Supplementary Note

Searching a combined set of spectra

We combined two sets of spectra. The first data set is the ISB18 standard protein mixture [1], consisting of 87,549 spectra derived from a mixture of 18 proteins, downloaded from https://regis-web.systemsbiology.net/PublicDatasets. We used the Orbitrap data set from Mix 7. The proteins come from six different species, including cow, horse, rabbit, chicken, *E. coli* and *B. licheniformis*. The second data set consists of 27,979 spectra derived from *Arabidopsis thaliana* [2], also generated on an Orbitrap LTQ mass spectrometer. The data was downloaded from PRIDE [3] via accession PXD000956.

The peptide database is also a concatenation of two separate databases: the 18 proteins from the ISB18 data set plus the *Arabidopsis* reference proteome from Uniprot (http://www.uniprot.org/proteomes/UP000006548), consisting of 33,469 proteins. To cleave the proteins to peptides, we used the tide-index command in Crux version 3.0 [4], using tryptic digestion with no proline suppression of cleavage, allowing up to two missed cleavages and up to two modifications (oxidation of M and phosphorylation of STY) per peptide, with a length range of 6–30 amino acids. The ISB18 proteins yield 10,758 peptides according to these rules. We identified 37 proteins in the *Arabidopsis* database that shared at least one peptide with the ISB18 data set, so these proteins were removed from the *Arabidopsis* database. The remaining proteins were cleaved to produce 19,932,616 peptides.

We searched the combined spectra against the ISB18 database and against the concatenated database using the Crux implementation of the Tide search engine [5]. We used exact p-value scoring, a precursor window of 10 ppm, isotope error of 1 Th, and we set the software to report the single best match per spectrum in a concatenated target-decoy search. All other search parameters were left at their default values. For "sub-sub," we used the Crux assign-confidence command to estimate the FDR among the set of targets scoring better than a given threshold as $\min(1, (\#\text{ decoys} + 1)/\#\text{ targets})$. For "all-sub," we implemented the "stable target-decoy approach in subset" method proposed by Clement et al. Note that, although the Methods section of Clement et al. suggests estimating their π_0 as $(\#\text{ decoys}_{\text{subset}} + 1)/(\#\text{ targets}_{\text{subset}} + 1)$, the corresponding R code only adds 1 to the numerator. Our implementation follows the R code. In both cases, we define the q-value for score x as the minimal FDR threshold at which x is deemed significant.

Contrary to what Clement et al. found, in this setting the relative performance of the all-sub and sub-sub is reversed: at a 1% FDR threshold, sub-sub accepts 11,416 PSMs, whereas all-sub accepts only 10,307. The source of all-sub's loss of statistical power is easy to track down. Prior to the FDR estimation procedure, the total number of target PSMs produced by the all-sub procedure is 10,307. Thus, the all-sub procedure accepts every single PSM that survives target-decoy competition. However, because of the large size of the *Arabidopsis* database, not enough subset PSMs survive to achieve the same statistical power as sub-sub.

References

[1] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J. E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J. K. Eng, R. Aebersold, and D. B. Martin. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *Journal of Proteome Research*, 7(1):96–1003, 2008.

[2] C. B. Engineer, M. Ghassemian, J. C. Anderson, S. C. Peck, H. Hu, and J. I. Schroeder. Carbonic anhydrases, EPF2 and a novel protease mediate CO2 control of stomatal development. *Nature*, 513(7517):246–250, 2014.

[3] J. A. Vizcaíno, A. Csordas, N. del Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, and T. Ternent. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, 44(D1):D447–D456, 2016.

[4] C. Y. Park, A. A. Klammer, L. Käll, M. P. MacCoss, and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research*, 7(7):3022–3027, 2008.

[5] B. Diament and W. S. Noble. Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research*, 10(9):3871–3879, 2011.