Supporting Information

Large-scale reanalysis of publicly available HeLa cell proteomics data in the context of the Human Proteome Project

Thibault Robin[1,2,4,5], Amos Bairoch[1,5], Markus Müller[3], Frédérique Lisacek[2,4,6] and Lydie Lane[1,5*]

1 CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, CH-1211 Geneva, Switzerland
2 Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, CH-1211 Geneva, Switzerland
3 Vital-IT Group, SIB Swiss Institute of Bioinformatics, Genopode building, Quartier Sorge, CH-1015 Lausanne, Switzerland
4 Computer Science Department, University of Geneva, Switzerland
5 Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, Switzerland
6 Section of Biology, University of Geneva, Switzerland

* Corresponding author. E-mail: Lydie.Lane@sib.swiss, Tel: +41 (0) 22 379 58 41

Supplementary File 1 (HeLa_Database.txt): HeLa customized variant protein database in the FASTA format that was used as input for the database search step.

Supplementary File 2 (HeLa_PSMs.txt): Tab-separated table describing the 1,225,780 identified PSMs. Each row represents a distinct PSM, providing information about both the peptide (protein, gene, variable modifications and variant/wild-type property) and the corresponding mass spectrum (experiment, file, index, scan, charge and universal spectrum identifier).

Supplementary File 3 (HeLa_Identifications.xlsx): Excel table summarizing the peptide, protein, phosphorylation site and N-terminal acetylation site identifications. For each of the four categories, the corresponding identifications were given a reference mass spectrum (referred by the universal spectrum identifier) based on the best score. For the protein identifications, the two best non-nested peptides with a length greater or equal to nine amino acids were additionally reported. For the peptide identifications, distinct rows were included to account for the presence of the PTMs detected by X!Tandem. For the phosphorylation site and N-terminal acetylation site identifications, the “Validated Sites” column describes the sites that were detected (and validated by PTMProphet for the phosphorylations) while the “New Sites” column describes the validated sites that were not reported in neXtProt (release 2018.01.17). A reference protein isoform is also specified to include fine details of site mapping.