ProteomeGenerator: A framework for comprehensive proteomics based on de novo transcriptome assembly and high-accuracy peptide mass spectral matching

Paolo Cifani[1], Avantika Dhabaria[1], Zining Chen[1], Akihide Yoshimi[2], Emily Kawaler[3], Omar Abdel-Wahab[2,4], John T. Poirier[1,4*], and Alex Kentsis [1,5,6*]

[1] Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY.

[2] Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center New York, NY.

[3] Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY.

[4] Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Langone Health, New York, New York.

[5] Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY.

[6] Departments of Pediatrics, Pharmacology, and Physiology & Biophysics, Weill Cornell Medical College, Cornell University, New York, NY.

[*] To whom correspondence should be addressed: John T. Poirier (e-mail: poirierj@mskcc.org, phone nr.: +1 646 888 3588), Alex Kentsis (e-mail: kentsisresearchgroup@gmail.com, phone nr.: +1 646 888 3860).

TABLE OF CONTENTS

SUPPLEMENTARY FIGURES

Figure S1. Comparison of proteogenomic databases generated using ProteomeGenerator and QUILTS. A) Number of tryptic peptides in the MS search space defined using ProteomeGenerator and QUILTS (in white is indicated the fraction of non-canonical peptides, not mapping in UniProt). B) Overlap of tryptic peptide composition of PGX database (generated using ProteomeGenerator, in red), UniProt (grey), and QUILTS (dotted line). The square contains an enhanced view of the overlapping region.

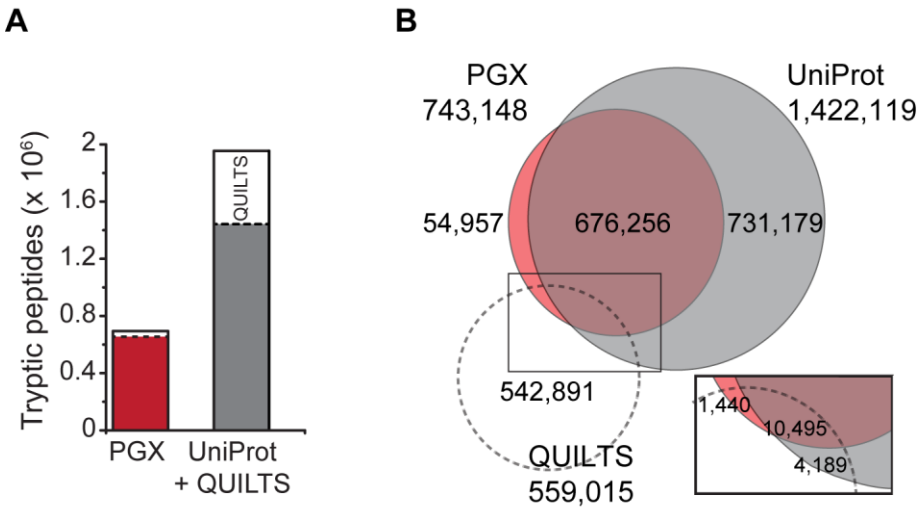


Figure S2. Subsampling analysis of (A) mass spectrometric and (B) RNA sequencing data. The plateau in the number of detected sequences indicates the limit of detection for the specific analytical method used, as observed for transcriptomic but not in proteomic data.

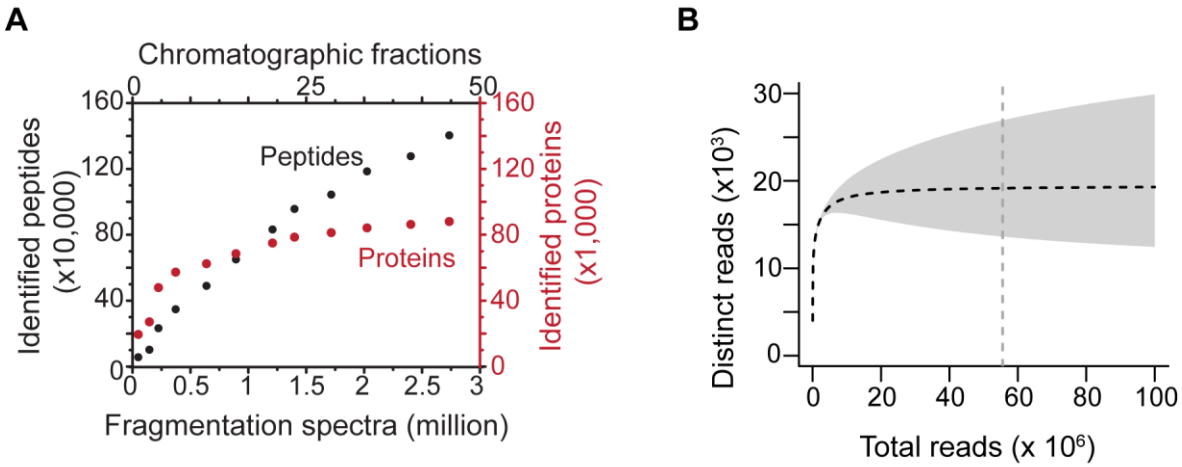


Figure S3. Composition of target and negative control protein databases. A) Amino acid frequency within each database. B) Aggregated frequency of amino acid with specific propertied. Aggregated frequency of Arg and Lys residues is plotted as a procy for the length of predicted tryptic peptides.

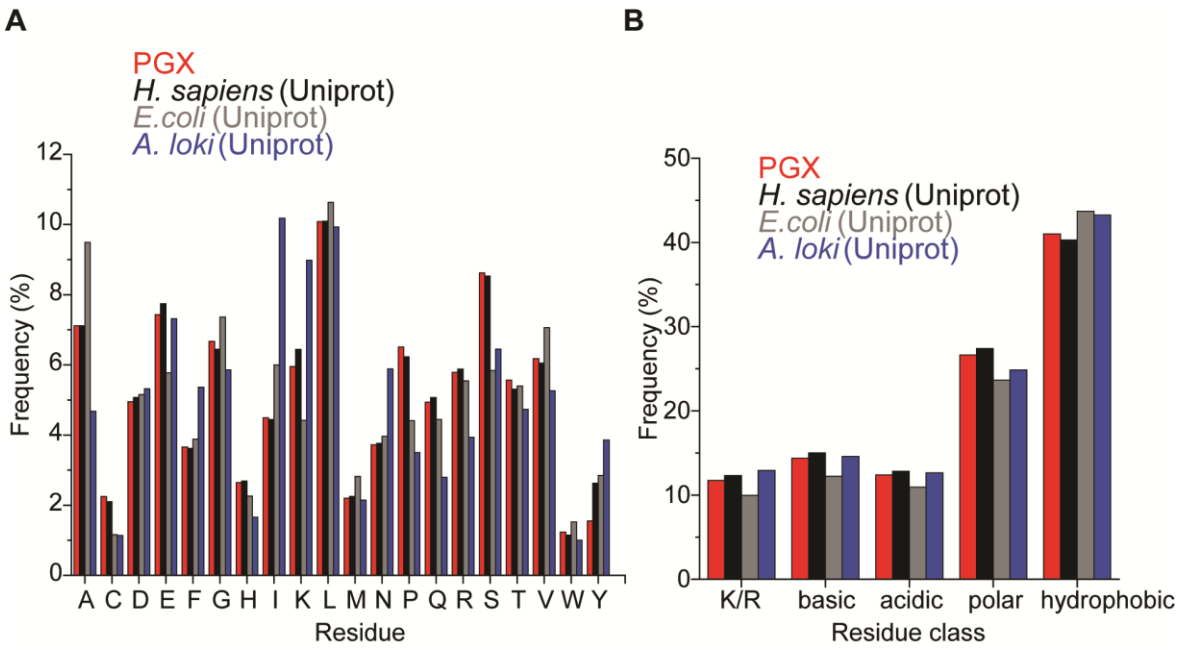


Figure S4. Frequency of isobaric peptides generated by the PGX (red) and UniProt (grey) databases, provided as proxy for likelihood of homeometric peptides.

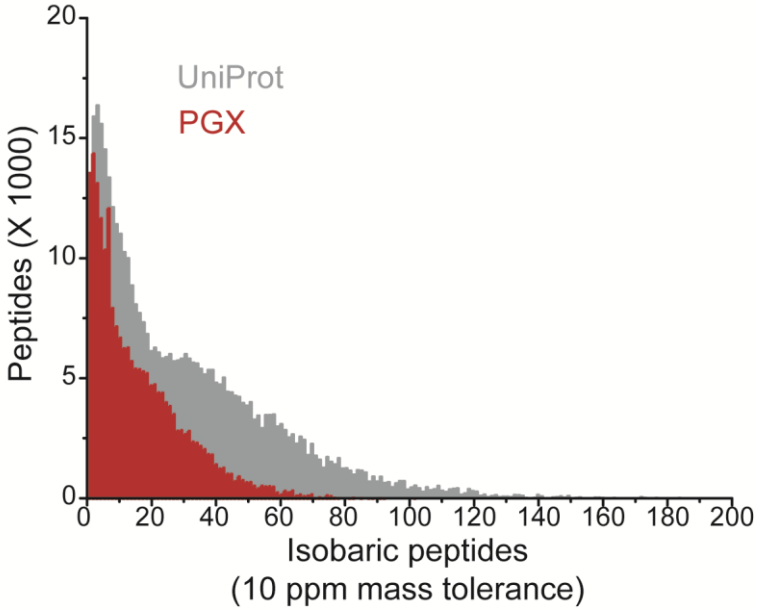


Figure S5. Comparison of peptides identified using as target database UniProt human proteome concatenated to ProteomeGenerator and QUILTS, respectively. A) Total number of peptide sequencies identified using ProteomeGenerator and QUILTS (PEAKS algoritm, FDR<0.01), showing superior sensitivity of ProteomeGenerator (sequences mapping in the proteogenomic database only, in UniProt only, and in both are indicated in red, gray, and black respectively). B) Overlap between non-canonical peptides identified using ProteomeGenerator and QUILTS. C) Loss of sensitivity arising from the use of concatenated PGX and UniProt databases, compared to setting each of the individually as target.

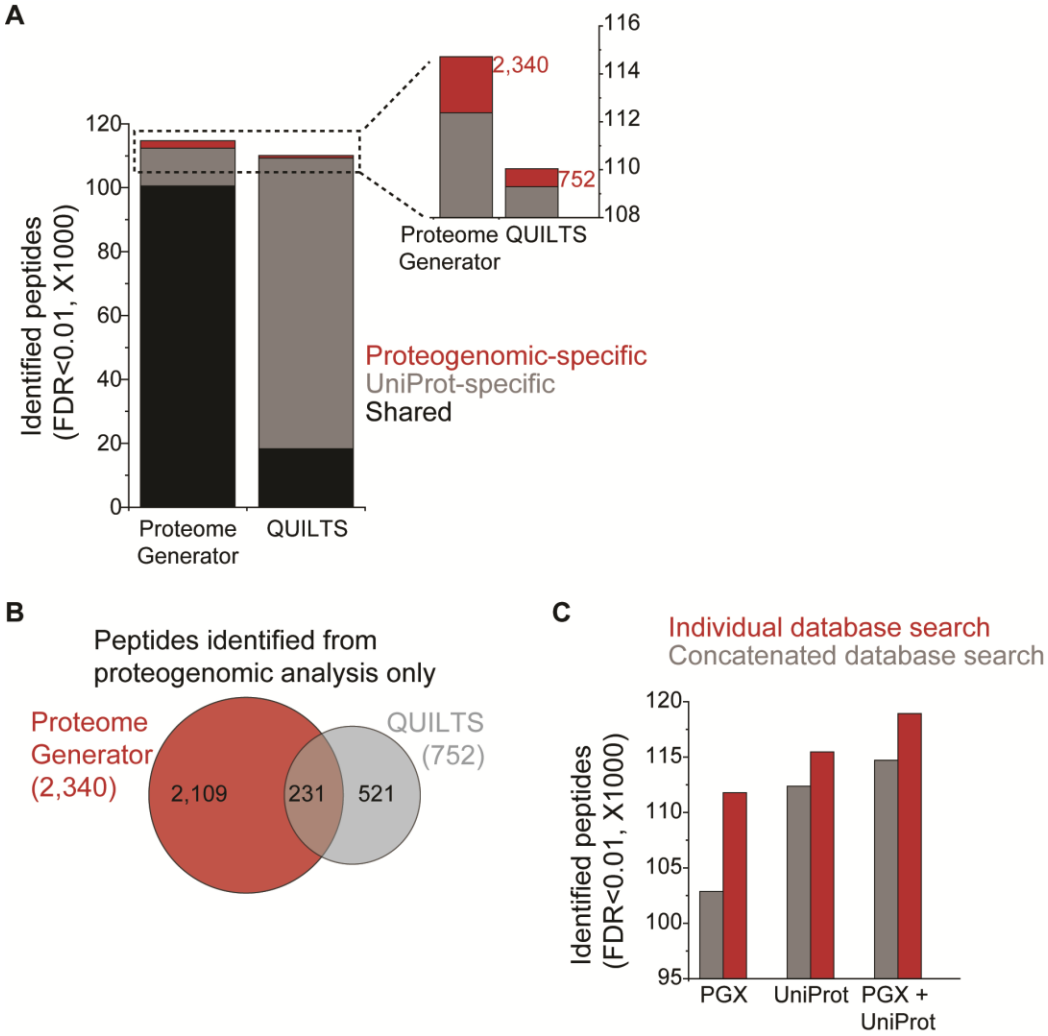
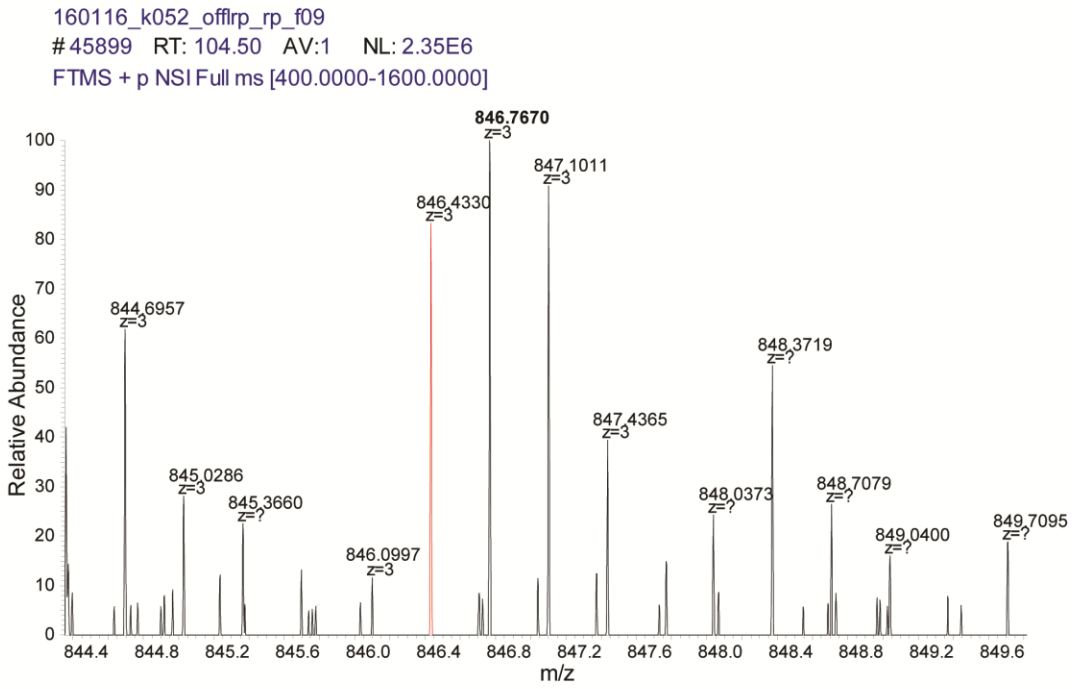
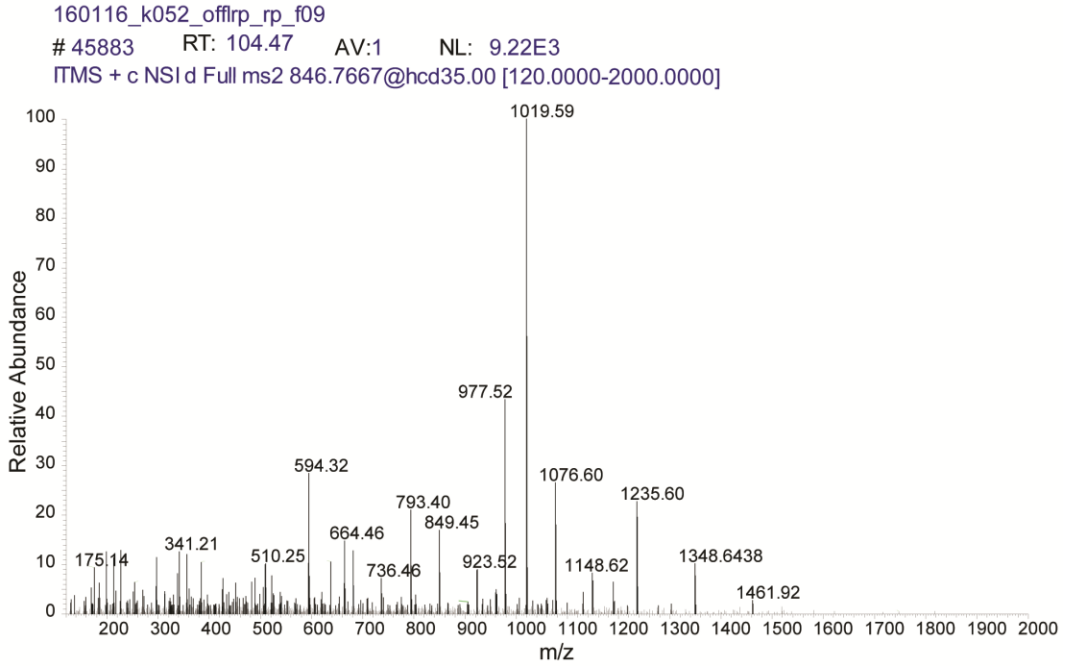


Figure S6. Raw spectra for peptide AGPDPGVSPAQVLLSEPEEAAALYR showing (A) precursor ion (in red: monoisotopic peak) within 5 ppm from expected m/z (846.4343). (B) Original fragmentation spectrum as per Figure 6C, without intensity axis resizing.

A



B



SUPPLEMENTARY TABLES

Table S1. Calibration of peptide-spectral matching using negative controls.

Table S2. Peptides matched to the PGX database at FDR<0.01, using PEAKS.

Table S3. Peptides matched to the UniProt database at FDR<0.01, using PEAKS.

Table S4. Identified peptides with PEAKS score higher than 50 and not mapping the reference UniProt database.

Table S5. Peptides matched to the PGX database at FDR<0.01, using MaxQuant.

Table S6. Peptides matched to the UniProt database at FDR<0.01, using MaxQuant.