large decoy sets for estimating incorrect PSM score distributions. The **Supplementary Methods** provide more details and introduce diagnostic plots to evaluate decoy quality.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Adriaan Sticker[1–4], Lennart Martens[2–5] & Lieven Clement[1,4,5]

[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. [2]VIB-UGent Center for Medical Biotechnology, Ghent, Belgium. [3]Department of Biochemistry, Ghent University, Ghent, Belgium. [4]Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium. [5]These authors contributed equally to this work. Correspondence should be addressed to L.M. (lennart.martens@vib-ugent.be) or L.C. (lieven.clement@ugent.be).

DATA AVAILABILITY STATEMENT.

Annotated code for all examples in our study is available on Github (https://github.com/compomics/search-all-assess-subset-paper). A user-friendly web-based R tool is available at http://iomics.ugent.be/saas/, and development of the R-package is hosted on GitHub (https://github.com/compomics/search-all-assess-subset). The proteomics data used in this work are available through the ProteomeXchange Consortium on the PRIDE partner repository with data set identifiers PXD001077 (*Pyrococcus furiosus*) and PXD000813 (*Plasmodium falciparum*). Source data for **Figure 1** are available online.

ACKNOWLEDGMENTS

This research was supported by the Ghent University Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to network," VLAIO SBO grant "INSPECTOR" (120025) and the concerted Research Action BOF12/GOA/014, Ghent University.

AUTHOR CONTRIBUTIONS

A.S., L.M. and L.C. designed the research; A.S. and L.C. developed the software; A.S. performed all data analyses; A.S., L.M. and L.C. wrote the manuscript; and all authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Noble, W.S. *Nat. Methods* **12**, 605–608 (2015).
2. Kim, S. & Pevzner, P.A. *Nat. Commun.* **5**, 5277 (2014).
3. Elias, J.E. & Gygi, S.P. *Nat. Methods* **4**, 207–214 (2007).
4. Bourgon, R., Gentleman, R. & Huber, W. *Proc. Natl. Acad. Sci. USA* **107**, 9546–9551 (2010).

Noble and Keich reply: We find much to agree with in Sticker *et al.*[1]. Overall, it is clear that we are engaged in the same general project: to first ensure the validity of our statistical confidence estimates and thereafter to maximize our statistical power in MS-based proteomics experiments. We also agree that controlling the false discovery rate (FDR) among matches to a large peptide database and then reporting results relative to a selected subset of peptides does not correctly control the FDR. Indeed, this point has been made previously on multiple occasions[2,3] and is well established in the statistical literature[4]. We also agree that the 'sub-sub' strategy—searching a subset database and evaluating the FDR within that subset—necessarily forces some matches between peptides in the subset and spectra that were generated by peptides outside of the database.

This leads to our two points of contention. First, Sticker *et al.*[1] claim that their proposed 'all-sub' strategy leads to improved statistical power relative to the sub-sub strategy. In support of this claim, they report empirical results on two data sets. We contend that all-sub is not always better than sub-sub. Accordingly, we constructed a different setup that allowed us to more accurately characterize false positive spectrum identifications. Specifically, we ran a concatenated set of spectra—from 18 purified proteins (ISB18)[5] and from the plant *Arabidopsis thaliana*[6]—against a corresponding concatenated database. Contrary to what Sticker *et al.*[1] found, in this setting the relative performance of the two methods is reversed: at a 1% FDR threshold, sub-sub accepts 11,416 peptide–spectrum matches (PSMs), whereas all-sub accepts only 10,307. We conclude that all-sub's loss of statistical power is due to the large size of the *Arabidopsis* database (**Supplementary Note**).

Second, in addition to claiming superior statistical power of the all-sub procedure, Sticker *et al.*[1] imply that the sub-sub strategy leads to invalid FDR control. As evidence, they point to the number of subset PSMs that matched a different peptide sequence in the complete search (all-all) and the subset search (sub-sub). However, their analysis does not account for the possibility that some of these PSMs may be incorrect in the all-all search and correct in the sub-sub search. Indeed, as the size of the competing, complement database increases, the probability that a correct match to the subset database will receive a lower score than an incorrect match in the complement database increases. This is precisely the effect that sub-sub aims to avoid. In the context of this simulation, Sticker *et al.*[1] are concerned that by forcing *Arabidopsis* spectra to match against the ISB18 database, we will create many false positive PSMs. Fortunately, in our experimental setup, we can directly observe this rate of false matching: among the 11,416 PSMs accepted by sub-sub, only 41 (0.36%) involve an *Arabidopsis* spectrum. This is well below the 1% FDR threshold. Furthermore, we note that in the subset database search, 1,127 of the accepted PSMs involving ISB18 spectra actually switch to matching *Arabidopsis* peptides when we search against the combined database. According to the arguments laid out by Sticker *et al.*[1], this rate of switching implies that that the actual sub-sub FDR is ~10%. However, in our setup, we know that those ISB18 spectra are definitely not correct when matched to *Arabidopsis* peptides.

Thus, though all-sub may provide superior statistical power in some settings, this is not always the case. Precisely characterizing the situations in which a given analysis strategy is optimal will require further research.

Data availability statement. All data used in this work are publicly available via the URLs listed in the **Supplementary Note**.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

William Stafford Noble[1,2] & Uri Keich[3]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA. [3]School of Mathematics and Statistics, University of Sydney, Sydney, New South Wales, Australia. Correspondence should be addressed to W.S.N. (william-noble@uw.edu) or U.K. (uri@maths.usyd.edu.au).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health awards R01 GM121818 and P41 GM103533 to W.S.N.

1. Sticker, A., Martens, L. & Clement, L. *Nat. Methods* **14**, 643–644 (2017).
2. Baker, P.R., Medzihradszky, K.F. & Chalkley, R.J. *Mol. Cell. Proteomics* **9**, 1795–1803 (2010).
3. Fu, Y. & Qian, X. *Mol. Cell. Proteomics* **13**, 1359–1368 (2014).
4. Efron, B. *Ann. Appl. Stat.* **2**, 197–223 (2008).
5. Klimek, J. *et al. J. Proteome Res.* **7**, 96–103 (2008).
6. Engineer, C.B. *et al. Nature* **513**, 246–250 (2014).