doi:10.1038/nature13302

Supplementary Discussion

Dependence of FDR on features of identified peptides

The distribution of FDRs across all identified peptides is not uniform and might depend on certain parameters such as length of peptides, number of missed cleavages and charge state of precursor ions. To test this, we collected PSMs identified at 1% FDR from all seven fetal tissues and re-examined FDRs for different groups of peptides based on the above-mentioned parameters. We observed that the FDRs vary across different subsets of peptides. For example, longer peptides (≥16) corresponded to a lower FDR than shorter peptides (Extended Data Figure 1d). Precursor ions with higher charge states (≥4) also corresponded to a lower FDR than those with lower charge states (Extended Data Figure 1e). Finally, the number of missed cleavages in peptides also affected the FDR although this effect was not very pronounced (Extended Data Figure 1f). These data confirm that the rate of false identification indeed depends on a number of features of peptides.

Effect of merging peptide lists on FDRs

The data analysis pipeline used in this study was developed based on analysis of *a set of samples* fractionated by either SDS-PAGE (*e.g.* 24 gel bands) or bRPLC (*e.g.* 24 fractions) as one unit. In our analysis, each unit of raw mass spectrometry files was searched against the protein sequence database resulting in one list of peptide identifications. At the end of all individual searches, the lists of identified peptides were merged to generate the final results. However, the FDR of the final list of identified peptides obtained in this fashion might be different from the FDR of each separate list. We decided to test this by analyzing two sets of samples (adult liver and adult CD8+ T cells where each sample was composed of 24 fractions generating ~400,000

spectra/sample) separately to generate a list of peptides at 1% FDR. When we merged these two lists of identified peptides, we observed that the FDR of the merged set showed a small increase in FDR to 1.1%.

Signal peptides cleavage sites

The exact sites of cleavage of signal peptides in proteins containing signal peptides are not experimentally determined but rather defined computationally based on their composition and predicted physicochemical properties. We have previously shown that this assignment can be erroneous and that semi-tryptic peptides can be used to annotate the correct cleavage sites[1]. We searched a custom database containing peptide sequences surrounding the cleavage sites of signal sequences annotated in Human Protein Reference Database[2] or predicted by SignalP4.0 (http://www.cbs.dtu.dk/services/SignalP/). This confirmed 201 annotated sites in addition to 128 cases where the cleavage site was located either upstream or downstream of the predicted cleavage site indicating that we do not yet know the exact N-termini of many of the mature proteins in the human proteome (Supplementary Table 1).

References

1. Molina, H. et al. A proteomic analysis of human hemodialysis fluid. Mol Cell Proteomics 4, 637-50 (2005).
2. Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13, 2363-71 (2003).