To generate the deepest human phosphoproteome to date, Ochoa et al.[3] reanalyzed 112 large-scale phosphoproteomics datasets from the PRIDE database[9], the largest data repository of MS-based proteomics data. They used machine learning to rank the biological relevance of more than 100,000 human phosphosites by assigning each of them a 'functional score', which integrates prior knowledge from experimental data, prediction tools and MS signal properties for prioritizing their biological importance (Fig. 1).

Bulk analysis of phosphoproteomics datasets can lead to high rates of false positives owing to their random origin. The systematic reanalysis performed by Ochoa et al.[3] controlled the rate of false positives and therefore serves as the largest high-quality and unified dataset, comprising 119,809 human phosphosites derived from 104 different cell types or tissues (Fig. 1). To learn which features are important for phosphosite function, they trained the machine-learning algorithm using a subset of phosphosites with annotated biological function. The resulting functional score for ranking phosphosites integrated information on 59 features in 4 main categories: MS evidence, evolutionary conservation, kinase regulation and protein structure (Fig. 1).

Interestingly, the relative importance of the selected features for phosphosite scoring differed between serine/threonine and tyrosine phosphorylations. Localization of phosphosites in the cytosolic portion of transmembrane proteins offered a high discriminative power in the model for tyrosine sites, whereas phosphosite evolutionary age, adjacent post-translational modifications and protein length were the most informative features for serine/threonine sites.

A major strength of the study by Ochoa et al.[3] is that the authors went to great lengths to validate their functional score in follow-up experiments. They convincingly demonstrated that the functional score of a phosphosite on a transcription factor strongly correlates with its activity and that high-scoring phosphosites on enzymes have functional consequences for cell fitness when mutated. They further developed these observations by highlighting the biological importance of several phosphosites in the transcription factor STAT1 and in glyceraldehyde 3-phosphate dehydrogenase.

Ochoa et al.[3] also showed that the functional score for known pathogenic mutations is generally high, especially for tyrosine sites. In line with this finding, we recently demonstrated that mutations near phosphorylated tyrosine residues can have oncogenic properties[10]. Integration of information about amino acid substitutions near phosphosites with the phosphosites' functional scores may help researchers prioritize and pinpoint disease-inducing mutations in phosphoproteins and may help predict the function of such sites.

The functional scores provided by Ochoa et al.[3] will be useful to systematically prioritize phosphosites for biological validation (Fig. 1). We hope that a user-friendly tool or web-based interface implementing the machine-learning algorithm to functionally score and rank phosphosites will be made available. With continuous improvements in phosphoproteomics technologies and their applications, many more phosphosites will be discovered in the future that should be incorporated into the model. We also expect that the approach of Ochoa et al.[3] will be extended to popular model organisms, such as mouse, zebrafish, roundworm, fruit fly and yeast. ❐

Giulia Franciosa, Ana Martinez-Val and Jesper V. Olsen ✉

Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
✉*e-mail: jesper.olsen@cpr.ku.dk*

Published online: 13 February 2020
https://doi.org/10.1038/s41587-020-0441-3

References

1. Olsen, J. V. et al. *Cell* **127**, 635–648 (2006).
2. Needham, E.J., Parker, B.L., Burykin, T., James, D.E. & Humphrey, S.J. *Sci. Signal.* **12**, eaau8645 (2019).
3. Ochoa, D. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0344-3 (2019).
4. Wilson, L. J. et al. *Cancer Res.* **78**, 15–29 (2018).
5. Hornbeck, P. V. et al. *Nucleic Acids Res.* **43**, D512–D520 (2015).
6. Casado, P. et al. *Sci. Signal.* **6**, rs6 (2013).
7. Beekhof, R. et al. *Mol. Syst. Biol.* **15**, e8981 (2019).
8. Francavilla, C. et al. *Nat. Struct. Mol. Biol.* **23**, 608–618 (2016).
9. Perez-Riverol, Y. et al. *Nucleic Acids Res.* **47D1**, D442–D450 (2019).
10. Lundby, A. et al. *Cell* **179**, 543–560.e26 (2019).

Competing interests

The authors declare no competing financial interests.

Check for updates

PODCAST

Forum: Molecular mapping of tumor heterogeneity

Brady Huggett talks with Zemin Zhang about a recent *Nature Biotechnology* paper detailing the use of spatial transcriptomics to provide insights into tumor architecture. The work was done by Itai Yanai and coauthors, and the paper can be read at https://doi.org/10.1038/s41587-019-0392-8.
https://play.acast.com/s/forum

Published online: 9 March 2020
https://doi.org/10.1038/s41587-020-0445-z