Supplementary Methods

William Stafford Noble*
Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington

Data sets

Plasmodium The *Plasmodium* data set is derived from a recent study of the erythrocytic cycle of the malaria parasite *Plasmodium falciparum*.[1] *P. falciparum* 3D7 parasites were synchronized and harvested in duplicate at three different time points during the erythrocytic cycle: ring (16±4 h postinvasion), trophozoite (26±4 h postinvasion), and schizont (36±4 h postinvasion). Parasites were lysed, and duplicate samples were reduced, alkylated, digested with Lys-C, and then labeled with one of six TMT isobaric labeling reagents. The resulting peptides were mixed together, then fractionated via strong cation exchange into 20 fractions, desalted and then analyzed via LC-MS/MS on an LTQ-Velos-Orbitrap mass spectrometer. All MS/MS spectra were acquired at high resolution in the Orbitrap. I focused on one of these fractions (number 13), consisting of 12,748 spectra.

SP-A The SP-A data set is derived from a recent study of of coding variants and isoforms of pulmonary surfactant protein A (SP-A).[2] SP-A proteins were purified from bronchial lavage fluid using calcium chloride precipitation and differential centrifugation. Proteins were further isolated on a gel and digested with trypsin, and the resulting peptides subjected to one-dimensional liquid chromatography tandem mass spectrometry using a Waters nanoAcquity UPLC and either a Waters G1 HDMS or G2 HDMS. The resulting 1503 spectra were downloaded from `https://discovery.genome.duke.edu/express/resources/2402/SPA_MSMS.sf3` and analyzed jointly.

Honeybee The honeybee data set consists of 63 RAW files generated as part of the Bromenshenk et al. study.[3] The data were collected from homogenized bee samples on an unspecified tandem mass spectrometry platform. The data sets consists of a total of 262,572 spectra.

Protein databases

Plasmodium The *Plasmodium* protein database was downloaded from NCBI Entrez on January 8, 2014, by searching for "Plasmodium falciparum 3D7." The resulting database contains 11,737 proteins and 8,669,430 amino acids. Peptides were digested by Lys-C with no missed cleavages allowed. Three static modifications were included: carbamidomethylation of cysteine and TMT labeling of lysine and N-terminal amino acids. At most one variable modifications per peptide was allowed for methionine oxidation (15.9949 Da). These digestion rules and variable modifications led to a total of 256,583 distinct peptides.

The Ensemble v75 human protein database was downloaded from `ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/pep`. The same digestion rules and modifications were used for this database as for *Plasmodium*, resulting in 579,698 distinct peptides.

*3720 15th Ave NE, Seattle, WA 98195-5065; Tel: 206 543-8930; Fax: 206 685-7301; Email: william-noble@uw.edu

Nature Methods: doi:10.1038/nmeth.3450

SP-A The human proteome database was downloaded on August 11, 2014, from `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes`. Seven variants and fragments of SP-A were extracted from this database and combined with six variants extracted from the multiple alignment at `http://pubs.acs.org/doi/suppl/10.1021/pr500307f/suppl_file/pr500307f_si_001.pdf` to create the SP-A database. Proteins were digested using trypsin, including proline suppression of cleavages. Only fully enzymatic peptides with two tryptic ends were included in the database, and up to two missed cleavages were allowed. One static modification (carbamidomethylation of cysteine) and no variable modifications were allowed. These rules yielded 99 peptides in the SP-A database and 2,799,144 peptides in the human proteome database.

Note that Foster et al. included in their search two variable modifications: oxidation of Met and deamidation of Asn/Gln. I also replicated this search setting, but we found that adding these modifications increases the size of the human database to 22,238,346 peptides. Due to this increased multiple testing burden, the total number of identified peptides in this case is slightly smaller (data not shown). Therefore, we opted to report results without including variable modifications.

Honeybee The Entrez IDs of the 984 proteins that comprise the viral database used in the initial analysis were provided in a technical report.[4] Unfortunately, because the report is only available as a scanned copy, optical character recognition analysis of the report introduced typos into many of the Entrez IDs. As a result, querying Entrez with the 984 IDs yielded a FASTA file containing only 642 proteins. These 642 proteins come from 15 distinct viruses and 13 *Nosema* species. In addition, a collection of 27,621 honeybee protein sequences was extracted from Entrez using the query "Apis mellifera"[Organism]. Peptide databases were created using tryptic digestion, considering peptides of length 7–50 amino acids and allowing up to 2 missed cleavages. These rules yielded a total of 24,031 viral peptides and 1,424,705 honey peptides. The two sets of peptides share 184 peptides in common.

All data sets and corresponding protein and peptide databases are available at `http://noble.gs.washington.edu/proj/multiple-testing`.

Search procedures

The *Plasmodium* spectra were searched with MS-GF+[5] using the high-res LTQ instrument profile ("-inst 1"), the CID fragmentation model ("-m 1"), a 50 ppm precursor monoisotopic mass tolerance ("-t 50ppm"), a minimum peptide length of 7 ("-minLength 7"), and disallowing isotopic offsets ("-ti 0,0").

The SP-A, *Plasmodium*, and honeybee spectra were searched with Tide[6] as distributed in the Crux toolkit.[7] Candidate peptides were selected using a 3 m/z precursor window for SP-A and honeybee and a 50 ppm window for *Plasmodium*, peptide length of 7–50 amino acids, and exact *P*-value scoring.[8]

Statistical confidence estimation

Decoy peptides were generated by shuffling the non-terminal amino acids of each distinct peptide, and shuffling multiples times if necessary to eliminate redundancies among the set of targets and decoys. Homopolymeric target peptides had no corresponding decoys, resulting in a slightly smaller final set of decoys than targets. False discovery rates were estimated using concatenated database search and setting the FDR to the number of accepted decoys divided by the number of accepted targets.[9]

References

[1] Pease, B. N. *et al.* Global analysis of protein expression and phosphorylation of three stages of *Plasmodium falciparum* intraerythrocytic development. *Journal of Proteome Research* **12**, 4028–4045 (2013).

[2] Foster, M. W. *et al.* Identification and quantitation of coding variants and isoforms of pulmonary surfactant protein A. *Journal of Proteome Research* **13**, 3722–3732 (2014).

[3] Bromenshenk, J. J. *et al.* Iridovirus and Microsporidian linked to honey bee colony decline. *PLOS One* **5**, e13181 (2010).

[4] Wick, C. H. *et al.* Iridescent virus and *Nosema ceranae* linked to honeybee colony collapse disorder. Tech. Rep. ECBC-TR-814, Chemical Biological Center, U. S. Army Research, Development and Engineering Command (2010).

[5] Kim, S., Gupta, N. & Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of Proteome Research* **7**, 3354–3363 (2008).

[6] Diament, B. & Noble, W. S. Faster sequest searching for peptide identification from tandem mass spectra. *Journal of Proteome Research* **10**, 3871–3879 (2011).

[7] McIlwain, S. *et al.* Crux: rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research* **13**, 4488–4491 (2014).

[8] Howbert, J. J. & Noble, W. S. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Molecular and Cellular Proteomics* **13**, 2467–2479 (2014).

[9] Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214 (2007).

Nature Methods: doi:10.1038/nmeth.3450