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

Chunhui Gu[1, 2], Ehsan Irajizad [1, 2], Yining Cai[2], Fu-Chung Hsiao[2], Jennifer Dennison[2], Jody Vykoukal[2], Hiro Katayama[2], Johannes Farhmann[2], Kim-Anh Do[1], Samir Hanash[2]*
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.

Background

"Missing proteins (MPs)" are proteins lacking sufficient supporting evidence from mass spectrometry or other direct protein methods (Baker et al., 2017). The lack of reliable proteomics techniques and high expression variation between different cells or tissues may be the reason why those MPs are considered missing. 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 non-missing (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] (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.,

[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".

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 k, is calculated by:

$$(NASF)_k = \frac{(SpC/L)_k}{\sum_{i=1}^{N}(SpC/L)_i} \times 10^6$$

The length-adjusted spectral count $(SpC/L)_k$ 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 10^6 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 (+) are 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 are 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. Pearson 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 with TPM > 0 was considered expressed, and a protein product with NASF > 0 was considered expressed. Obviously, those are 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.

~~Missing protein distribution among other cancer types (?)~~

~~We explored the distribution of MP for additional four common cancer, namely small cell lung cancer (SCLC), lung adenocarcinoma, pancreatic cancer, and breast cancer. A pre-filtering was applied to empirically control the false discovery rate (FDR) at the group level 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 numbers 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 half of the samples (Yunshun Chen, 2022).~~

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”, 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 for this part were similar when using data with the “two-SpC-rule” (Table S2). However, the

Correlation of protein and RNA

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 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 lifetime of RNA and protein are different also causes the non-perfect correlation between protein and RNA

There are more single-count proteins in the missing protein. The less report of missing proteins. Should use more than simply abandon one-hit proteins. Require special consideration about the number of peptides matched when calculating FDR. And there is yet common belief about how such information should be used in calculating FDR. As the development, some missing proteins could be as regular proteins. The others may be disease-specific proteins that could be used as biomarkers.

Ask for new model for detecting of missing protein (Wu et al., 2018).

There is no difference between the RNA-protein joint distribution using count-filtered data and without count-filtered data. (another support for missing protein)

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.

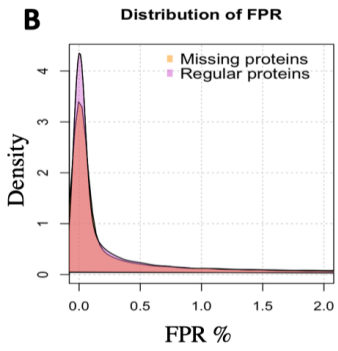


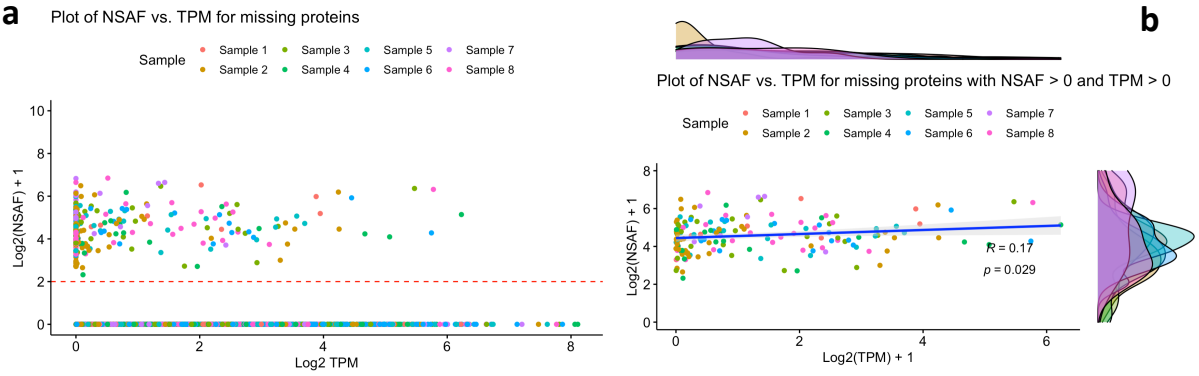
Figure 1 (need a better figure): False discovery rate (FDR) for missing proteins and regular proteins.

There is only a minor difference in FDR distribution between 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.**



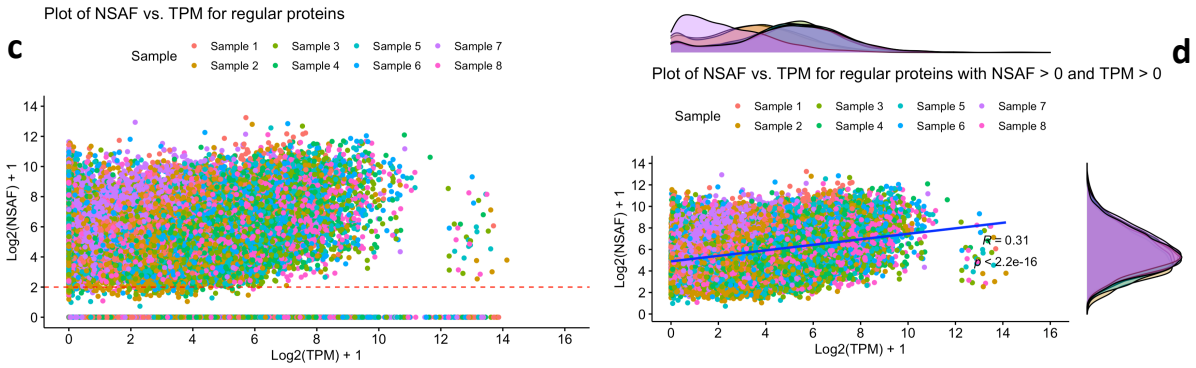


Figure 2. The scatter plot of protein-RNA-product matched pairs. (a): All protein-RNA pairs for missing proteins. (b): protein-RNA pairs with TPM > 0 and NSAF > 0. (c): All protein-RNA pairs for regular proteins. (d): All protein-RNA pairs for regular proteins with TPM > 0 and NSAF > 0.

Supplementary figures:

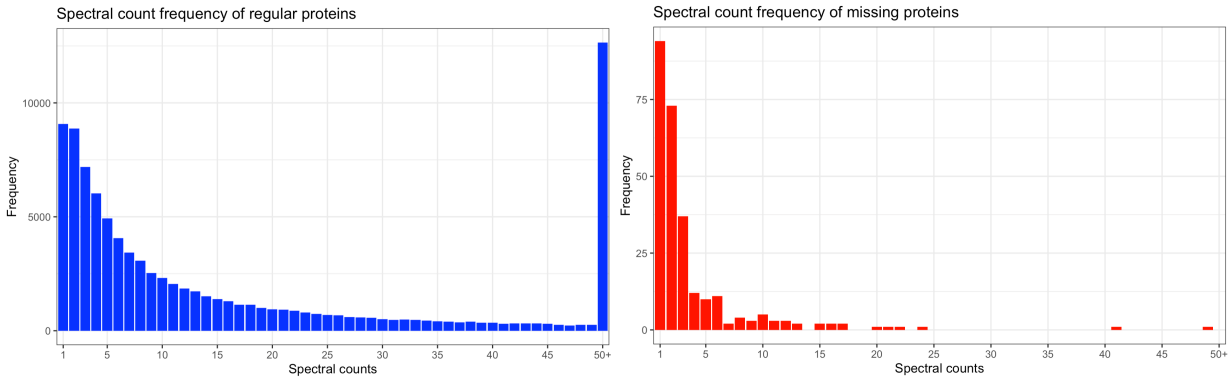


Figure S1: The peptide-spectral-matching (PSM) count distribution for regular proteins (Left blue) and missing proteins (Right red)

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.**

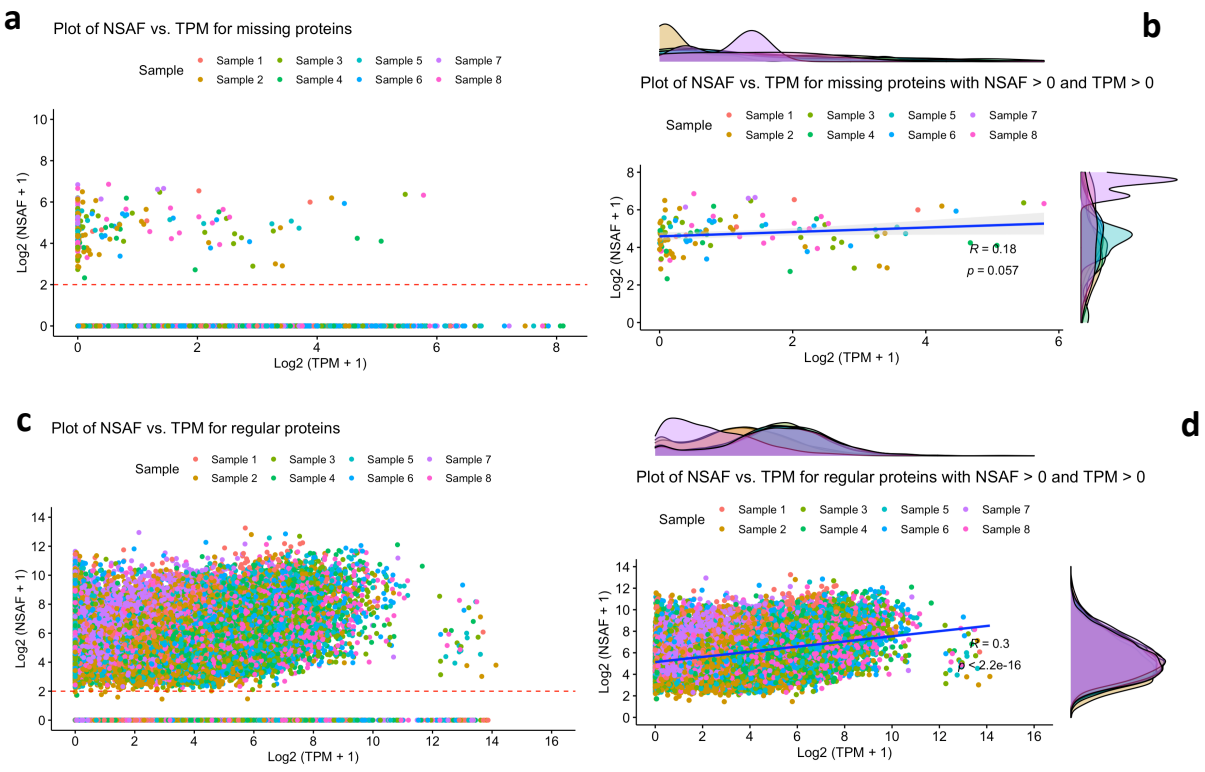


Figure S2. The scatter plot of protein-RNA-product matched pairs. Only protein products with spectral count >= 2 were kept. (a): All protein-RNA pairs for missing proteins. (b): protein-RNA pairs with TPM > 0 and NSAF > 0. (c): All protein-RNA pairs for regular proteins. (d): All protein-RNA pairs for regular proteins with TPM > 0 and NSAF > 0.

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.

Supplementary statistical methods

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

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

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

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

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

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