactive Virtual Environment (MassIVE) ID: MSV000079978
Data User Citation: *Journal of Proteome Research* **2021** 20 (1), 261-269 DOI: 10.1021/acs.jproteome.0c00369