**Using Protein Localization Studies to Assess the GeTPRA Framework**

Name: Yifan Zhao

Student ID: 260761037

Supervisor: Dr. Uri David Akavia

Department of Biochemistry

McGill University

**Abstract**

Towards the improvement of human genome-scale metabolic models (GEMs), it is important to consider the biological roles of protein isoforms in the context of human metabolism. A systemic framework of gene-transcript-protein-reaction associations (GeTPRA) was developed to facilitate model integration with transcript-level compatible data characterizing the functional roles of protein isoforms1. A potential problem associated with the generation of the GeTPRA framework is its employment of a prediction algorithm to obtain subcellular protein localization data. This study aims to assess the reliability of the subcellular localization data to provide suggestions for future updates of the GeTPRA framework and for the development and reconstruction of more robust human GEMs. Our analysis of protein localization information of a total of 17275 peptide sequences from two tandem mass spectrometry-based experimental datasets demonstrates that there are approximately equal proportions of GeTPRA protein localization predictions supported or not supported by experimental evidence. This result implies that more caution needs to be taken with the use of prediction algorithms alone when integrating protein localization data in the existing metabolic models. It is therefore necessary to review a substantial number of protein localization studies and to employ a hierarchical decision-making model based on reliability scores of such studies for the future improvements of the GeTPRA frameworks as well as human GEMs.

**Introduction**

Metabolism is a complex network of biochemical reactions that can be quantified using a constraint-based approach2. Many transcriptome, proteome, and metabolome datasets have been generated in an effort to understand the mechanistic details of metabolic reactions. As computational platforms to accommodate these data, genome-scale metabolic models (GEMs) together with constraint-based flux balance analysis have been developed as tools for providing a holistic view and snapshots of the global metabolism of organisms under particular genetic and environmental conditions3.

GEMs are computational models, representations of enzymatic reaction networks, describing the entire metabolic reactions and pathways on a genome level of a cell, tissue, organ, or organism4. The human GEMs more specifically represent a curated knowledge-base and have helped uncover the molecular basis of metabolism. Several generic human GEMs, including EHMN, HMR series, and Recon series, have been reconstructed3. As an extension to the Recon series, information on protein isoform has also been incorporated to reconstruct Recon 2M.2 by a Korean research team from KAIST in 20171. Ryu, Kim et al. was the first to develop a systematic framework that addresses the biological roles of protein isoforms in human metabolic network. They first refined Recon2Q by model integration with transcript-level data, resulting in an updated human GEM called Recon 2M.1 that is transcript-level data-compatible. Using the newly refined Recon 2M.1, they generated gene-transcript-protein-reaction associations (GeTPRAs) which were deployed to further upgrade Recon 2M.1 to Recon 2M.2. It was demonstrated in their study that personal GEMs with GeTRPA information enabled more accurate simulation of cancer metabolism and prediction of anticancer targets.

The current GeTPRA framework relies on a prediction algorithm, Wolf PSort5, to provide subcellular location (SL) information for all peptide sequences in the framework. Wolf PSort is a protein SL prediction program, computing based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs5. To evaluate the reliability of Wolf PSort prediction, Ryu, Kim, *et* *al.* compared the SL prediction with immunohistochemistry evidence available at Human Protein Atlas (HPA)6. Among the 1,106 metabolic genes included in the GeTPRA framework, 498 genes are available at HPA­, and 371 out of the 498 genes have SL prediction consistent with HPA experimental evidence1. While this result is in support of the use of Wolf PSort prediction algorithm, the comparison analysis itself could be problematic due to the following reasons: 1) Antibodies used for immunochemistry experiments in HPA do not necessarily recognize all of the protein isoforms of the same gene, and therefore, subcellular location information of protein isoforms is not well defined in HPA database. HPA provides experimental protein SL data at gene level instead of at transcript level6. 2) HPA database only covers 45% of all of the metabolic genes in the GeTPRA framework although 74.5% of the covered metabolic genes have consistent SL data. Because of these two reasons, the use of HPA database to evaluate the SL data reliability seems to have oversimplified the problem. Towards the development of more thorough and robust GeTPRA and reconstruction of a better human GEM, this work is aimed to address the problems abovementioned to provide suggestions for future updates of the GeTPRA framework.

**Materials and Methods**

*Obtaining and Processing Protein Localization Data*

The experimental data chosen in this work come from two recent studies that cover data obtained from tandem mass spectrometry technique, including a sub-cellular fractionation-based study7 and a biotinylation study aimed at inner mitochondrial membrane (IMM) proteome using an *in situ*-generated radical probe with genetically targeted peroxidase (APEX)8 (accessed Feb 11, 2019).

A Python script, *inputFastaGenerator.py*, was used to fetch and process the experimental data from the two studies using the Pandas package9. Data from Shekari *et al*.’s study (Table S2) was parsed into a Pandas data frame to generate a list containing peptide sequences with corresponding subcellular location information. Repeated entries of peptide sequence-subcellular location were removed, since the relative quantity of each peptide in its subcellular compartment is irrelevant in this work. The resulted list of unique peptide sequence-subcellular location entries was then used to write a FASTA file to make queries using command line BLASTP10, 11. Data from Lee *et al*.’s study was processed in a similar manner, with a FASTA file containing unique peptide sequences as input to command line BLASTP.

*BLASTP Analysis*

Command line BLASTP (Version 2.8.1, build Nov 26, 2018) was run on a MacOS version 10.14.1 machine with the following parameters: evalue 100, outfmt 510, 11. The E-value cutoff was set to 100 because many of the peptide sequences obtained from the experimental studies are relatively short (Table S1). A Python script, *peptide\_seq\_length\_counter,py*, was used to summarize the peptide length information (Table S1). 93.0% of the peptide sequences obtained from Shekari *et al.*’s study has lengths smaller than or equal to 20 peptides, and 791 peptides are 6-peptide in length which is the shortest length found this dataset. 75.7% of the peptide sequences obtained from Lee *et al*.’s study has lengths smaller than or equal to 20 peptides, and 49 peptides in this dataset are 8-peptide in length which is the shortest length found in this dataset.

A bash script, *runblast.sh*, was used to streamline the process of creating a BLASTP database using the Ensembl database of all protein sequences in human genome12 (release 95, accessed Jan 26, 2019), inputting the query FASTA files to BLASTP against all peptide sequences in the Ensembl library12, and generating BLASTP output in xml format. A python script, *MSMS\_blast\_parser.py*, was used to process BLASTP output using the Pandas and Biopython packages9, 13. The parsed output is converted into a csv file containing the following hit information for each query peptide sequence: Ensembl Gene (ENSG) ID, Ensembl Transcript (ENST) ID, Gene Name, Match Sequence, Subject Sequence, Alignment Length, Identities, E Value. A *entry\_counter.py* script was written to look at the number of ENSTs and ENSGs that were mapped to three subcellular locations based on tandem mass spectrometry evidence (Table 1). The subcellular locations being mapped to are mitochondria, nucleus and cytoplasm. Light and heavy microsome experimental data were excluded due to ambiguity and lack of consensus with respect to how these locations may correspond to the subcellular locations included in the GeTPRA framework (ER membrane, peroxisome, lysosome, Golgi apparatus), in addition to the three locations that are included. A detailed discussion on whether and how these organelles could be categorized as heavy or light microsomes is beyond the scope of this study. Furthermore, each entry in the csv file for query peptide sequences is also annotated with the subcellular location information obtained from the two experimental studies aforementioned7, 8 (Table S2, Table S3).

*Comparison of GeTPRA and Experimental Protein Localization Data*

A Python script, *compare\_with\_GeTPRA.*py, was used to parse the GeTPRA framework1 using the Pandas package9 into a data frame with the following information extracted: Entrez gene ID, ENSG, ENST, Transcript Type, Predicted Subcellular Location, Experimental Evidence on SLs. Subsequently, the parsed BLASTP output csv files are individually processed. For each ENST ID in the GeTPRA framework, its existence in the BLASTP output files was checked and experimental SL information was appended to corresponding entry if ENST ID was present in the BLASTP output files. The output is a summary data table of comparison between the GeTPRA framework and experimental protein localization information in csv format (Table S4, Table S5).

In order to assess the coverage of GeTPRA entries with subcellular location information supported by the tandem mass spectrometry evidence from the primary literature7, 8, a *check\_MS\_evidence.py* script was used to search for all entries in the GeTPRA framework1 that has SLs supported by experimental evidence (Table S6), and all of the unique entries, in terms of ENST, that are supported (Table S7), regardless of whether the ENSTs of the entries are covered in the two experimental studies. For all of the entries with ENSTs covered at least one of the two experimental studies, the *check\_MS\_evidence.py* script was used to collect all the entries that include mitochondria/nucleus/cytoplasm for subcellular location but are not supported by experimental evidence from either of the two studies (Table S8). The unique entries, in terms of ENSTs, that are not supported was also found using the same script (Table S9).

Lastly, *coverage\_histogram\_generator.py* script was used to generate histograms of transcript coverage per gene in the GeTPRA framework by the tandem mass spectrometry evidence, and of number of transcripts per gene in the GeTPRA framework (Figure 1, Figure 2).

**Results**

In this work, a total number of 17275 peptide sequences were obtained from the two tandem mass spectrometry experimental studies, with 686 sequences from the APEX mediated proximity biotinylation study8 and 16589 sequences from the sub-cellular fractionation-based study7. The peptide sequences were sent as input through command-line BLASTP to make queries about which gene and transcript ID these sequences may correspond to, with E-value cutoff set to 100. The threshold for E-value is higher than that in a typical BLASTP analysis due to the fact that many of the input peptide sequences are very short in length, as summarized in the Materials and Methods section and Table S1. BLASTP analysis mapped the input peptide sequences to 14750 transcript IDs which correspond to 3449 gene IDs (Table 1). There were 1086 transcript IDs overlapped between the two datasets, corresponding to 328 overlapped gene IDs (Table 1). These transcript IDs were then annotated with subcellular locations based on tandem mass spectrometry evidence7, 8 (Table S2, Table S3).

The gene IDs, transcript IDs and subcellular location predictions in the GeTPRA framework were subsequently analyzed and compared to their localization data from the primary literature7, 8. In order to assess the inclusivity of the GeTPRA framework and the coverage on the GeTPRA entries by the two datasets, two histograms were generated, looking at transcript coverage per gene in the framework, as well as the number of transcripts per gene that are included in the framework (Figure 1, Figure 2). Overall, the two datasets covered 4137 out of 11668 GeTPRA entries (35.5%) (Table 2, Table S5). It should be noted that in the GeTPRA framework, a transcript ID can have multiple entries if it is predicted in more than one subcellular location. Out of the 3445 unique transcripts included in GeTPRA, 1234 (35.7%) of them were covered by the two datasets (Table 2, Table S4).

The subcellular locations of all transcripts included in GeTPRA were compared to their localization data from the two tandem mass spectrometry studies aforementioned7, 8. Due to the aforementioned limitations in the experimental studies, only location predictions that include mitochondria/nucleus/cytoplasm (m/n/c) were analyzed in this work. For all the GeTPRA entries that have transcripts covered by experimental data, 36.0% have location predictions supported by experimental evidence (Table 2, Table S6). For all the transcript IDs that are covered by the experimental datasets, 75.8% have supported location predictions. Another question of interest is, for the GeTPRA entries or unique transcripts that are covered by experimental data, whether there is any m/n/c location prediction in the GeTPRA framework that are not found in the experimental datasets. For convenience purposes in this discussion, a subcellular location prediction is considered “not supported” if a GeTPRA entry includes m/n/c but such information is not found in the two experimental datasets. Surprisingly, 39.7% of all covered GeTPRA entries and 72.0% of all covered GeTPRA unique transcripts are not supported by experimental data, showing that there are approximately equal proportions of subcellular location prediction supported and not supported by experimental evidence (Table 2, Table S8, Table S9).

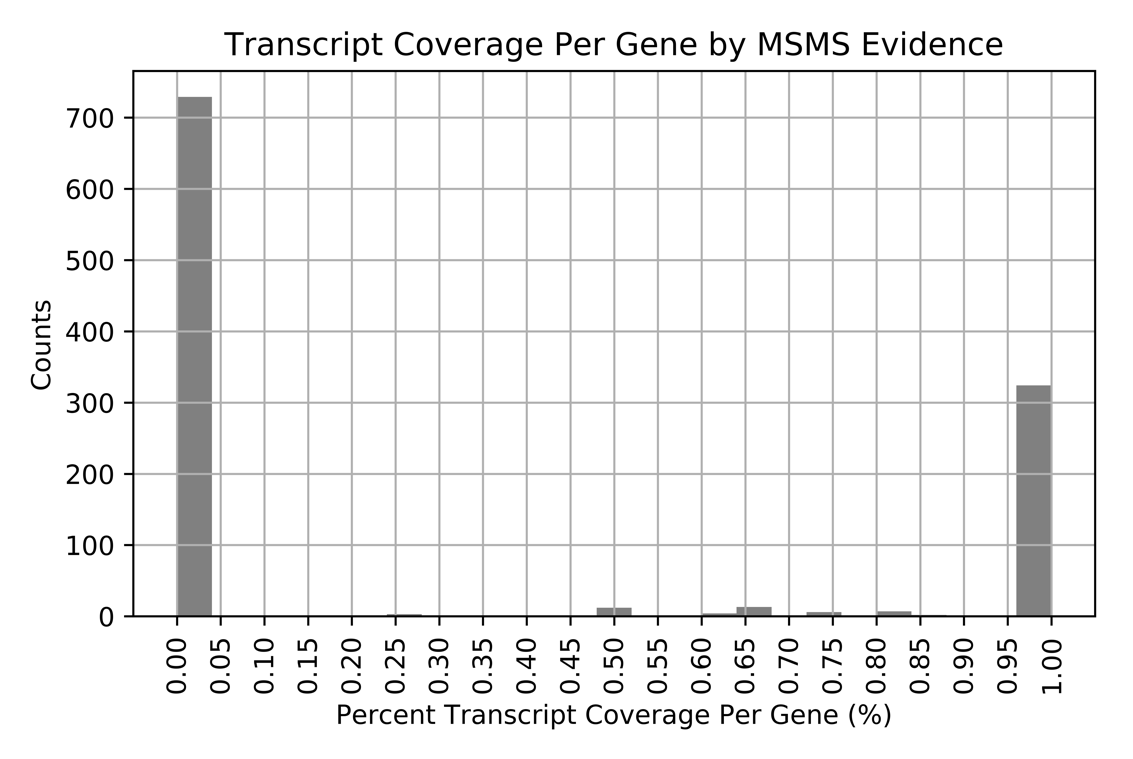
**Discussion**

The cellular proteome is compartmentalized and spatiotemporally regulated to a high degree14. Integrating protein spatial distribution into existing human GEMs can significantly increase the accuracy of model predictions and is useful for detection of false-positive predictions of protein interactions14. As reviewed by Helen Tung in her BIOC 462 Literature Review, inaccurate annotations of protein localization on gene-protein reaction associations may lead to inaccurate predictions in gene knockout model generations, resulting in false positives in *in silico* cell death screens and inaccurate quantitative phenotypes in gene deletion or knockdown simulations. While there is currently no definitive source of information for the subcellular localization for most proteins15, fractionation, biotinylation and immunofluorescence are the three main method commonly employed to identify subcellular localization. A detailed review for these three methods can be found in Tung’s BIOC 462 Literature Review: *Using Protein Localization Studies to Improve Genome-scale Metabolic Models*. Previous work by Nisha Kabir has compared GeTPRA protein localization prediction with protein isoform subcellular location data available in Human Protein Atlas which uses immunofluorescence evidence to identify endogenous protein distribution across cellular compartments6. Because HPA localization data is not specific to protein isoforms at transcript level and the exact epitope binding sequences was unknown, the entire antigenic sequences were used to as input to query through BLASTP to obtain transcript IDs of protein isoforms that may bind to any particular antibody, as described in Kabir’s BIOC 396 Lab Report. ,

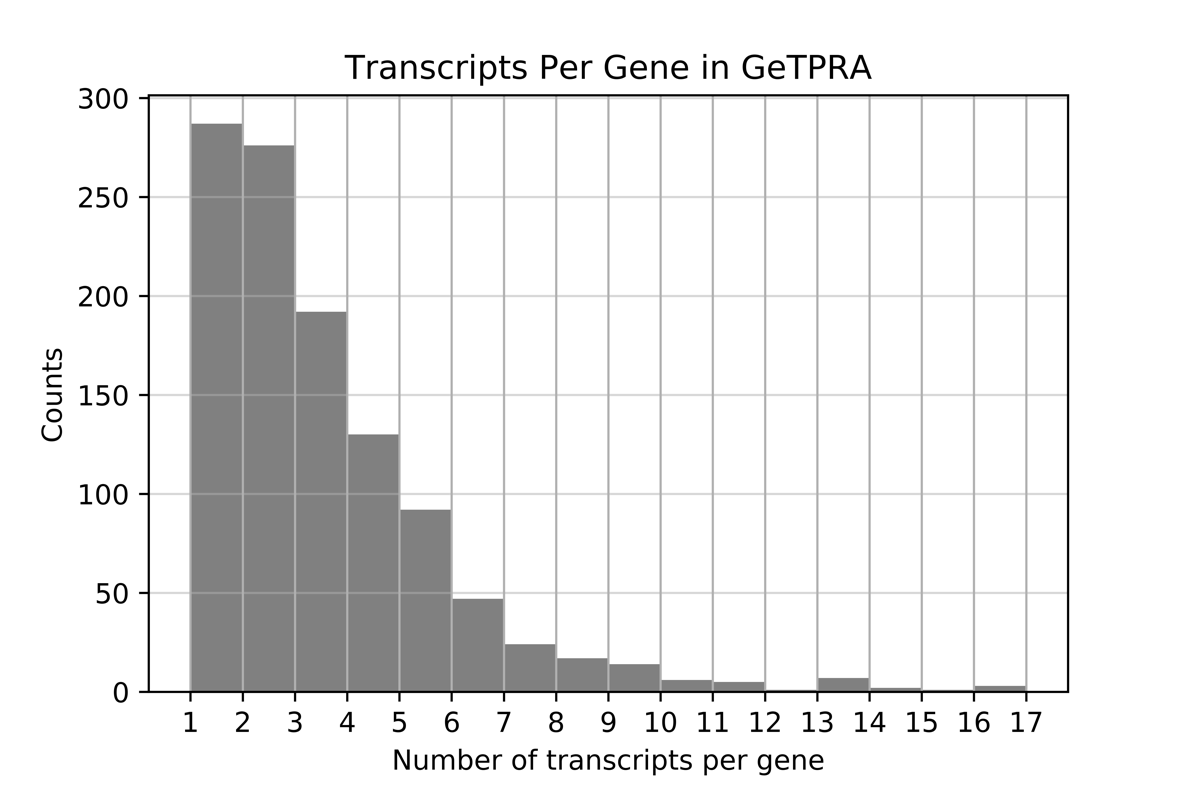
In this work, protein localization information from two datasets were analyzed, encompassing a sub-cellular fractionation-based method7 and a proximity biotinylation method using an *in situ*-generated radical probe with genetically targeted peroxidase (APEX)8. Both the fractionation-based method and the APEX mediated biotinylation method were followed by tandem mass spectrometry to identify the peptides present in different subcellular compartments. The results from this work suggest that the reliability of protein localization predictions in the current GeTPRA framework is of question. While the usage of predictive algorithms such as Wolf PSort is convenient and enables automation for the generation or update of existing human GEMs, it is problematic to rely solely on predictive algorithms when integrating protein localization information in the existing metabolic models.

*Limitations and Further Directions*

There are several limitations in this study. First, only two experimental studies were analyzed and compared to due to time restriction. In order to achieve greater accuracy when updating the existing metabolic frameworks, a substantial number of protein localization studies need to be reviewed and incorporated. Second, as described in the Materials and Methods section, the expectation value threshold was set to a relatively high value during BLASTP analysis due to the short lengths of peptide sequences, which may lead to false positive results. Third, it needs to be emphasized that the type of cells used in experimental studies should not be ignored. For generic human GEMs such as the Recon series3, proteomics data across multiple cell types need to be incorporated to achieve greater accuracy. Moreover, conflicting protein localization result may arise depending on the type of cell and the experimental protocols used in any particular study. For future updates of the GeTPRA framework and the reconstruction of a more robust metabolic models, a hierarchical decision-making model can be employed, where the inclusion of localization information with experimental evidence should be prioritized based on factors such as the reliabilities of the experimental protocol and the cell type employed, while exercising more caution with the use of predictive algorithms to generate localization prediction.



**Figure 1 | Transcript coverage per gene by MS/MS Evidence**



**Figure 2 | Transcripts per gene in GeTPRA**

|  |  |  |
| --- | --- | --- |
|  | Number of gene IDs | Number of transcript IDs |
| APEX study8 (Lee *et al.,* 2017) | 468 | 1541 |
| Fractionation-based study7 (Shekari *et al*., 2017) | 3309 | 14295 |
| Total Number of unique Gene/Transcript IDs | 3449 | 14750 |
| Number of overlapped Gene/Transcript IDs | 328 | 1086 |

**Table 1 | Summary statistics of gene and transcript IDs that were mapped to three subcellular locations based on tandem spectrometry evidence7, 8**

The subcellular locations being mapped to are mitochondria, nucleus and cytoplasm. Crude membrane, light and heavy microsome experimental data were not included in this summary.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SL prediction supported | SL prediction not supported\* | Total number of GeTPRA entries/ENSTs covered | Total number of GeTPRA entries/ENSTs |
| Number of GeTPRA Entries | 1489 | 1642 | 4137 | 11668 |
| Number of ENSTs in GeTPRA | 935 | 888 | 1234 | 14750 |

**Table 2 | Validation of GeTPRA location prediction against two experimental evidence7, 8 with respect to GeTPRA entries and transcript IDs and their respective coverage by the experimental datasets**

The detailed lists of GeTPRA entries and unique transcript IDs that are either supported or not supported by experimental evidence can be found in Supplementary Tables 6-9. \*Due to limitations in the experimental studies aforementioned, in this study, only location predictions that include mitochondria/nucleus/cytoplasm (m/n/c) were analyzed. A subcellular location prediction is considered “not supported” if a GeTPRA entry includes m/n/c but such information is not found in the two experimental datasets.

**References:**

1. Ryu, J. Y.; Kim, H. U.; Lee, S. Y., Framework and resource for more than 11,000 gene-transcript-protein-reaction associations in human metabolism. **2017,** *114* (45), E9740-E9749.

2. Lewis, N. E.; Nagarajan, H.; Palsson, B. O., Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology* **2012,** *10*, 291.

3. Cho, J. S.; Gu, C.; Han, T.; Ryu, J. Y.; Lee, S. Y., Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring. *Current Opinion in Systems Biology* **2019**.

4. O'Brien, E. J.; Monk, J. M.; Palsson, B. O., Using Genome-scale Models to Predict Biological Capabilities. *Cell* **2015,** *161* (5), 971-987.

5. Adams-Collier, C. J.; Harada, H.; Park, K.-J.; Fujita, N.; Horton, P.; Obayashi, T.; Nakai, K., WoLF PSORT: protein localization predictor. *Nucleic Acids Research* **2007,** *35* (suppl\_2), W585-W587.

6. Thul, P. J.; Lindskog, C., The human protein atlas: A spatial map of the human proteome. *Protein science : a publication of the Protein Society* **2018,** *27* (1), 233-244.

7. Shekari, F.; Nezari, H.; Larijani, M. R.; Han, C.-L.; Baharvand, H.; Chen, Y.-J.; Salekdeh, G. H., Proteome analysis of human embryonic stem cells organelles. *Journal of Proteomics* **2017,** *162*, 108-118.

8. Lee, S.-Y.; Kang, M.-G.; Shin, S.; Kwak, C.; Kwon, T.; Seo, J. K.; Kim, J.-S.; Rhee, H.-W., Architecture Mapping of the Inner Mitochondrial Membrane Proteome by Chemical Tools in Live Cells. *Journal of the American Chemical Society* **2017,** *139* (10), 3651-3662.

9. McKinney, W., *Data Structures for Statistical Computing in Python*. 2010.

10. Information, N. C. f. B.; Camacho, C., *BLAST (r) Command Line Applications User Manual*. National Center for Biotechnology Information (US): 2008.

11. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *Journal of molecular biology* **1990,** *215* (3), 403-10.

12. Pietrelli, A.; Lopez, G.; Ezkurdia, I.; Wesselink, J.-J.; Rodriguez, J. M.; Maietta, P.; Valencia, A.; Tress, M. L., APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research* **2012,** *41* (D1), D110-D117.

13. Dalke, A.; Wilczynski, B.; Chapman, B. A.; Cox, C. J.; Kauff, F.; Friedberg, I.; Chang, J. T.; de Hoon, M. J. L.; Cock, P. J. A.; Hamelryck, T.; Antao, T., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009,** *25* (11), 1422-1423.

14. Thul, P. J.; Akesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Ait Blal, H.; Alm, T.; Asplund, A.; Bjork, L.; Breckels, L. M.; Backstrom, A.; Danielsson, F.; Fagerberg, L.; Fall, J.; Gatto, L.; Gnann, C.; Hober, S.; Hjelmare, M.; Johansson, F.; Lee, S.; Lindskog, C.; Mulder, J.; Mulvey, C. M.; Nilsson, P.; Oksvold, P.; Rockberg, J.; Schutten, R.; Schwenk, J. M.; Sivertsson, A.; Sjostedt, E.; Skogs, M.; Stadler, C.; Sullivan, D. P.; Tegel, H.; Winsnes, C.; Zhang, C.; Zwahlen, M.; Mardinoglu, A.; Ponten, F.; von Feilitzen, K.; Lilley, K. S.; Uhlen, M.; Lundberg, E., A subcellular map of the human proteome. *Science (New York, N.Y.)* **2017,** *356* (6340).

15. Lund-Johansen, F.; de la Rosa Carrillo, D.; Mehta, A.; Sikorski, K.; Inngjerdingen, M.; Kalina, T.; Roysland, K.; de Souza, G. A.; Bradbury, A. R.; Lecrevisse, Q.; Stuchly, J., MetaMass, a tool for meta-analysis of subcellular proteomics data. *Nature methods* **2016,** *13* (10), 837-40.