Expression Atlas update – an integrated database of gene and protein expression in humans, animals and plants

Robert Petryszak[*,1], Maria Keays[*], Y. Amy Tang[*], Nuno A. Fonseca[*], Elisabet Barrera[*], Tony Burdett[*], Anja Füllgrabe[*], Alfonso Muñoz-Pomer Fuentes[*], Simon Jupp[*], Satu Koskinen[*], Oliver Mannion[*], Laura Huerta[*], Karine Megy[*], Catherine Snow[*], Eleanor Williams[*], Mitra Barzine[*], Emma Hastings[*], Hendrik Weisser[**], James Wright[**], Pankaj Jaiswal[***], Wolfgang Huber[*], Jyoti Choudhary[**], Helen E. Parkinson[*] and Alvis Brazma[*]

[*] European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK
[**] Welcome Trust Sanger Institute, Hinxton, UK
[***] Oregon State University, Corvallis, US

MATERIAL AND METHODS

Human Proteome Label Free Analysis

Mass spectrometry raw data from the draft map of the human proteome (1) was downloaded from the PRIDE (2) repository (PXD000561), comprising 85 experimental samples from 30 human adult and fetal tissues. Raw data was converted to mzML format and was processed in an OpenMS workflow (3). Spectra were searched using Mascot (4, v2.5) and MS-GF+ (5, v10089) against a combined FASTA database containing: GENCODE (6, v20) coding sequences, UniProt (7) human proteome sequences (May 2014), and a set of contaminant sequences. This database was concatenated with an equal number of shuffled decoy sequences. Results from both search engines were processed using MascotPercolator (8, 9, v2.13) and Percolator (10, v2.08) respectively. All database searches were performed with a precursor tolerance of 10 ppm and a fragment tolerance of 0.02 Da. Up to three missed cleavages were allowed. The fixed modification carbamidomethyl (+57.0214) was specified for all cysteine residues. In addition, the following variable modifications were used in the searches: N-terminal acetylation (+42.01056), N-terminal carbamidomethyl (+57.0214), deamidation of asparagine and glutamine residues (+0.984), oxidation of methionine (+15.9949), and N-terminal conversion of glutamine and glutamic acid to pyro-glutamine (-17.0265, -18.0106). Percolated results were parsed, merged and filtered so that every peptide spectrum match (PSM) had the same identification in both search engines. The worst posterior error probability (PEP) was retained in each case. The PSMs were then filtered to a 1% false discovery rate (FDR), a maximum PEP of 0.05, and a minimum peptide length of 7 amino acids. PSMs matching contaminant or decoy sequences were also removed. GENCODE CDS and UniProt accessions were mapped to Ensembl gene identifiers, and these genes were clustered to remove entries only matching a subset of peptides from another gene. Label free quantification for each mapped Ensembl gene was conducted in each individual tissue. This was calculated as the summed precursor intensities of the most intense three unique peptides in each gene cluster, these were then converted into within sample abundances by dividing by the total summed quantification of all proteins in each sample. These abundance values were then normalised using the ten genes displaying the lowest co-efficient of variation across all tissues, with the median value being taken across replicates.

[1] To whom correspondence should be addressed. Tel: +44 (0)1223 492 696 Fax: +44 (0)1223 494 468 Email: rpetry@ebi.ac.uk Present Address: Robert Petryszak, Functional Genomics, European Bioinformatics Institute EMBL, Hinxton, Cambridge, CB10 1SD. UK

REFERENCES

1. Kim MS, et al. (2014) A draft map of the human proteome. *Nature [2014, 509(7502):575-581]*, DOI: 10.1038/nature13302

2. Vizcaíno JA, et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research* 2013 Jan;41(Database issue):D1063-1069.

3. Sturm M, et al. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics. 2008 Mar 26;9:163*, DOI: 10.1186/1471-2105-9-163.

4. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis. 1999 Dec;20(18):3551-67*, PMID: 10612281

5. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJ, Pevzner PA (2010). The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol Cell Proteomics. 2010 Dec;9(12):2840-52*, DOI: 10.1074/mcp.M110.003731.

6. Harrow J, et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res. 2012. 22: 1760-1774*, DOI: 10.1101/gr.135350.111

7. The Uniprot Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Research (28 January 2015) 43 (D1): D204-D212*, DOI: 10.1093/nar/gku989

8. Brosch M, Yu L, Hubbard T and Choudhary J (2009). Accurate and sensitive peptide identification with Mascot Percolator. *Journal of proteome research 2009;8;6;3176-81*, DOI: 10.1021/pr800982s

9. Wright JC, Collins MO, Yu L, Käll L, Brosch M and Choudhary JS (2012). Enhanced peptide identification by electron transfer dissociation using an improved Mascot Percolator. *Molecular & cellular proteomics : MCP 2012;11;8;478-91*, DOI: 10.1074/mcp.O111.014522

10. Granholm V, et al (2014). Fast and accurate database searches with MS-GF+Percolator. *J Proteome Res. 2014 Feb 7; 13(2): 890–897*, DOI: 10.1021/pr400937n

11. Fiona Cunningham et al. (2015). Ensembl 2015. *Nucleic Acids Research 2015 43 Database issue:D662-D669*, DOI: 10.1093/nar/gku1010