Exploring missing proteins expression in gastric cancers

Chunhui Gu1, 2, Irajizad Ehsan1, 2, Yining Cai2, Fu-Chung Hsiao2, Jennifer Dennison2, Jody Vykoukal2, Hiro Katayama2, Johannes Farhmann2, Kim-Anh Do1, 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 or other direct protein methods [1]. The number of MPs have been constantly reduced due to the development of new detection techniques and through efforts from the growing community [2]. Gastric cancer accounts for 1.5% of all newly-diagnosed cancers in the united states [3] and better understanding of the expression fingerprint of proteins, including MPs, can help us to better understand the gastric cancer.

**Objective**

To explore the distribution characteristics of expressed missing proteins in gastric cancer primary cell samples and evaluate how it is differ from regular proteins.

**Methods**

A total of 198 MPs were detected in 8 different gastric-cancer primary cell samples. Normalized spectral abundance factors (NSAFs) [4] were calculated using mass spectrometry (MS) spectral counts (SCs). Transcripts per million (TPM) [5] were calculated using RNA-seq count data from the same 8 samples and were matched with their MS proteomics data to check the association between proteomics expression and DNA expression for both MPs and regular proteins on the gene level [6].

**Results**

Six of the eight samples showed a similar level of missing protein detection (Table 1). The proportion of protein products with corresponding RNA products was 54.70% and 58.04% respectively for missing proteins with and without the two potential outliers (sample 1 and sample 7 for their noticeably fewer detected proteins), which in comparison were 90.38% and 91.02% for regular proteins. The protein expressions of MPs showed an clear truncated pattern with lacking of low-abundance expression (indicated by a large gap region in below red dot line in Figure 1a). There is a significant linear association between protein expression and RNA expression for missing proteins with RNA products (R=0.17, p = 0.029) but not as strong as in regular proteins (R=0.38, p<2.2e-16). Missing proteins gene CTAGE1 were detected and had RNA products in all 6 non-outlier samples.

**Discussion**

The highly-truncated expression distribution pattern of missing proteins could be partially explained by the insensitivity of count-based-MS proteomics in low-abundance proteins [7]. However, this cannot explain alone by seeing only a mild truncated pattern in regular proteins. The either-feast-or-famine expression pattern in missing proteins in this study suggested that the detected missing protein is very likely to be disease-specific proteins. The significant association between RNA-Seq and proteomics. The detections of missing proteins, 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 screening.

Table 1. The proportions of proteins products with RNA products for missing proteins and regular proteins

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Sample 1** | **Sample 2** | **Sample 3** | **Sample 4** | **Sample 5** | **Sample 6** | **Sample 7** | **Sample 8** | **Total**  **with outlier** | **Total without outlier** |
| **Missing Proteins** | **RNA product (+)** | 8  (50.00%) | 39  (59.09%) | 24  (55.81%) | 20  (55.56%) | 22  (57.89%) | 18  (64.29%) | 7  (25.93%) | 25  (56.82%) | 163  (54.70%) | 148  (58.04%) |
| **RNA product (-)** | 8 | 27 | 19 | 16 | 16 | 10 | 20 | 19 | 135 | 107 |
| **Total protein products** | 16 | 66 | 43 | 36 | 34 | 28 | 27 | 41 | 298 | 255 |
| **Regular Proteins** | **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%) | 31,661 (91.02%) |
| **RNA product (-)** | 334 | 746 | 604 | 471 | 437 | 381 | 643 | 483 | 4099 | 3122 |
| **Total protein products** | 3,489 | 6,774 | 6,723 | 5,201 | 5,600 | 4,538 | 4,336 | 5,948 | 42,608 | 34,783 |

**\* Sample 1 and Sample 7 were potential outliers sample 1 and sample 7 for their noticeably fewer detected proteins.**

Graphical user interface

Description automatically generated with medium confidence

Figure 1. The scatter plot of protein-RNA-product matched pairs. Top left: All protein-RNA pairs for missing proteins. Top right: protein-RNA pairs with TPM > 0 and NSAF > 0. Bottom left: All protein-RNA pairs for regular proteins. Bottom right: All protein-RNA pairs for regular proteins with TPM > 0 and NSAF > 0.

1. Baker, M.S., et al., *Accelerating the search for the missing proteins in the human proteome.* Nature Communications, 2017. **8**(1): p. 14271.

2. Omenn, G.S., et al., *Progress on Identifying and Characterizing the Human Proteome: 2019 Metrics from the HUPO Human Proteome Project.* J Proteome Res, 2019. **18**(12): p. 4098-4107.

3. American-cancer-society. *Key statistics about stomach cacner*. 2022 [cited 2022 11/15]; Available 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>).

4. Paoletti, A.C., et al., *Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors.* Proceedings of the National Academy of Sciences, 2006. **103**(50): p. 18928-18933.

5. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biology, 2016. **17**(1).

6. Edfors, F., et al., *Gene-specific correlation of RNA and protein levels in human cells and tissues.* Mol Syst Biol, 2016. **12**(10): p. 883.

7. Lundgren, D.H., et al., *Role of spectral counting in quantitative proteomics.* Expert Rev Proteomics, 2010. **7**(1): p. 39-53.