1. Raw Data Source

The 2212 raw files generated in the Pandey lab, were downloaded from ProteomeXchange via PRIDE ftp. These files make up 85 fractionated experimental samples from 30 different human tissues.

http://www.ebi.ac.uk/pride/archive/projects/PXD000561/files

2. Spectral Processing, Conversion and Merging

Each Raw file was converted to standard mzML format using the ProteoWizard converter. Following conversion, all spectra were centroided using the OpenMS centroiding tool. Finally, the multiple fractions for each experiment were merged into 85 mzML files.

http://proteowizard.sourceforge.net/

http://open-ms.sourceforge.net/

3. Sequence Database Generation

The FASTA sequence database contained four parts. The complete human GRCh38 CDS translated sequences (93246), the UniProt human reference proteome from May2014 (88708), a selection of common contaminate sequences (247) and a set of decoy randomised sequences (182201) generated using the MIMIC tool. To account for isobaric peptides all isoleucine (I) and leucine (L) residues were replaced with a common amino acid J.

http://www.ensembl.org/Homo_sapiens/Info/Index

http://www.uniprot.org/proteomes/UP000005640

https://github.com/percolator/mimic

4. Sequence Database Searching

All 85 experiments were searched against the sequence database with Mascot v2.4 and MSGFplus v10036 search engines, from within an OpenMS workflow, using the parameters:

- Precursor tolerance = *10ppm*
- Fragment tolerance = *0.02Da*
- Allowed missed cleavages = *3*
- Enzyme = *Trypsin*
- Mass type = *Monoisotopic*
- Instrument / Fragmentation = *ESI-Trap / HCD*
- Fixed Modifications = *Carbamidomethyl [C]*
- Variable Modifications = *Acetyl [N-term], Carbamidomethyl [N-term], Deamidation [NQ], Oxidation (M), Pyro-Glu [N-term QE]*

The search results for both search engines were post processed using the Percolator algorithm. The standalone Mascot Percolator v2.08 was used for the Mascot results and msgf2pin and Percolator v2.08-1 were used for the MSGFplus results.

http://www.matrixscience.com/

http://proteomics.ucsd.edu/Software/MSGFPlus/

http://www.sanger.ac.uk/resources/software/mascotpercolator/

https://github.com/percolator

5. Results Processing and Filtering

The results from the two search engines for each experiment were first filtered to remove all peptide spectrum matches (PSMs) with a q-value (FDR) greater than 0.01 (1%) and a posterior error probability (PEP) greater than 0.05. The peptide sequences were then further filtered to have a minimum length of 6 amino acids and to be identified by both search engines. This filtering left no decoy PSMs amongst the results. The resulting peptides were then inferred into protein identifications and these proteins clustered to remove any subset proteins (proteins identified with no unique peptides). The final list of proteins was again filtered with the requirement that every protein had at least 3 identified peptides.

6. Protein Quantification

Protein quantification was conducted using two different methods. The first used was spectral counting, in which all the significant PSMs matching a peptide in a protein were summed, this value was not normalised and although correlates to protein abundance is not an accurate measurement and is biased by protein size. The second is a Top3 intensity based method in which the precursor intensities for the most intense 3 peptides matched to each protein are summed together. To normalise this value into a within sample abundance this value is divided by the total for all proteins in an experiment. Again due to lack of replicates and the data being generated in different way on different instruments these values are not considered accurate quantification.

7. ID Mapping and Tissue Averaging

In this last step the identified and quantified proteins had their accessions mapped to Ensembl gene identifiers. The 85 experiments were then collapsed into 30 individual tissues, taking the median values of the spectral count and Top3 quantifications for multiple experiments.