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

Chunhui Gu1, 2, Ehsan Irajizad 1, 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 (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 association between protein expression and RNA expression for MPs (rho=0.2, p = 0.01) and regular proteins (rho=0.3, 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 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.

# Background

The advance of mass spectrometry (MS)-based high throughput proteomics made it possible to spontaneously identify and quantify an unprecedentedly wide range of proteins. A considerable meanwhile credible protein parts list in humans is crucial for the understanding of protein functions and interaction between proteins. In 2010, the Human Proteome Project (HPP) was launched by the Human Proteome Organization (HUPO) as a flagship project for fulfilling this goal (Deutsch et al., 2016; Omenn et al., 2021). The protein parts list from the Chromosome-centric part of HPP (C-HPP) was curated as the netXProt knowledgebase. In the neXtProt, proteins were classified by five levels of protein evidence: Evidence at protein level (PE1), Evidence at transcript level (PE2), Inferred from homology (PE3), Predicted (PE4) and Dubious (PE5) (Gaudet et al., 2015).

The missing proteins (MPs) are proteins lacking sufficient supporting evidence from mass spectrometry or other direct protein methods (Baker et al., 2017). From HPP, the MPs refer to PE2-4 proteins in the neXtProt (Omenn et al., 2019). PE5 proteins are considered highly unlikely to be translated (Baker et al., 2017) and not included as the MPs. The definition above is sometimes called the narrow definition of the MP while a broad definition can refer to the large fraction of the proteomics that is not revealed in a single proteomics experiment run. In this study, we use the narrow definition of the MPs. Many factors could contribute to why MP is hard to be detected by MS-based proteomics due to the sophisticated multi-step processing from raw MS data to final data at protein level. Biology factors such as typical low-abundance, sequence ambiguity (peptide can be matched to more than one proteins), or high expression variation between different cells or tissues (Baker et al., 2017; Nesvizhskii, 2010; Zhou et al., 2018) are especially crucial for causing MPs to be missing.

So far, 18,407 (90.4%) of the 20,359 proteins that were predicted to be coded by the human genome have been credibly detected (PE1, called regular proteins in this study) (neXtProt, 2022; Omenn et al., 2021). The number of MPs has been constantly reduced due to the development of new detection techniques and through efforts from the growing community from 5,511 in 2012-02 to 1,343 in 2022-10 (Omenn et al., 2021). Given the current progress, it was estimated that 95% completion of the protein list by around 2030-40 by PeptideAtlas (Baker et al., 2017), and the problem of MPs is deemed as solvable in a long run (Xu et al., 2015).

Globally, gastric cancer is the sixth most common cancer (1.09 million new cases) and the fourth leading cause of cancer death (769, 000 deaths) in 2020 (Bray F, 2018; Ferlay J, 2018). Gastric cancer accounts for 1.5% of all newly-diagnosed cancers in the united states (American-cancer-society, 2022). For lacking effective screening and early-diagnosis approaches, gastric cancers are often diagnosed at an advanced stage and are aggressive with poor 5-year survival (Li et al., 2022). Conventional biomarkers uncovered yet for gastric cancer, such as CEA and CA199, are not highly specific and can only detect gastric cancer at advanced stages. As a result, novel biomarkers, such as cancer-related genes, circulating cell-free DNA, and MicroRNA, are proposed but their application in the clinical setting is still debatable (Durães et al., 2014; Matsuoka & Yashiro, 2018). Compared to genomics, proteomics could reflect more accurately about what is the status quo in our body, which leads to potential biomarkers or treatment targets. Exploring the fingerprints of proteins, including MPs, can help us to better understand gastric cancer.

In this study, we tried to explore the distribution characteristics of expressed MPs in gastric-cancer primary cell samples and evaluate how it associates with regular proteins.

# Statistical Methods

## Protein score and false discovery rate calculation

Need additional information

## Data processing

After peptide spectral matching (PSM), the raw spectral counts (SpCs) were acquired for each protein for each sample (Lundgren et al., 2010). A 4% global false discovery rate (FDR) was used for controlling all detected proteins. “One-hit-wonders” refer to proteins that are identified by only a single peptide. “One-hit-wonders” complexify protein inference when the single peptide is shared by multiple proteins and how to treat them is still an open debate (Gupta & Pevzner, 2009; Huang et al., 2012; Veenstra et al., 2004). To accommodate this, we used both the data with and without the two-SpC-rule[[1]](#footnote-1) (keep a protein if only it has at least two spectral counts).

## Missing proteins query

Missing proteins (MPs) are defined as protein entries that belong to PE2 (Evidence at transcript level), PE3 (Inferred from homology), and PE4 (Predicted) categories in neXtProt (Omenn et al., 2019). The query for MPs was done by using “queryId=NXQ\_00204” in advanced searching in neXtProt. In total, 1343 MPs (1135 in PE1, 195 in PE2, and 13 in PE4) were retrieved (see supplementary file 1).

## 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 to match the magnitude of RNA data and didn’t change its property. 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).

## Supporting missing proteins by RNA-Seq data

By Central Dogma, proteins are translated from RNA, which means theoretically whenever there is a protein product it should have a corresponding RNA product.

The relation between protein product and RNA product in a sample can be described in a way like a 2X2 contingency table (Table 2). A protein-RNA-pair-at-gene-level (protein-RNA-pair for short) could fall into one of four cases jointly defined by the detection of the protein product and the detection of its RNA product. In case 1, both the protein product (+) and RNA product (+) were detected. In case 2, a protein product (+) is detected but not its RNA product (-). Case 1 is what is expected from Central Dogma while case 2 is the opposite. Case 3 simply stands for the truth that not every RNA product will be translated into proteins and is not the focus of this study. Unlike a real contingency table, the number of case 4 where both protein product and RNA product were not detected is unknown due to the number of expressed genes in a sample is not fixed.

We were interested in the proportion of MP products supported by their corresponding RNA product, which is case 1 / (case 1 + case 2). We wished it could indirectly support the detection of those MPs from another perspective. However, even with perfect detection, it is impossible to eliminate all case 2 protein-RNA-pair due to the different disintegration rates of RNA and protein (Ron Milo, 2015). Therefore, the relative comparison of the proportion of protein products with RNA products between MPs and regular proteins was more important than their absolute values.

Table 1. Relation between a protein product and its RNA product

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RNA product (+)** | **RNA product (-)** |  |
| **Protein product (+)** | Case 1 (desired) | Case 2 (not desired) | Case1 + Case2 |
| **Protein product (-)** | Case 3 (common and not interested) | Case 4 (hard to detect) | --- |
|  | Case 1 + Case 3 | --- | --- |

Scatterplots were used to check the distribution of case 1, 2 and 3 for both MPs and regular proteins. Spearman correlations between RNAs and proteins were calculated for RNA-supported proteins (case 1 proteins).

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 (gene symbol ID) to check the association between proteomics expression and DNA expression for both MPs and regular proteins (Edfors et al., 2016). R “BiomaRt” package was used to map ensemble ID to gene symbol ID for RNA-Seq data (see supplementary statistical methods for more details).

an RNA product of TPM > 0 was considered to be expressed, and an NSAF > 0 was used for considering a protein product to be expressed. Obviously, those were not strict rules for deciding the expression of RNA and protein. However, the main objective is the relative difference in expression patterns for MPs and regular proteins. So, as long as the same rule was used for MPs and regular proteins, it should be fine and a relatively loose rule help to keep more information for sake of the exploratory nature of this study.

# Result

## Detection of missing proteins

There was only a minor difference in the FDR distributions between MPs and regular proteins (Figure 1). However, the proportion of single-count proteins in the MPs was higher than in regular proteins and overall has lower SpCs (Supplementary Figure S1).

When using data without the “two-SpC-rule”, a total of 163 MP products were detected in 8 gastric-cancer primary cell samples. 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) (Table 2). However, two samples (Sample 1 and Sample 7) had noticeably fewer detected proteins both for MPs (8 and 7) and regular proteins (3155 and 3693) and were considered as potential low-profiling samples. The proportion of protein products with RNA products was 54.70% and 58.04% respectively for MPs with and without the two low-profiling samples, which were 90.38% and 91.02% for regular proteins. The results were similar when using data with the “two-SpC-rule” (Table S2). However, the number of detected protein products decrease much more for MPs than regular proteins (25% loss from 148 to 111 for MPs and only 9% loss from 31,661 to 28,798 for regular proteins).

## Correlation of protein and RNA

When using data without the “two-SpC-rule”, the protein expressions of MPs showed a clear truncated pattern by lacking low-abundance expression (NSAF < 3) indicated by a large gap region in the low end of the distribution (region below the red dotted line in Figure 1a). There were significant associations between protein expression and RNA expression for MPs (rho=0.2, p = 0.01) and regular proteins (rho=0.3, p<2.2e-16). The results from data with “two-SpC-rule” were similar except the truncated pattern observed only for MPs now can be seen in regular proteins too (Figure S2).

## Expressed missing proteins and regular proteins behavior

Using data without the “two-SpC-rule”, 30 MPs were decided to be expressed by the criterion that it must had non-zero TPM and NSAF values in at least two of the non-low-profiling six samples and. Using a similar criterion but require expression in all 6 samples, 2,375 regular proteins were deemed as expressed. For each detected MP, the RNA and protein expressions of individual samples were close to each other (Figure 3). The full list of expressed missing protein can be found in supplementary file 2.

Calendar

Description automatically generated

Figure 3: The distribution of NSAF and TPM for detected missing proteins in the gastric cancer primary cell. 30 unique missing protein were detected with a total of 63 protein products (represented by points in the scatterplot). For each detected MP, the RNA and protein expressions of individual samples were close to each other (points under the same label were clustered with each other). The figure only showed label for missing proteins that is expressed in at least 3 samples of the six non-low-profiling samples for clarity, but the results for other missing proteins were similar.

(need to redraw the plot to avoid label covering point)

# 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 “either-feast-or-famine expression” pattern in MPs in this study suggested that the detected MPs were likely to be gastric-cancer-specific proteins.

The significant association between RNA-Seq and proteomics suggests the validity of our findings. The detections of MPs, such as 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.

The more one-wonder-hit proteins and overall lower SpC distribution in the MPs ask for special consideration about the number of peptides when controlling FDR while detecting more truly expressed MPs (a lower false negative rate FNR). There is yet no common belief about how such information should be used in controlling FDR. Although several algorithms were proposed to try to solve this issue by incorporating such information in calculating protein scores and then FDR, they are not perfect (Gupta & Pevzner, 2009; The et al., 2016). The difference in SpC distributions between MPs and regular proteins could also imply different protein score distributions. Separately modeling protein scores and controlling FDR could also be a solution for escalating missing protein detection while keeping low false discoveries in regular proteins like in (Wu et al., 2018).

Despite that we had tried to address as many important issues as possible, there are still some weaknesses of this study. Currently, there were around 3% (43 out of 1343) MPs that were not assigned with a gene symbol ID in the neXtProt, and the relationship between those MPs and their corresponding RNA expression could not be fathomed by this strategy. Another concern was how well RNA expression can support protein detection since RNA and protein certainly have different lifespans. Typically, mRNAs degrade at a much fast speed (3-8 minutes for the majority of mRNAs) than proteins (only 2-7% of proteins are actively degraded) (Ron Milo, 2015). Therefore, simply using RNA expression at a single-time point could lead to a more conservative protein detection and using longitudinal multi-omics analysis may further improve the results (Zhou et al., 2019).

Our study has shown the potential of integrating information from MPs and transcriptomics for a better understanding of gastric cancer genesis and development. The credible detection of MPs in patient gastric cancer tissue is the major barrier to explore how what roles MPs play in gastric cancer. Future studies shall continue focusing on improving the credible detection of MPs in gastric cancer by integrating multi-omics data and protein-protein interaction.

Chart

Description automatically generated

Figure 1 (need a better figure): False discovery rate (FDR) for missing proteins and regular proteins.

Table 2. 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.**

**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**

Figure 2. 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. Blue line: Simple linear regression. Spearman correlations and corresponding p-value were calculated for (b) and (d).

Supplementary figures:

Chart, histogram

Description automatically generatedChart

Description automatically generated

Figure S1: The peptide-spectral-matching (PSM) count distribution for regular proteins (Left blue) and missing proteins (Right red)

Table S1. The proportions of protein products with RNA products for missing proteins and regular proteins after only keeping proteins with spectral count >= 2.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **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 (+)** | 5  (50.00%) | 34  (69.39%) | 18  (58.06%) | 14  (50.00%) | 14  (60.87%) | 14  (73.68%) | 3  (17.65%) | 17  (62.96%) | 119  (58.33%) | 111  (62.71%) |
| **RNA product (-)** | 5 | 15 | 13 | 14 | 9 | 5 | 14 | 10 | 85 | 66 |
| **Total protein products** | 16 | 49 | 31 | 28 | 23 | 19 | 17 | 27 | 204 | 177 |
| **Regular Proteins** | **RNA product (+)** | 2,640  (90.32%) | 5,545  (89.62%) | 5,661  (91.45%) | 4,297  (91.44%) | 4,548  (92.76%) | 3,762  (92.05%) | 3,325  (86.21%) | 4,985  (92.49%) | 34,763  (90.92%) | 28,798  (91.55%) |
| **RNA product (-)** | 283 | 642 | 529 | 402 | 355 | 325 | 532 | 405 | 3,473 | 2,658 |
| **Total protein products** | 2,923 | 6,187 | 6,190 | 5,201 | 4,903 | 4,087 | 3,857 | 5,390 | 38,236 | 31,456 |

**\* Sample 1 and Sample 7 were considered as potential low-profiling samples for their noticeably fewer detected proteins.**

**b**

**a**

Chart, scatter chart

Description automatically generated Chart

Description automatically generated

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

**d**

**c**

Figure S2. 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. Blue line: Simple linear regression. Spearman correlations and corresponding p-value were calculated for (b) and (d).

# Reference

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>

Bray F, F. J., Soerjomataram I, Siegel RL, Torre LA, Jemal A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, *68(6):394–424*. <https://doi.org/https://doi.org/10.3322/caac.21492> PMID:30207593

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>

Deutsch, E. W., Overall, C. M., Van Eyk, J. E., Baker, M. S., Paik, Y.-K., Weintraub, S. T., Lane, L., Martens, L., Vandenbrouck, Y., Kusebauch, U., Hancock, W. S., Hermjakob, H., Aebersold, R., Moritz, R. L., & Omenn, G. S. (2016). Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *Journal of Proteome Research*, *15*(11), 3961-3970. <https://doi.org/10.1021/acs.jproteome.6b00392>

Durães, C., Almeida, G. M., Seruca, R., Oliveira, C., & Carneiro, F. (2014). Biomarkers for gastric cancer: prognostic, predictive or targets of therapy? *Virchows Arch*, *464*(3), 367-378. <https://doi.org/10.1007/s00428-013-1533-y>

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>

Ferlay J, C. M., Soerjomataram I, Mathers C, Parkin DM, Piñeros M, Znaor A, Bray F. (2018). Estimating the global cancer incidence and mortality in 2020: GLOBOCAN sources and methods. *Int J Cancer*, *144(8):1941–1953*. <https://doi.org/https://doi.org/10.1002/ijc.31937> PMID:30350310

Gaudet, P., Michel, P.-A., Zahn-Zabal, M., Cusin, I., Duek, P. D., Evalet, O., Gateau, A., Gleizes, A., Pereira, M., Teixeira, D., Zhang, Y., Lane, L., & Bairoch, A. (2015). The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Research*, *43*(D1), D764-D770. <https://doi.org/10.1093/nar/gku1178>

Gupta, N., & Pevzner, P. A. (2009). False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule. *Journal of Proteome Research*, *8*(9), 4173-4181. <https://doi.org/10.1021/pr9004794>

Huang, T., Wang, J., Yu, W., & He, Z. (2012). Protein inference: a review. *Briefings in Bioinformatics*, *13*(5), 586-614. <https://doi.org/10.1093/bib/bbs004>

Li, Y., Feng, A., Zheng, S., Chen, C., & Lyu, J. (2022). Recent Estimates and Predictions of 5-Year Survival in Patients with Gastric Cancer: A Model-Based Period Analysis. *Cancer Control*, *29*, 10732748221099227. <https://doi.org/10.1177/10732748221099227>

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>

Matsuoka, T., & Yashiro, M. (2018). Biomarkers of gastric cancer: Current topics and future perspective. *World J Gastroenterol*, *24*(26), 2818-2832. <https://doi.org/10.3748/wjg.v24.i26.2818>

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>

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*, *73*(11), 2092-2123. <https://doi.org/10.1016/j.jprot.2010.08.009>

neXtProt. (2022, 2022-08-18). *Protein existence*. Retrieved 10/01 from <https://www.nextprot.org/about/protein-existence>

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>

Omenn, G. S., Lane, L., Overall, C. M., Paik, Y.-K., Cristea, I. M., Corrales, F. J., Lindskog, C., Weintraub, S., Roehrl, M. H. A., Liu, S., Bandeira, N., Srivastava, S., Chen, Y.-J., Aebersold, R., Moritz, R. L., & Deutsch, E. W. (2021). Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *Journal of Proteome Research*, *20*(12), 5227-5240. <https://doi.org/10.1021/acs.jproteome.1c00590>

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>

Ron Milo, R. P. (2015). *Cell biology by the numbers*. CRC Press.

The, M., MacCoss, M. J., Noble, W. S., & Käll, L. (2016). Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J Am Soc Mass Spectrom*, *27*(11), 1719-1727. <https://doi.org/10.1007/s13361-016-1460-7>

Veenstra, T. D., Conrads, T. P., & Issaq, H. J. (2004). What to do with“one-hit wonders”? *Electrophoresis*, *25*(9), 1278-1279. <https://doi.org/10.1002/elps.200490007>

Wu, G., Wan, X., & Xu, B. (2018). A new estimation of protein-level false discovery rate. *BMC Genomics*, *19*(S6). <https://doi.org/10.1186/s12864-018-4923-3>

Xu, A., Li, G., Yang, D., Wu, S., Ouyang, H., Xu, P., & He, F. (2015). Evolutionary Characteristics of Missing Proteins: Insights into the Evolution of Human Chromosomes Related to Missing-Protein-Encoding Genes. *Journal of Proteome Research*, *14*(12), 4985-4994. <https://doi.org/10.1021/acs.jproteome.5b00450>

Zhou, L., Wong, L., & Goh, W. W. B. (2018). Understanding missing proteins: a functional perspective. *Drug Discovery Today*, *23*(3), 644-651. <https://doi.org/https://doi.org/10.1016/j.drudis.2017.11.011>

Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., Zhang, M. J., Rao, V., Avina, M., Mishra, T., Johnson, J., Lee-McMullen, B., Chen, S., Metwally, A. A., Tran, T. D. B., Nguyen, H., Zhou, X., Albright, B., Hong, B.-Y., Petersen, L., Bautista, E., Hanson, B., Chen, L., Spakowicz, D., Bahmani, A., Salins, D., Leopold, B., Ashland, M., Dagan-Rosenfeld, O., Rego, S., Limcaoco, P., Colbert, E., Allister, C., Perelman, D., Craig, C., Wei, E., Chaib, H., Hornburg, D., Dunn, J., Liang, L., Rose, S. M. S.-F., Kukurba, K., Piening, B., Rost, H., Tse, D., McLaughlin, T., Sodergren, E., Weinstock, G. M., & Snyder, M. (2019). Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*, *569*(7758), 663-671. <https://doi.org/10.1038/s41586-019-1236-x>

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>

1. 1.The two-SpC-rule used is different from the “two-peptide-rule”, which requires two unique peptides for each identified protein. Since this study is spectral-count-based and the data is already FDR-adjusted, the two-SpC-rule is easy to be implemented and should behave similar to the “two-peptide-rule”. [↑](#footnote-ref-1)