**CPTAC GENCODE V34 harmonized proteomics README**

Yongchao Dou, Felipe Leprevost, Zhiao Shi, Sara Savage, Alexey Nesvizhskii, and Bing Zhang

Matched protein database, gencode.v34.pc\_translations.fa, was download from GENCODE. Only proteins from coding transcripts in GENCODE V34 Basic (CHR) annotation were retained and others were discarded. MSFragger v3.4(Kong et al., 2017), the Philosopher v4.0.1(da Veiga Leprevost et al., 2020) toolkit, and the TMT-Integrator(Djomehri et al., 2020) pipeline were used by the Michigan University team in the CPTAC pan-cancer working group to process and quantify the mass spectrometry data using above protein database. Gene and phosphosite intensities reported by the Michigan pipeline from the analysis of global and phosphoproteomics data were normalized across cancer types by median centering of the medians of reference intensities of each cancer type.

Single phosphosites were re-annotated to the selected primary and secondary protein isoforms. First, if the original selected isoform protein sequence matched the primary selected isoform sequence, only the protein isoform ID was changed. If the protein sequences did not match, the primary selected sequence was searched for all peptides identified for the phosphorylation site. If at least one peptide matched exactly once to the sequence, that peptide was used to update the site position. Otherwise, for peptides that matched more than one location, the one that matched the fewest locations was selected and the first matching position was used to update the site position. Finally, if no peptides exactly matched the selected protein sequence, all I’s were changed to L’s in both the sequence and peptides and the matching step was performed again. All sites with no peptides that could be mapped to the selected protein sequence were discarded after this step. Because some re-annotated site IDs were no longer unique, the data row with the fewest missing values was selected for that site and all others discarded. The site ID finally consisted of the Ensembl gene ID, Ensembl protein ID, site position based on the selected protein ID, fifteenmer (+/- 7 amino acids) based on the selected protein ID, and a flag for whether the protein is a primary (1) or secondary (2) selected sequence.

Related file(s):

XXXXXX\_phospho\_site\_abundance\_log2\_reference\_intensity\_normalized\_Normal.txt

XXXXXX\_phospho\_site\_abundance\_log2\_reference\_intensity\_normalized\_Tumor.txt

XXXXXX\_proteomics\_gene\_abundance\_log2\_reference\_intensity\_normalized\_Normal.txt

XXXXXX\_proteomics\_gene\_abundance\_log2\_reference\_intensity\_normalized\_Tumor.txt

Contact:

If you have questions about this pipeline, please contact Yongchao Dou at BCM: [ydou@bcm.edu](mailto:ydou@bcm.edu)