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 (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 missing proteins 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 different gastric-cancer primary cell samples. Normalized spectral abundance factors (NSAFs) (Paoletti et al. 2006) were calculated using mass spectrometry (MS) spectral counts (SCs). 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 MS proteomics data to check the association between proteomics expression and DNA expression for both MPs and regular proteins on the gene level (Edfors et al. 2016).

**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 low-profiling samples (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 a clear truncated pattern by lacking low-abundance expression (indicated by a large gap region below the red dot line in Figure 1a). There is a significant linear association between protein expression and RNA expression for missing proteins (R=0.17, p = 0.029) and regular proteins (R=0.38, p<2.2e-16). Missing gene-protein product CTAGE1 was detected and had RNA products in all 6 non-low-profiling 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 (Lundgren et al. 2010). However, this cannot explain alone 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 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.

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** | **Total without low-profiling samples** |
| **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 considered as potential low-profiling samples for their noticeably fewer detected proteins.**

Graphical user interface

Description automatically generated

Figure 1. The scatter plot of protein-RNA-product matched pairs. (a): All protein-RNA pairs for missing proteins. (b): protein-RNA pairs with TPM > 0 and NSAF > 0. (c): All protein-RNA pairs for regular proteins. (d): All protein-RNA pairs for regular proteins with TPM > 0 and NSAF > 0.

Reference:

Key statistics about stomach cacner. 2022. American Cancer Society; [accessed 2022 11/15]. <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 MS, Ahn SB, Mohamedali A, Islam MT, Cantor D, Verhaert PD, Fanayan S, Sharma S, Nice EC, Connor M et al. 2017. Accelerating the search for the missing proteins in the human proteome. Nature Communications. 8(1):14271.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X et al. 2016. A survey of best practices for rna-seq data analysis. Genome Biology. 17(1).

Edfors F, Danielsson F, Hallström BM, 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.

Lundgren DH, Hwang SI, Wu L, Han DK. 2010. Role of spectral counting in quantitative proteomics. Expert Rev Proteomics. 7(1):39-53.

Omenn GS, Lane L, Overall CM, Corrales FJ, Schwenk JM, Paik YK, Van Eyk JE, Liu S, Pennington S, Snyder MP et al. 2019. Progress on identifying and characterizing the human proteome: 2019 metrics from the hupo human proteome project. J Proteome Res. 18(12):4098-4107.

Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D, Conaway RC, Conaway JW, Florens L, Washburn MP. 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.