Exploring missing proteins expression in gastric cancers and their potential as biomarkers

Chunhui Gu1, 2, Irajizad Ehsan1, 2, Yining Cai2, Fu-Chung Hsiao2, Samir Hanash2\*

1. The University of Texas MD Anderson Cancer Center, Department of Biostatistics

2. The University of Texas MD Anderson Cancer Center, Department of Clinical Cancer Prevention

# Abstract

**Background:** Missing proteins (MPs) are proteins lacking sufficient supporting evidence from mass spectrometry (MS) or other direct protein methods (Baker et al., 2017). The number of MPs has been constantly reduced due to the development of new detection techniques and through efforts from the growing community (Omenn et al., 2019). Gastric cancer accounts for 1.5% of all newly-diagnosed cancers in the united states (American-cancer-society, 2022) and exploring the fingerprints of proteins, including MPs, can help us to better understand gastric cancer.

**Objective:** To explore the distribution characteristics of expressed MPs in gastric-cancer primary cell samples and evaluate how it associates with non-missing (regular) proteins.

**Methods**: A total of 198 MPs were detected in 8 gastric-cancer primary cell samples. Normalized spectral abundance factors (NSAFs) (Paoletti et al., 2006) were calculated using MS spectral counts. Transcripts per million (TPM) (Conesa et al., 2016) were calculated using RNA-Seq count data from the same 8 samples and were matched with their NSAFs on the gene level to check the association between proteomics expression and DNA expression for both MPs and regular proteins (Edfors et al., 2016).

**Results**: Six of the eight samples showed a similar level of MP and regular protein detection (min-max range: [18, 39] for MPs and [4,157, 6,119] for regular proteins). The proportion of protein products with RNA products was 54.70% and 58.04% respectively for MPs with and without the two potential low-profiling samples for their noticeably fewer detected proteins, which were 90.38% and 91.02% for regular proteins. The protein expressions of MPs showed a clear truncated pattern by lacking low-abundance expression indicated by gap region in the low-end of distribution. There is a significant linear association between protein expression and RNA expression for MPs (R=0.17, p = 0.029) and regular proteins (R=0.38, p<2.2e-16). Several MP genes, such as CTAGE1, were consistently detected with protein products and their RNA products.

**Discussion:** The highly-truncated expression distribution pattern of MPs could not be completely explained by the insensitivity of count-based-MS proteomics in low-abundance proteins (Lundgren et al., 2010) by seeing only a mild truncated pattern in regular proteins. The significant association between RNA-Seq and proteomics suggests the validity of our findings. The detections of MPs, such as Q9HC47 (CTAGE1), were supported by the clear association between proteomics and RNA-Seq data in gastric cancer and should be further explored their potential as biomarkers in gastric cancer.

# Statistical Methods

After peptide spectral matching (PSM), the raw spectral count was acquired for each protein for each sample. Raw spectral count (SpC) data for cell components such as total cell extraction (TCE), media, nuclear, exosome, and surface were available. The availability of samples in data of different components was different (see supplementary table 1 for more details).

## Data normalization

The normalized spectral abundance factors (NSAFs) were calculated using the raw spectral count data (Zybailov et al., 2006). The NSAF for a given protein , is calculated by:

The length-adjusted spectral count for protein k is the total number of spectral counts matched to the protein divided by the amino acids length of the protein. This value is then divided by the sum of length-adjusted spectral counts of all N proteins in a sample to get NASF. The NASF used in this study was scaled by , which didn’t change the property and the reason for doing this will be mentioned later. The NASF can be used to measure the relative abundance of proteins within a sample and the relative abundance of a specific protein between samples, since it is normalized based on the sequencing depth of each sample and the length of the proteins (Neilson et al., 2013).

## Missing proteins distribution for components

## Due to the relatively insensitive detection of low-abundance proteins for the SpCs-based label-free proteomics technique (Lundgren et al., 2010), a pre-filtering was applied to control the false discovery rate (FDR) by only keeping proteins with valid count reads in a certain number of samples (Bourgon et al., 2010). The application of pre-filtering is highly recommended in widely used high-throughput count data analysis packages, such as the edgeR (Yunshun Chen, 2022) and DESeq2 (Michael I. Love, 2022). Considering the number of samples in total and in each cancer type are different for different cell line component datasets, the pre-filtering approach we used needed to control for those factors to make expressed proteins from different component datasets comparable. For example, the pre-filtering in DESeq2 only keeps rows with a total sum greater than 10. For the same truly expressed protein in different component datasets, the dataset with more samples in it is obviously more likely to keep the protein left after pre-filtering using this approach than the component dataset with far fewer samples. The pre-filtering we used for this section was keeping a protein if only it had a valid abundance value (NSAF > 0) in at least N samples, where N was the number of samples in the smallest cancer type group in that dataset (Yunshun Chen, 2022).

## For each cell line component data,

## Correlation between RNA-Seq data

# Result

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **IP0981** | **IP0982** | **IP0993** | **IP0995** | **IP0999** | **IP7100** | **IP7103\*** | **IP7105** | **Total**  **with outlier** | **Total without outlier** |
| **Protein** | **RNA product (+)** | 8   (50.00) | 38  (60.32) | 24  (60.00) | 17  (56.67) | 22  (64.71) | 17  (65.38) | 7  (30.43) | 25  (71.4%) | 158  (57.88) | 151  (60.40) |
| **RNA product (-)** | 8 | 25 | 16 | 13 | 12 | 9 | 16 | 16 | 115 | 99 |
| **Total protein products** | 16 | 63 | 40 | 30 | 34 | 26 | 23 | 41 | 273 | 250 |
| **RNA** | **Protein product (+)** | 8 (1.94% | 38 (9.18%) | 24  (5.16%) | 17  (4.47%) | 22  (4.26%) | 17  (3.48%) | 7  (2.37%) | 25  (5.39%) | 158  (4.60%) | 151  (4.81%) |
| **Protein protduct (-)** | 404 | 376 | 441 | 363 | 495 | 471 | 288 | 439 | 3277 | 2989 |
| **Total RNA products** | 412 | 414 | 465 | 380 | 517 | 488 | 295 | 464 | 3435 | 3140 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **IP0981** | **IP0982** | **IP0993** | **IP0995** | **IP0999** | **IP7100** | **IP7103\*** | **IP7105** | **Total**  **with outlier** | **Total without outlier** |
| **Protein** | **RNA product (+)** | 3155 (90.43%) | 6028  (88.99%) | 6119  (91.02%) | 4730  (90.94%) | 5162  (92.20%) | 4157  (91.6%) | 3693  (85.17%) | 5465  (91.88%) | 38,509  (90.38%) | 34816 (90.97%) |
| **RNA product (-)** | 334 | 746 | 604 | 471 | 437 | 381 | 643 | 483 | 4099 | 3456 |
| **Total protein products** | 3,489 | 6,774 | 6,723 | 5,201 | 5,600 | 4,538 | 4,336 | 5,948 | 42,608 | 38,272 |
| **RNA** | **Protein product (+)** | 3,155  (13.66%) | 6,028  (26.38%) | 6,119  (25.72%) | 4,730  (20.88%) | 5,162  (20.40%) | 4,157  (17.26%) | 3,693  (21.37%) | 5,465  (22.72%) | 38,509  (21.03%) | 34,816  (20.99%) |
| **Protein (-)** | 19,940 | 16,823 | 17,675 | 17,925 | 20,144 | 19,923 | 13,590 | 18,589 | 144,609 | 131,019 |
| **Total RNA products** | 23,095 | 22,.851 | 23,794 | 22,655 | 25,306 | 24,080 | 17,283 | 24,054 | 183,118 | 165,835 |

# Supplementary statistical methods

**b**

**a**

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

# Chart, scatter chart Description automatically generated Chart, scatter chart Description automatically generated

**d**

**c**

# 

American-cancer-society. (2022). *Key statistics about stomach cacner*. Retrieved 11/15 from <https://www.cancer.org/cancer/stomach-cancer/about/key-statistics.html#:~:text=The%20American%20Cancer%20Society's%20estimates,6%2C690%20men%20and%204%2C400%20women>)

Baker, M. S., Ahn, S. B., Mohamedali, A., Islam, M. T., Cantor, D., Verhaert, P. D., Fanayan, S., Sharma, S., Nice, E. C., Connor, M., & Ranganathan, S. (2017). Accelerating the search for the missing proteins in the human proteome. *Nature Communications*, *8*(1), 14271. <https://doi.org/10.1038/ncomms14271>

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, *107*(21), 9546-9551. <https://doi.org/10.1073/pnas.0914005107>

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1). <https://doi.org/10.1186/s13059-016-0881-8>

Edfors, F., Danielsson, F., Hallström, B. M., Käll, L., Lundberg, E., Pontén, F., Forsström, B., & Uhlén, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol*, *12*(10), 883. <https://doi.org/10.15252/msb.20167144>

Lundgren, D. H., Hwang, S. I., Wu, L., & Han, D. K. (2010). Role of spectral counting in quantitative proteomics. *Expert Rev Proteomics*, *7*(1), 39-53. <https://doi.org/10.1586/epr.09.69>

Michael I. Love, S. A., and Wolfgang Huber. (2022). *Analyzing RNA-seq data with DESeq2*. Retrieved 10/10 from <https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>

Neilson, K. A., Keighley, T., Pascovici, D., Cooke, B., & Haynes, P. A. (2013). Label-Free Quantitative Shotgun Proteomics Using Normalized Spectral Abundance Factors. In (pp. 205-222). Humana Press. <https://doi.org/10.1007/978-1-62703-360-2_17>

Omenn, G. S., Lane, L., Overall, C. M., Corrales, F. J., Schwenk, J. M., Paik, Y. K., Van Eyk, J. E., Liu, S., Pennington, S., Snyder, M. P., Baker, M. S., & Deutsch, E. W. (2019). Progress on Identifying and Characterizing the Human Proteome: 2019 Metrics from the HUPO Human Proteome Project. *J Proteome Res*, *18*(12), 4098-4107. <https://doi.org/10.1021/acs.jproteome.9b00434>

Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S., Zhu, D., Conaway, R. C., Conaway, J. W., Florens, L., & Washburn, M. P. (2006). Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proceedings of the National Academy of Sciences*, *103*(50), 18928-18933. <https://doi.org/10.1073/pnas.0606379103>

Yunshun Chen, D. M., Matthew Ritchie, Mark Robinson, Gordon Smyth. (2022). edgeR: differential analysis

of sequence read count data user’s guide. Retrieved 10/10/2022, from

Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., & Washburn, M. P. (2006). Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res*, *5*(9), 2339-2347. <https://doi.org/10.1021/pr060161n>