BoxCar method enables single shot proteomics at a depth of 10,000 proteins in 100 minutes

Florian Meier[1], Philipp E. Geyer[1,2], Sebastian Virreira Winter[1], Juergen Cox[3] and Matthias Mann[1,2]*

[1] *Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany*

[2] *NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark*

[3] *Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany*

*To whom correspondence may be addressed: mmann@biochem.mpg.de

ABSTRACT

Great advances have been made in sensitivity and acquisition speed on the Orbitrap mass analyzer, enabling increasingly deep proteome coverage. However, these advances have been mainly limited to the MS2 level whereas ion beam sampling for the full MS scan remains extremely inefficient. Here we report a novel acquisition method, termed BoxCar, in which filling multiple narrow mass-to-charge segments increases the mean ion injection time more than ten-fold compared to a standard full scan. In 1 hour analyses, the method provided MS1 evidence for more than 90% of the proteome of a human cancer cell line that had previously been identified in 24 fractions, and quantified more than 6,200 proteins in ten of ten replicates. In brain tissue, we achieved broad proteome coverage in only 100 min and sensitivity extended into the low attomol range. MS2 acquisition strategies will also benefit from the MS1 level improvements of BoxCar.

INTRODUCTION

The proteome of a mammalian cell comprises expression products from over 10,000 protein-coding genes with copy numbers ranging from tens to over one million per cell and body fluids such as plasma have an even larger dynamic range[1–3]. Mass spectrometry (MS)-based proteomics has become very powerful in elucidating this highly complex structural and functional entitiy[4,5]. However, further improvements in its core technology – mass spectrometry – are urgently needed to more fully capture the proteome in its entire complexity.

In shotgun proteomics, MS1 scans measure the mass-to-charge ratio (*m/z*) and signal intensity of peptides, whereas the function of MS2 scans is to determine fragmentation patterns for peptide identification in databases[6]. Innovations over the last decade have almost exclusively focused on the sensitivity and speed of peptide fragmentation by tandem mass spectrometry (MS2)[7,8], which includes the development of data-independent acquisition[9,10] and peptide-centric analysis[11]. This is particularly the case for Orbitrap mass analyzers, which have become much faster, but whose major bottleneck remains the charge capacity of the upstream ion storage device (C-trap) of only about one million charges[12,13]. With modern high transmission ion sources and typical peptide amounts on column, this value is often reached in less than 1 ms. This is less than 1% of the required transient time for a high resolution mass spectrum (128 ms), implying that more than 99% of the generated ions are not used for mass analysis. In contrast, when only a small proportion of the total ion current is admitted, high ion currents and high resolution can directly be used to increase proteome sampling depth[14-16].

The ion current in MS-based proteomics is generated by electrospraying hundreds of co-eluting peptides into the mass spectrometer and, in MS1 mode, ions from the full mass range are trapped together. As a consequence of the high dynamic range of the proteome and the different ionization properties of peptides, the total signal is concentrated within a few peaks rather than evenly distributed across the full mass range. We thus reasoned that distributing the maximal charge capacity of one million evenly over multiple narrow *m/z* segments would limit the proportion of high abundant species in the C-trap and greatly increase ion injection (or 'filling') times for less abundant species. Our strategy is built on the ability of the quadrupole–Orbitrap mass analyzer to be filled sequentially with different mass windows, which are then analyzed

together in a single scan[17,18]. We termed this novel acquisition mode 'BoxCar´, and show that it drastically improves performance in MS1 mode.

RESULTS

Development of the method

To balance the absolute ion count across the full mass spectrum, we decomposed it into narrow *m/z* segments, which are each allotted the same proportion of the total incoming ion current – for instance 100,000 charges for each of ten segments. In the mass spectrometer, each segment is sequentially selected by the analytical quadrupole, its ions combined in the C-trap, and the ions from all segments mass analyzed together in the Orbitrap (**Fig. 1a**). However, combining adjacent segments in a single scan causes artefacts at the respective edges and precludes accurate quantification. Instead, we allotted the full mass range to multiple, subsequent scans (three in **Figure 1b**), each of which is defined by interspaced, rectangular 'BoxCar' patterns for mass selection. The term BoxCar refers to the mathematical on/off shape of the transmission function. The superimposition of the BoxCar scans in an acquisition cycle reconstructs a full scan with the difference that the boxes with low ion signal benefit from a dramatic increase in fill time, while the ion proportion from the few boxes with very high ion current is restricted. In **Figures 1c and d**, the summed ion injection time of the BoxCar scan (177 ms) dwarfs the ion injection time of the standard full scan (1 ms) and the average ion injection times of individual boxes is 22-fold higher than in the standard full scan. As a result, the low abundant species with poor signal-to-noise ratios (S/N) were now detected at 30 to 60-fold increased S/N. Furthermore, we observed an additional isotope pattern at nominal m/z 862, which had completely escaped detection in the full scan.

Next, we systematically investigated the effect of the acquisition parameters on the number of quantified peptide features in a single shot experiment using a 'statistical Design of Experiment' approach (**Supplementary Tab. 1, Supplementary Fig. 1**). Small boxes have longer injection times, greater S/N, dynamic range and improved 'isolation' of very high abundance peaks. However, this needs to be balanced by the total available fill time, as well as the overall cycle time in a topN acquisition scheme. To achieve cycle times around one second for the full

acquisition cycle of one full scan, BoxCar scans and data-dependent MS/MS scans, we here limited the number of BoxCar scans to two or three per acquisition cycle. For the same reason, we limited the number of MS/MS scans to five – with precursors selected from full scans. This yielded on average eight data points per chromatographic peak with two BoxCar scans for the 45 min gradient. We further scaled box widths inversely proportional to the *m/z* density of tryptic peptides (**Supplementary Tab 2**).

Having established sensible parameters for BoxCar, we did a first comparison to the established shotgun method with standard full scans. Analyzing a HeLa digest, BoxCar doubled the number of detected multiply charged isotope patterns and increased the dynamic range by nearly one order of magnitude (**Supplementary Fig. 2, Supplementary Tab. 3**).

Reproducible and accurate protein quantification

Ion transmission through the mass selecting quadrupole inevitably deviates from an ideal rectangular shape, particularly at the edges of the isolation window. To alleviate these 'edge effects' we incorporated a small overlap of neighboring boxes on the *m/z* axis. Further correcting for non-uniform ion transmission efficiencies was crucial for the detection and accurate quantification of isotope patterns, in particular those that reach across the borders of a single box. To achieve this we adapted the post-processing workflow in MaxQuant[19] (**Online Methods**, **Fig. 2a**). Our algorithm first assembles all BoxCar scans and the preceding full scans on a common high-resolution *m/z* grid. Next, all ion signals from these scans are integrated over the entire LC runtime to determine the empirical ion transmission function by comparison to the full scan. The resulting relative ion transmission factors for each *m/z* bin are then applied to correct the relative ion abundances in each individual BoxCar scan. Finally, for each acquisition cycle, a single high dynamic range spectrum is calculated as the weighted average of the standard full scan and the intensity-corrected BoxCar scans.

Using the full information over the entire elution time makes our algorithm robust against inter-scan and time-dependent fluctuations in the ion current. Further downstream analysis in MaxQuant remained unchanged, while greatly benefiting from the increased spectral quality and dynamic range.

To test the quantitative accuracy of the resulting hybrid spectra, we first analyzed a mixture of heavy and light SILAC-labeled HeLa lysate, in which BoxCar accurately quantified the pipetting ratio of 1:3 throughout the entire *m/z* range (**Supplementary Fig. 3**). To further benchmark the BoxCar method for label-free quantification, we assessed the quantitative reproducibility and accuracy between ten technical replicate injections of a whole proteome digest and compared it to our standard method (**Fig. 2b**). In total, we identified 4,302 protein groups, of which on average 4,156 were quantified per 45 min run. Median coefficients of variation (CV) were 17% and 23% for the non-normalized peptide intensities with BoxCar and the standard method, respectively, and 8% on the protein level after label-free normalization for both. Pairwise comparisons of all replicates revealed an excellent quantitative reproducibility of the BoxCar method with a median Pearson correlation of 0.98 on the protein level, virtually identical to the overall correlation of standard full scan quantifications to each other (r = 0.98). Likewise, BoxCar–BoxCar correlations were very similar to BoxCar–full scan correlations, both on the peptide and protein level, further confirming the absence of any systematic bias (insets **Fig. 2b, Supplementary Fig. 4**). In a two-proteome experiment (HeLa and E.coli lysate mixed in 1:2 and 1:12 ratios), BoxCar performed particularly well for low abundant E.coli proteins that were quantified with either low accuracy or missing values with the standard method, indicating an approximately 10-fold increased dynamic range (**Supplementary Fig. 5**).

Extended dynamic range in body fluids

Analysis of body fluid proteomes is one of the most important clinical applications of proteomics, but also very challenging due to their extreme dynamic range, which exceeds ten orders of magnitude for plasma[20]. The finite intra-scan dynamic range severely limits our ability to detect low abundant proteins in standard single run measurements[21,22]. In trapping instruments such as the Orbitrap analyzer, this is mainly due to peptides from highly abundant proteins, such as serum albumin. **Figure 3a** visualizes detected peptide features in standard full scans and the total ion chromatogram, which is dominated by individual abundant peptides for much of the elution time. While these ions elute, the mass spectrometer is essentially blind to all other peptides, resulting in sparse regions in the *m/z* – retention time plane. In stark contrast, the BoxCar method effectively limits the proportion of abundant species by confining them to narrow mass windows with an equalized number of charges (**Fig. 3b**).

In the 16-22 min elution window, for instance, injection times for the standard MS spectrum were only 0.15 ms - corresponding to a duty cycle of 0.12% for a 128 ms transient. In the BoxCar measurement, ion injection time scaled with the expected ion current in each *m/z* window up to the defined maximum of 21 ms. This 140-fold enhanced ion fill time revealed previously hidden, co-eluting isotope features and thus potential peptides (**Fig. 3a,b**). Gains are particularly high in less populated *m/z* ranges, however, the median gain across the full mass range was more than 10-fold (**Fig. 3b**).

In total, we detected 249,162 multiply charged isotope patterns with BoxCar, four-fold more than the 59,491 seen by the standard method (**Fig. 3c**). Patterns identified in both methods exhibited more isotope peaks and were detected for longer elution times (**Supplementary Fig. 6**). The dynamic range improvement was uniform over the retention time but increased with *m/z*, for an overall ten-fold gain, consistent with the gain in ion fill time (**Fig. 3 d,e**).

Deep proteome quantification with BoxCar

For very deep proteome coverage, peptide pre-fractionation is generally employed as it increases the overall amount of material that can be loaded and the time available for peptide sequencing[8,23]. Given the greatly improved dynamic range and number of detected features of the BoxCar method, we asked if comprehensive proteome characterization could be reached by combining single BoxCar runs with a peptide library[24,25]. In such an approach, peptide identifications would be transferred from the library using the 'match between runs' feature of MaxQuant[26], while the quantitative information is provided by BoxCar single runs (**Fig. 4a**). Unlike a standard shotgun approach, the library strategy is not limited by the MS/MS scan speed of the mass spectrometer and, with BoxCar, a much larger proportion of the incoming ion current is used for quantification.

To test this concept, we acquired a library of HeLa peptides from 24 high-pH reversed-phase fractions[27], comprising 128,720 sequence unique peptides from 9,210 protein groups. Matching was facilitated by sub-ppm absolute average mass accuracies as well as median absolute elution time differences of 1.7 s after alignment (compared to median peak widths of 11.3 s). On average, we assigned 80,714 isotope patterns to peptide sequences in 45 min BoxCar runs (**Supplementary Tab. 3**), corresponding to a maximum of 50 peptide features per second

(**Supplementary Fig. 7**). At typical MS/MS identification rates, this would otherwise have required a sequencing rate of more than 60 Hz. Likewise, matching the library to standard shotgun runs was previously limited by the dynamic range of the full scan, which is now expanded up to one order of magnitude throughout the LC gradient and mass range (**Supplementary Fig. 7**).

Each BoxCar run identified on average 60,228 unique peptide sequences and 7,775 protein groups from 1 µg of peptides (**Fig. 4b**). Combining the ten BoxCar single runs retrieved two thirds of all unique peptide sequences and 93% of all protein groups present in the library, with 7,508 proteins identified by two or more peptides in at least one replicate. We reasoned that this level of coverage should be sufficient to estimate the copy numbers and stoichiometries of most protein complexes in the cell. As an example, we studied the six-membered minichromosome maintenance (MCM) complex, a key factor in DNA replication[28] (**Fig. 4c**). In eukaryotes, MCM associates with the four-membered GINS complex and the initiation factor CDC45, also known as the CMG complex[29]. From our 45 min BoxCar single runs, we estimated an abundance of about 750,000, 150,000 and 55,000 copies per cell for MCM, GINS and CDC45, in agreement with our previous results from fractionated samples[30]. Moreover, all members of the MCM and GINS complexes were quantified with the expected 1:1 stoichiometry (**Fig. 4d**).

Alleviating the 'missing value' problem

The semi-stochastic nature of precursor selection in shotgun proteomics results in not every peptide being quantified in every run – often referred to as the 'missing value' problem[31,32]. The complete quantification matrix of ten HeLa runs analyzed with a standard shotgun method comprises 37,869 unique peptide sequences and 5,050 protein groups of which only 11,626 peptides and 2,789 proteins where quantified in each of the 45 min runs (**Fig. 5 a,b**). 'Matching between runs' in MaxQuant greatly mitigated this problem by transferring peptide identifications between experiments and lead to complete quantification of 25,653 peptides and 4,180 protein groups in ten replicates. BoxCar in conjunction with the peptide library achieved even more consistent profiles across all replicates, providing quantification of approximately 70% of all peptides and >95% of all proteins from the entire shotgun experiment in every replicate. In total, BoxCar quantified 7,222 proteins per 45 min run with a median CV of 10% and more than

doubled the proportion peptides and proteins quantified in ten of ten runs to 36,377 and 6,216, respectively.

Next, we calculated the completeness of the data matrices as a function of the peptide and protein abundance rank (**Fig. 5 c,d**). For the standard shotgun method, both curves decline fast and reach 60% and 75% at the full depth. With matching between runs the curve is much shallower and only 11% of the peptide and 6% of the protein data points were missing. The BoxCar workflow outperformed both methods in terms of proteome coverage and reproducibility. From the 54,000 most abundant peptides, only 10% of the values were missing and the protein data completeness does not drop below 95% until a depth of 7,450 proteins. Even at the full depth of 8,253 protein groups, the fraction of missing data points was as low as 12.5%. Our results compare favorably to reported sparsity levels in data-independent acquisition methods, which are furthermore reached at lower proteome coverage[33–35].

Ten thousand protein groups in 100 min

The excellent depth and reproducibility in single shot analyses of cell lines encouraged us to investigate the performance of BoxCar in a highly complex tissue sample from mammalian brain[36]. First, we generated a deep peptide library from fresh frozen mouse cerebellum comprising 153,405 unique peptide sequences from 11,328 protein groups with a median sequence coverage of 30%. We then analyzed quintuplicates of un-fractionated cerebellum digest in 100 min gradients. BoxCar yielded over 100,000 identified features per single run, 60% more than the analogue shotgun/library approach and spanning more than one additional order of magnitude (**Supplementary Fig. 7**). Overall, the BoxCar library-based workflow identified on average 72,896 unique peptide sequences and over 10,000 protein groups in each replicate, which is equivalent to the identification of 100 proteins per minute over the entire gradient (**Fig. 6a**). A total of 9,270 protein groups were detected in all five replicates and 10,207 in at least three out of five replicates (**Fig. 6b**). We found evidence for 65% of all unique peptide sequences and 92% of all library proteins in our 100 min BoxCar runs, with 9,583 proteins being identified with two or more peptides in at least one replicate. Protein abundances, as approximated by summed and normalized peptide intensities, spanned over six orders of magnitude and we

estimate the amount of the 1,200 least abundant proteins to be less than ten attomol (**Fig. 6c, Supplementary Fig. 8**). Label-free protein quantification was highly reproducible with a median CV of 9.9% (20% for non-normalized peptide and summed peptide intensities).

Coverage of biological functions and neural compartments in the single run measurements in comparison with the cerebellum library was almost complete for structural proteins; for instance, 100% of all proteins constituting the node of Ranvier and 98.5% of those associated with actomyosin (**Fig. 6d**). Likewise, proteins involved in neuronal activity were broadly covered, including 97.9% and 98.5% of those annotated as 'synaptic vesicles' and 'postsynaptic density', respectively.

Bioinformatic analysis indicated that high and medium abundant proteins were associated with key metabolic functions, such as the tricarboxylic acid cycle, and important neuronal functions, for instance vesicle fusion (**Fig. 6c**). Members of transcription factor complexes were among the lower abundant proteins, often with estimated copy numbers of less than 10,000 per cell (**Supplementary Fig. 8**). An array of proteins that we had previously reported to be specifically enriched in cerebellum as compared with other brain regions[36] were distributed across the entire abundance range. These included regulatory proteins (Syn2, Sfn), scaffolding proteins (Grid2ip, Homer3), as well as proteins that are crucial for cerebellar development (Ca8, Fat2).

DISCUSSION

A key limitation of trapping-based mass spectrometers – including the Orbitrap - is the limited charge capacity of the ion trap[13] which excluded the majority of ions from MS1 level analysis. Here, we address this limitation with a novel acquisition method that utilizes the incoming ion current far more efficiently by redistributing the proportion of ions allocated across the mass range. The BoxCar method increases MS1 level dynamic range by one order of magnitude on average and even more in the case of very high dynamic range proteomes such as body fluids. BoxCar does not require any changes to the hardware and the post-processing is fully integrated in MaxQuant. Insertion of BoxCar scans lengthens the acquisition cycle which we balanced by reducing the number of MS/MS scans and transferring identifications from a library. Ongoing

improvements in resolving power[37] can be fully harnessed and would allow BoxCar scans to be inserted into an acquisition cycle with less effect on its duration.

Library-based workflows are well-established for DIA[11,38] and the increased spectral quality and the extended dynamic range of the BoxCar method make this strategy also very attractive on the MS1 level, even though stringent control for false matches remains challenging and currently relies on reproducible chromatography as well as project-specific libraries. In a human cancer cell line, we consistently quantified almost the entire contents of the protein library in only 1h, using 4% of the sample amount. Likewise, the 10,000 protein groups detected in 100 min gradients of the cerebellum proteome broadly covered the relevant biological functions from minimal input material. Thus, the combination of BoxCar scans with a peptide library essentially delivers the results of a pre-fractionation approach without the attendant consumption of sample and measurement time.

While demonstrated here in the context of proteomics, the BoxCar method is generic and could seamlessly be applied in other fields with similar dynamic range challenges, for instance metabolomics. Note that all the advantages of BoxCar follow from its greater dynamic range at the MS1 level. Therefore, they are not limited to shotgun proteomics, but should also benefit any method relying on accurate and repeatable picking of peptide precursors, such as targeting and isobaric tagging[39,40]. Furthermore, MS2-level data-independent acquisition methods would also profit from BoxCar by much improved precursor information.

Acknowledgements

Author contributions

F.M. and M.M. conceptualized the method; P.G. contributed to the development and J.C. designed and implemented the post-processing algorithms; F.M. and P.G. performed the experiments; and F.M., P.G., S.V.W., J.C., and M.M. analyzed the data; F.M. and M.M. wrote the manuscript with input from all authors.

Competing financial interests

F.M., J.C., and M.M. are inventors on a patent application covering the method described herein (Applicant: Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V; Application number: PCT/EP2018/051290; Status: pending).

References

1. Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537,** 328–338 (2016).

2. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).

3. Schwenk, J. M. *et al.* The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J. Proteome Res.* **16,** 4299–4310 (2017).

4. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355 (2016).

5. Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Mol. Cell* **49,** 583–590 (2013).

6. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682 (1999).

7. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14,** 35–48 (2012).

8. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24,** 11–17 (2015).

9. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1,** 39–45 (2004).

10. Chapman, J. D., Goodlett, D. R. & Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom. Rev.* **33,** 452–470 (2014).

11. Ting, Y. S. *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol. Cell. Proteomics* **14,** 2301–2307 (2015).

12. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem.* **8,** 61–80 (2015).

13. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **85,** 5288–96 (2013).

14. Yi, E. C. *et al.* Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23**, 3205–3216 (2002).

15. Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. Cell. Proteomics* **16,** 2296–2309 (2017).

16.   Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* acs.jproteome.7b00602 (2017). doi:10.1021/acs.jproteome.7b00602

17.   Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10,** M111.011015 (2011).

18.   Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods* **10,** 744–746 (2013).

19.   Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–72 (2008).

20.   Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13,** 942 (2017).

21.   Geyer, P. E. *et al.* Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2,** 185–195 (2016).

22.   Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11,** 786 (2015).

23.   Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **4,** 587–599.e4 (2017).

24.   Pasa-Tolic, L., Masselon, C., Barry, R. C., Shen, Y. & Smith, R. D. Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* **37,** 621–636. (2004).

25.   Zhang, B., Kall, L. & Zubarev, R. A. DeMix-Q: Quantification-centered Data Processing Workflow. *Mol. Cell. Proteomics* **15**, 1467-1478 (2016).

26.   Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol. Cell. Proteomics* **11,** M111.014050-M111.014050 (2012).

27.   Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694-705 (2017).

28.   Tye, B. K. MCM proteins in DNA replication. *Annu. Rev. Biochem.* **68,** 649–686 (1999).

29.   Moyer, S. E., Lewis, P. W. & Botchan, M. R. Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 10236–10241 (2006).

30.   Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11,** 319–24 (2014).

31.   Bruderer, R. *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol. Cell. Proteomics* **14,** 1400–1410 (2015).

32. Rost, H. L., Malmstrom, L. & Aebersold, R. Reproducible quantitative proteotype data matrices for systems biology. *Mol. Biol. Cell* **26,** 3926–3931 (2015).

33. Röst, H. L. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13,** 1–14 (2016).

34. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34,** 1130-1136 (2016).

35. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8,** 291 (2017).

36. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci.* **18,** 1–16 (2015).

37. Grinfeld, D., Aizikov, K., Kreutzmann, A., Damoc, E. & Makarov, A. A. Phase-Constrained Spectrum Deconvolution for Fourier Transform Mass Spectrometry. *Anal. Chem.* **89,** 1202-1211 (2017).

38. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics* **11,** O111.016717-O111.016717 (2012).

39. Gallien, S., Kim, S. Y. & Domon, B. Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction. *Mol. Cell. Proteomics* **14,** 1630–1644 (2015).

40. Erickson, B. K. *et al.* A Strategy to Combine Sample Multiplexing with Targeted Proteomics Assays for High-Throughput Protein Signature Characterization. *Mol. Cell* **65,** 361–370 (2017).
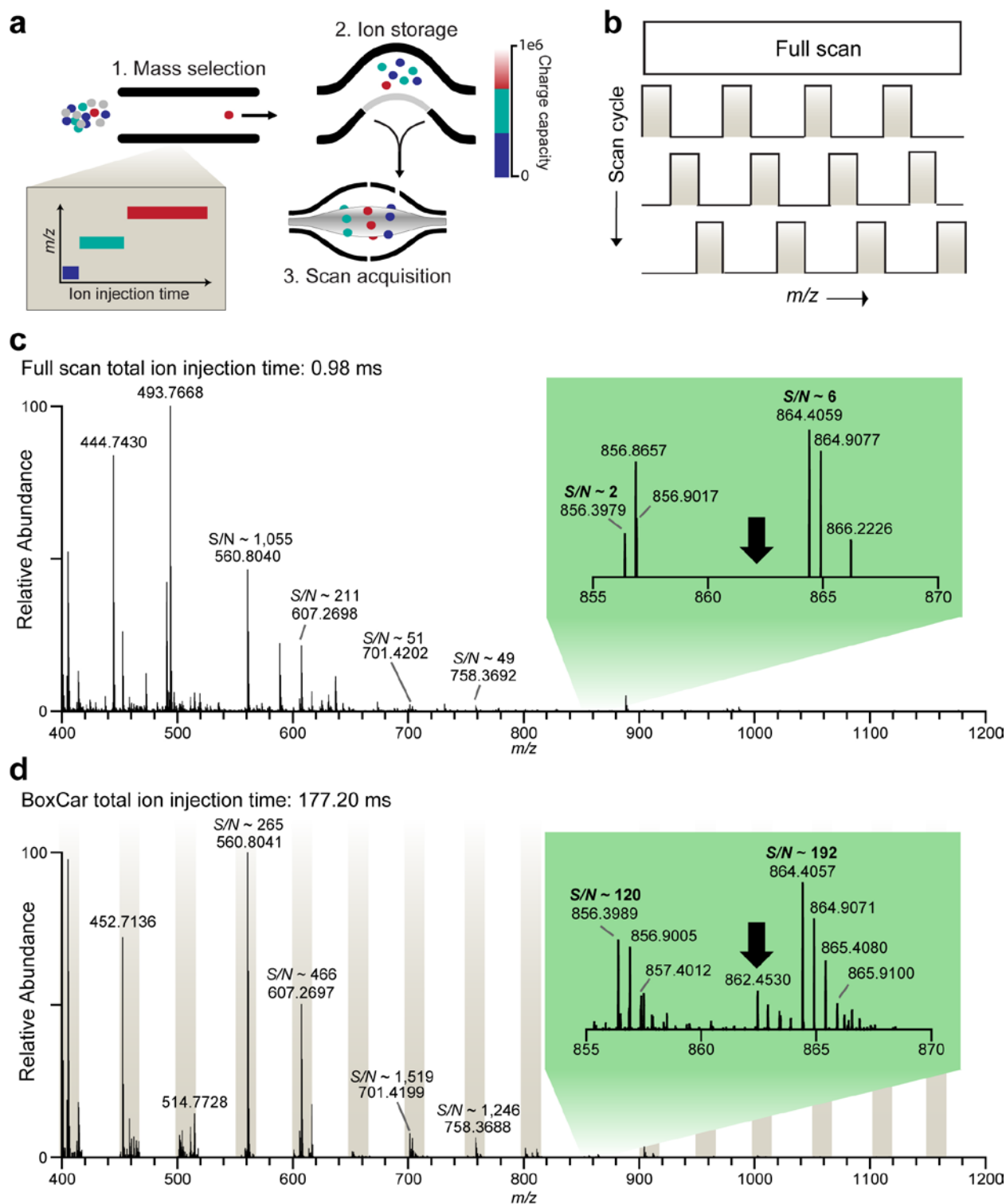
Figures

Figure 1. The BoxCar acquisition method. (**a**) The high dynamic range of the incoming ion population is normalized by (1) sequentially filtering narrow *m/z* windows in the analytical

quadrupole, (2) accumulating the ions in the C-trap and (3) analyzing the stored ion population in a single Orbitrap scan. (**b**) The full mass range is covered by an interspaced, rectangular boxcar function for mass selection in subsequent scans. (**c**) Representative example of a standard full scan in shotgun proteomics. (**d**) The adjacent BoxCar scan composed of 16 narrow isolation windows ('boxes'). The green insets highlight an apparently empty *m/z* region in the full scan that was filled for 16 ms in the BoxCar scan. Signal-to-noise ratios (S/N) are provided for comparison.
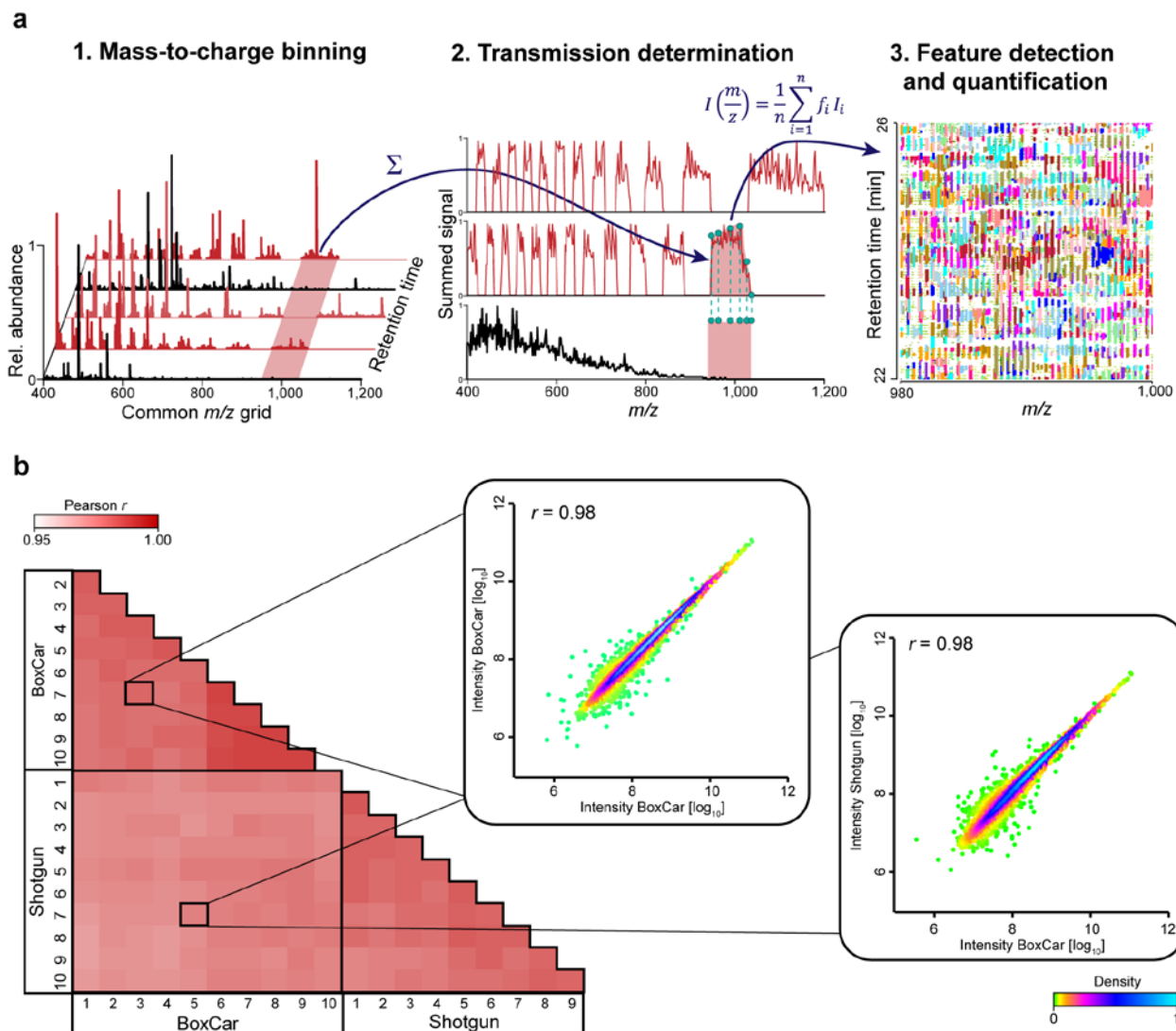
Figure 2. Label-free protein quantification with BoxCar scans. (**a**) MaxQuant assembles the full scan (black traces) and BoxCar scans (red traces) of an acquisition cycle to a single high dynamic range scan. First, all spectra are transformed to a common high-resolution *m/z* grid and the signals from each scan (here: one full scan and two BoxCar scans) are integrated over the entire LC elution time (step 1). From the integrated signals, the shape of the quadrupole transmission function for each BoxCar scan is globally determined by a pointwise comparison to the full scan (step 2). The resulting relative transmission factors for each *m/z* bin in each BoxCar scan are used as weights for calculating the average signal intensity from the full scan and the

17

BoxCar scans. These hybrid spectra are taken as a replacement for standard full scans in all subsequent processing steps without further adjustments (step 3). See main text and Online Methods for further details. (**b**) Pearson correlation analysis of label-free protein intensities in ten replicate 45 min HeLa runs with BoxCar or a standard shotgun method (N=4,292 proteins). The median correlation coefficients were 0.98 for BoxCar and shotgun replicates, and 0.98 for the cross-correlated protein intensities. Density plots illustrate the protein intensity correlation of two representative pairs.
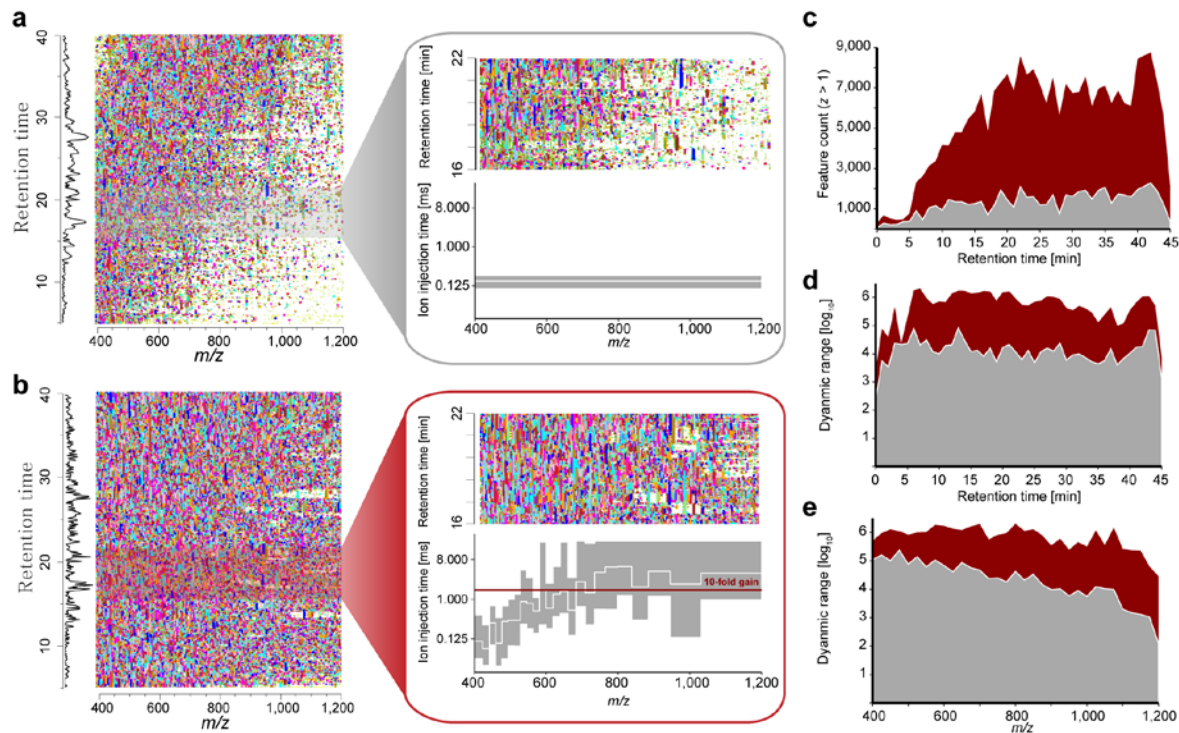
Figure 3. High dynamic range analysis of a human plasma sample in single runs. (**a**) Isotope patterns detected in the *m/z*-retention time plane with standard full scans are color-coded. (**b**) Same for BoxCar scans, revealing many more isotope patterns. Median ion injection times (full scans *N*=604, BoxCar scans: *N*=347) are illustrated in the insets as a function of *m/z* (white line). Top and bottom of the grey boxes represent the 75% and 25% percentiles, respectively. (**c**) Detected multiply charged features per minute over the entire gradient with the standard (grey) and BoxCar method (red). (**d**) Dynamic range of the detected features as a function of retention time and (**e**) *m/z*.
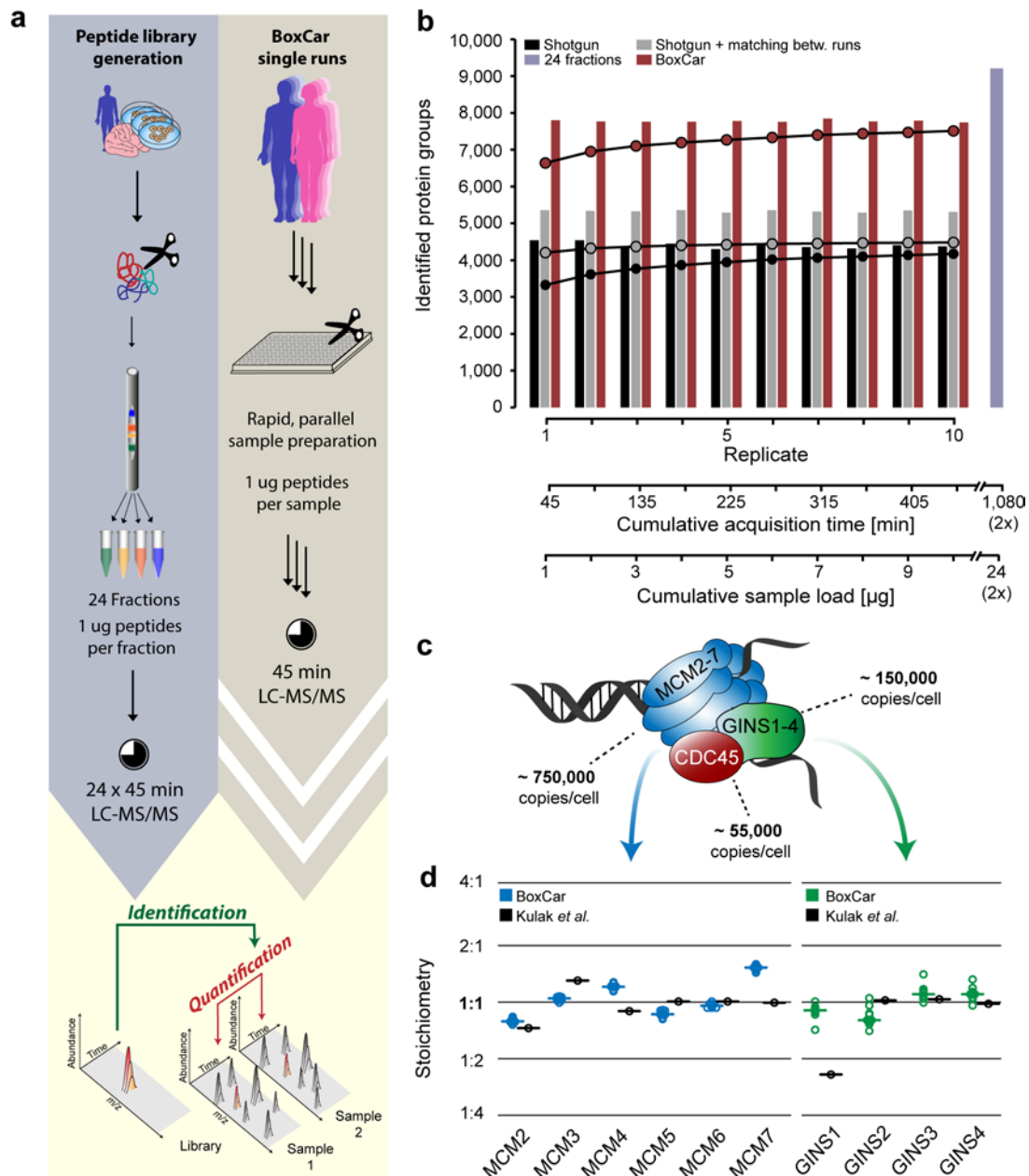
Figure 4. Library approach for deep proteome coverage in BoxCar single runs. (**a**) Schematic workflow for peptide library generation from fractionated protein digests (left panel) and high-throughput proteome quantification in BoxCar single runs (right panel). (**b**) Comparison of protein identifications in ten 45 min analyses of a HeLa digest with a standard shotgun method, with 'matching between runs' and with the BoxCar approach using a library. Line plots indicate the cumulative number of proteins identified with two or more peptides in at least one replicate. Cumulative peptide load and measuring times (bottom axes) show that BoxCar minimizes the

requirements for acquisition time and starting material, while the proteome coverage reaches the level of 24 high-pH reversed-phase fractions. (**c**) Representation of the CMG complex during DNA replication together with copy numbers estimated from BoxCar single runs. (**d**) Stoichiometry of the MCM and GINS complexes reconstructed from relative copy numbers in BoxCar single runs and in six SCX fractions[30]. Horizontal bars indicate medians (*N*=10 replicate injections).
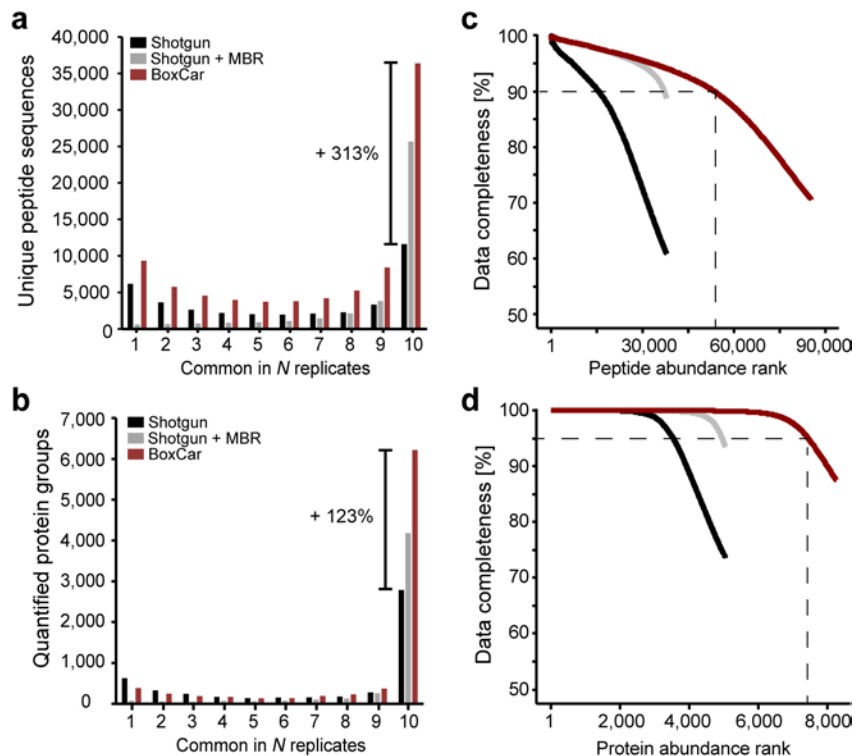
Figure 5. Assessment of missing value rates in ten replicate 45 min analyses of HeLa digest. (**a**) Number of unique peptide sequences quantified in specific numbers of replicates (*N*) with our standard shotgun method, with 'matching between runs' (MBR) and with BoxCar in conjunction with a matching library. (**b**) Same for quantified protein groups. (**c**) Completeness of the peptide quantification matrix as a function of descending peptide abundance in all three experiments. (**d**) Same for quantified protein groups.
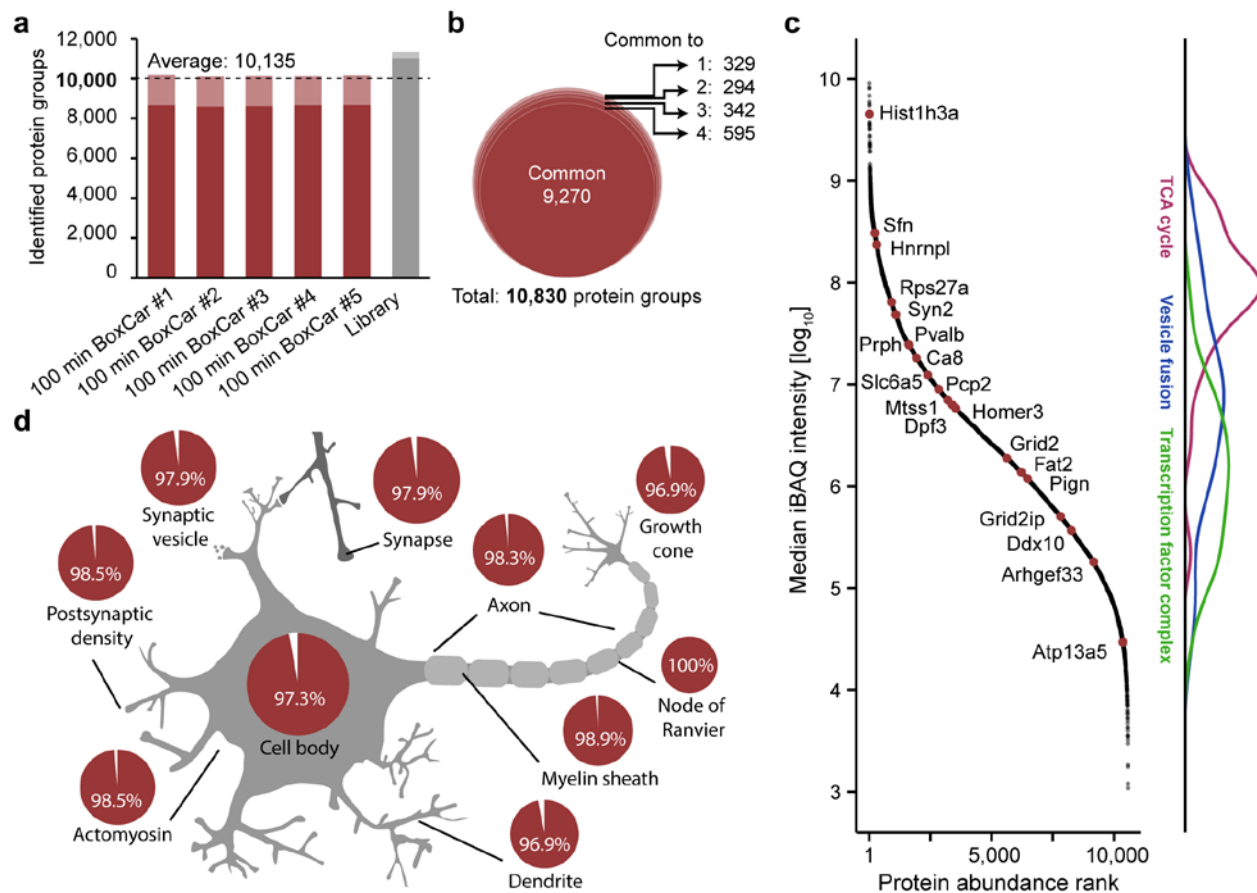
Figure 6. In-depth proteome analysis of rodent cerebellum in 100 min BoxCar single runs. (**a**) Number of proteins identified in each replicate and in 24 high-pH fractions. Bright bars indicate proteins identified by a single peptide. (**b**) A total of 10,850 protein groups were identified in 100 min BoxCar single runs and 97% of the proteins were identified in at least two replicates. (**c**) Ranking of cerebellum proteins by their median, normalized intensity in BoxCar single runs. Differentially expressed brain proteins[36] are highlighted in red and the density distribution of key metabolic and neuronal Gene Ontology (GO) annotations is projected onto the right axis. (**d**) Relative coverage of GO cellular compartments in BoxCar single runs as compared with the full library.

ONLINE METHODS

Cell culture

Human malignant epithelial cells (HeLa S3, ATCC) were cultured in Dulbecco's modified Eagle's medium (DMEM), fortified with 10% fetal bovine serum, 20 mM glutamine and 1% penicillin-streptomycin (all PAA Laboratories, Germany). For the SILAC experiment, cells were cultured in arginine- and lysine-free DMEM and the medium was fortified with equal concentrations of either natural L-arginine and L-lysine (light) or their stable-isotope labeled counterparts (Arg10, $[^{13}C_6{}^{15}N_4]$arginine, and Lys8, $[^{13}C_6{}^{15}N_2]$lysine; both Cambridge Isotope Laboratories, USA) following our standard protocol[41]. Escherichia coli (strain: XL1 blue) was grown in LB medium at 37°C until log phase and harvested at OD 0.5. Cells were collected by centrifugation at 200 x g for 10 min and washed once with cold phosphate buffered saline (PBS) before re-centrifugation. Cell pellets were flash frozen in liquid nitrogen and stored at -80°C until needed.

HeLa and E. coli sample preparation

HeLa cells were lysed, reduced and alkylated in SDC buffer with chloroacetamide (PreOmics GmbH, Germany) as previously described[30]. Briefly, the samples were boiled at 95°C for 10 min, followed by 15 min sonication at maximum power (Bioruptor, Diagenode, Belgium) to disrupt cells and shear nucleic acids. Protein digestion was performed overnight at 37°C by adding equal amounts of the proteolytic enzymes LysC and trypsin in a 1:100 (w/w) ratio to the sample. The reaction was stopped with five volumes of isopropanol/1% trifluoroacetic acid (TFA).

Mouse cerebellum sample preparation

All animal experiments were performed according to the institutional regulations of the Max Planck Institute of Biochemistry, and approved by the government agencies of Upper Bavaria. The cerebellum was dissected from an individual male mouse (strain: C57BL/6) at postnatal day 5 and snap frozen in liquid nitrogen. The frozen cerebellum was ground using a mortar and pistil

and suspended in 100 µL SDC reduction and alkylation buffer. Further sample processing and enzymatic digestion was performed as described above.

Plasma sample preparation

Blood sampling complied with all relevant ethical regulations, and was approved by the ethics committee of the Max Planck Society (24th February 2015, Heidelberg, Germany). Donors provided written informed consent. The human plasma samples from a single healthy donor was prepared as previously reported[21]. In brief, SDC reduction and alkylation buffer was added to blood plasma in 25-fold excess (v/v) and the mixture was boiled at 95°C for 10 min. The samples were digested for 1 h at 37°C with LysC/trypsin in a 1:100 ratio (w/w) before quenching the digestion with isopropanol/1% TFA.

Peptide purification for mass spectrometry

Following our previously published standard protocol (ref. 30), acidified protein digests were loaded onto a styrenedivinylbenzene-reversed phase sulfonate (SDB-RPS) sorbent. Following two consecutive purification and desalting steps with isopropanol/1% TFA and ddH$_2$O/0.1% TFA, the peptides were eluted with 80% ACN/1% ammonia. The eluate was vacuum-centrifuged to dryness (SpeedVac, Eppendorf, Germany) and re-constituted in 2% acetonitrile/0.1% TFA in ddH$_2$O for single run analyses or reversed-phase chromatography at basic pH.

High pH reversed-phase fractionation

Peptide pre-fractionation was performed on a fully automated 'spider fractionator' (PreOmics GmbH, Germany) coupled online to an EASY-nLC1000 chromatography system (Thermo Fisher Scientific, Germany). A detailed description of the setup is available in ref. 27. Approximately 60 µg of purified peptides were loaded onto a 30 cm reversed-phase column with an inner diameter of 250 µm and separated at a flow rate of 2 µL/min with a binary buffer system (buffers A and B) at pH 10 (all PreOmics GmbH). The gradient started from 3% B and was stepwise increased to 30% B within 45 min and to 60% within 17 min, and then further increased to 95% B within 5 min, held constant for 3 min and then decreased to 3% B within 10 min for re-equilibration. For each sample, 24 fractions were automatically concatenated and collected into 0.2 mL tubes by shifting the rotor valve every 90 s. The pooled fractions were vacuum-

centrifuged to dryness subjected to LC-MS/MS analysis after re-constitution in 2% acetonitrile/0.1% TFA in ddH$_2$O without any further purification.

Liquid chromatography and mass spectrometry

Nano-flow reversed-phase liquid chromatography was performed on an EASY-nLC 1000 or EASY-nLC 1200 ultra-high pressure system coupled online to a Q Exactive HF mass spectrometer equipped with a nano-electrospray ion source (all Thermo Fisher Scientific). Approximately 1-2 µg purified peptides were loaded on a 40 cm column (75 µm inner diameter) with a pulled emitter, packed in-house with ReproSil-Pur C$_{18}$-AQ 1.9 µm resin (Dr. Maisch GmbH, Germany). To reduce the backpressure at high flow rates and enhance separation efficiency the column compartment was kept at 60 °C. For 45 min gradients, mobile phases A and B were 100/0.1% water/formic acid (v/v) and 60/40/0.1% acetonitrile/water/formic acid (v/v/v). Peptides were separated at a constant flow rate of 450 nL/min with a linear gradient of 3-40% B within 35 min, followed by a linear increase from 40-90% B within 8 min and a 2 min plateau before re-equilibration. For 100 min gradients, mobile phase A was as above and mobile phase B was 80/20/0.1% acetonitrile/water/formic acid (v/v/v). Peptide separation was performed at a constant flow rate of 350 nL/min with a binary gradient of 2-5% B for the first 2 min, followed by a linear increase to 35% B over 93 min and column washing for a total of 5 min with a maximum of 98% B before re-equilibration.

Standard shotgun LC-MS experiments were performed with a data-dependent top15 method. Full MS scans were acquired from *m/z* 300-1,650 at a resolution of 60,000 at *m/z* 200 with an AGC target of 3e6 charges. For higher energy collisional dissociation MS/MS scans, the normalized collision energy was set to 27 eV and the resolution to 15,000 at *m/z* 200. Precursor ions were isolated in a 1.4 or 1.5 Th window and accumulated for a maximum of 25 ms or until the AGC target of 1e5 ions was reached. Precursors with unassigned charge states, charge 1+ or charge 6+ and higher were excluded from sequencing. Previously targeted precursors were dynamically excluded from re-sequencing for 20 s.

Implementation of BoxCar scans on a quadrupole Orbitrap mass spectrometer

The Q Exactive HF mass spectrometer was operated with the instrument control software Tune v2.6 (build 2640, Thermo Fisher Scientific). Customized MS acquisition methods were built in the 'text method interface' of Xcalibur (v3.1.66.10, Thermo Fisher Scientific), which was accessed through 'XmlMode' as provided by Thermo Fisher Scientific. A detailed step-by-step procedure is available in the **Supporting Protocol**. We implemented BoxCar scans using the built-in multiple ion injection method with the normalized collision energy set to zero. The scan sequence for a single BoxCar scan was as follows: (1) the mass range of the first 'box' was selected in the analytical quadrupole and the ions were accumulated in the C-trap until either the AGC target value or the maximum ion injection time was reached (2) the selection window of the analytical quadrupole was switched to the next 'box' and the ions were accumulated as before (3) step 2 was repeated as often as required (4) the combined ion population was injected into the Orbitrap cell and mass analyzed in a single transient. Within one LC-MS experiment, the number of BoxCar scans as well as the number and dimensions of all boxes per BoxCar scan were fixed. For each BoxCar scan, the total AGC target value was adjusted below 1e6 to not exceed the space charge limit of the C-trap and evenly distributed across all boxes. Likewise, the maximum total ion injection time given by the Orbitrap transient time was evenly allotted to all boxes. The overall acquisition cycle comprised a standard full scan, two or more BoxCar scans covering a range of *m/z* 400-1,200 and five data-dependent MS/MS scans. Suitable precursor ions were selected from the preceding full scan, applying the same criteria and settings as in the standard shotgun experiments.

Investigation of BoxCar scan parameters

To assess the effect of different BoxCar acquisition parameters on the number of detectable MS1 features we utilized a statistical 'Design of Experiment' (DoE) approach. The DoE was designed and analyzed with the MODDE software package version 10.0.0 (Umetrics, Sweden). We varied the number of BoxCar scans per acquisition cycle, the number of 'boxes' per Boxcar scan, the maximum ion injection time per scan, and the mass resolution. We chose a D-optimal design with 33 design runs and 5 center points (**Supplementary Tab. 1**). All experiments were run with

45 min gradients of 1 µg peptides from a HeLa digest in randomized order and the number of detected isotope patterns (m/z 400-1200, z>1) was extracted from the MaxQuant output tables. In total, we included N=37 data points to fit a quadratic model with multiple linear regression. R^2 and Q^2 values >0.978 indicated a valid model with excellent predictive power and a very high reproducibility (0.995). The results are represented as 4D response contour curves (**Supplementary Fig. 1**).

Analysis of HeLa proteomes in single runs

Single runs of 1 µg HeLa digest were measured with the 45 min gradient using both, the standard shotgun method and BoxCar. BoxCar runs were acquired at a MS1 resolution of 120,000 at m/z 200 and the acquisition cycle included two BoxCar scans à 12 boxes each with a 1 Th overlap at the edges. The maximum ion injection time for each box was set to 21.3 ms to reach an AGC target of 16,666 ions. The width of the individual boxes scaled inversely proportional to the m/z density of tryptic peptides (**Supplementary Tab. 2**). To assess the quantitative accuracy of BoxCar, we mixed HeLa and E.coli digests as well as heavy and light SILAC HeLa digests as specified in the main text and analyzed in each case 1 µg of the peptide mixtures with the 100 min gradient. The BoxCar parameters were set as above, but with three BoxCar scans à 12 boxes per acquisition cycle and the AGC target set so 83,333 ions per box.

Analysis of human plasma digests in single runs

Purified peptides from human plasma were acquired with a standard shotgun method and BoxCar for comparison, using the 45 min gradient. The MS1 resolution was set to 120,000 at m/z 200 for BoxCar and the acquisition cycle comprised two BoxCar scans à 12 boxes (scaled width, 1 Th overlap) with a maximum ion injection time of 21.3 per box with the individual AGC target set to 25,000.

Library generation and deep proteome coverage in single runs

To increase the proteome coverage in single runs of HeLa digest, we first generated a peptide library from two separate high-pH reversed-phase fractionations as described above and measured each fraction with the 45 min gradient using our standard shotgun acquisition method. The library raw files were analyzed together with the single run HeLa dataset from above.

To generate a deep library of cerebellum peptides we fractionated the purified peptides five times and measured each fraction with the 100 min gradient using standard acquisition parameters. Then, we acquired quintuplicate single runs of 1 µg digested proteins with our standard shotgun method as well as BoxCar. The BoxCar parameters were set as above for the 100 min gradient.

Adaptation of MaxQuant for BoxCar scans

MaxQuant contains a new data analysis mode for processing BoxCar scans, which assembles the separate BoxCar scans and the preceding full scan into a single high dynamic range scan. For this purpose the BoxCar scans as well as the full scans are first transformed to a common m/z grid. This is facilitated by having the same mass resolution in each of these scans. The *m/z*-dependent bin size of the common *m/z* grid is chosen to be the local mean of the bin sizes of the full scan and the BoxCar scans in the raw data. Note that the typical spacing of the raw data points in 'profile mode' is about ten times more finely grained than the mass resolution. The intensities on the new *m/z* grid with common binning are calculated from the original scans as linear interpolations from the closest *m/z* values above and below the *m/z* value on the new grid. Next, BoxCar scans for each range combination as well as the full scans are summed up over the whole retention time range. The summed full scan as well as the summed BoxCar scans allow us to calculate a transmission function for each BoxCar range as follows.

Since the full scan summed over the whole LC-MS run, as well as the summed BoxCar scans have the same common *m/z* grid, we can divide intensity values point wise. The transmission function for BoxCar scan *j* is the point-wise ratio between the intensities of the full scan and the BoxCar scan:

$\text{transmission}_j(m/z) = \text{boxcar\_intensity}_j(m/z) / \text{full\_scan\_intensity}(m/z).$

These transmission functions are used to calculate a single high dynamic range scan, by using the transmission function as a weight for a weighted average of the full scan and all BoxCar scans from one acquisition cycle. Importantly, the algorithm does not require any user input but rather adapts to the experimental design, preserving full flexibility in choosing the scan range for full scans and BoxCar scans, the number of BoxCar scans as well as the number of boxes per scan. All subsequent parts of the MaxQuant computational workflow take these scans as input as a replacement for the commonly used MS1 survey scans.

Data analysis

MS Raw files were processed with MaxQuant version 1.5.3.55 or 1.6.0.15 and the extracted MS/MS spectra were matched by the Andromeda search engine against tryptic peptides (max. 2 missed cleavages) derived from human or mouse reference proteomes (Uniprot, version 2016/05) and a list of common contaminants at a false discovery rate (FDR) of 1% on peptide spectrum match and protein levels. The search included cysteine carbamidomethylation as fixed modification and methionine oxidation, as well as acetylation of the protein N-terminus as variable modifications. The required minimum peptide length was 7 amino acids and the maximum peptide mass was limited to 4600 Da. Maximum mass tolerances were 4.5 ppm for precursor ions after non-linear recalibration and 20 ppm for fragment ions. MaxLFQ[42] was performed separately in parameter groups with a minimum ratio count of 1. If applicable, peptide identifications were transferred between samples by 'match between runs' within a 0.3 min (45 min gradients) or 0.5 min (100 min gradients) window after retention time alignment. The matching type was set to 'matching from' for library runs and 'matching from and to' for single runs. To minimize the potential for false matching (**Supplementary Note 1**), peptide libraries were generated from the very same cell culture batch or specimen and, if applicable, processed in parallel with the samples for single run experiments and analyzed using the same chromatography method. Identifications in the library were stringently filtered for a FDR <1% on the peptide spectrum match and protein group level, which implies an even lower FDR on the peptide level. To estimate the rate of potentially false matches in the current study, we devised a decoy model on the level of MS-features, by offsetting masses by 40 ppm and permuting all retention times with a minimal shift of 180 s. This resulted in a conservative FDR estimate of < 5% at the matched feature level.

MS metadata, such as ion injection times, were accessed and extracted from raw files with the MSFileReader (v3.0.31, Thermo Fisher Scientific). Further downstream analysis of the MaxQuant output tables and metadata was performed in either Perseus[43] (v1.6.0.8) or the R statistical computing environment[44] (v3.2.1). Copy numbers and absolute protein abundances were estimated with the 'Proteomic Ruler' as previously described[45]. Decoy database hits were strictly excluded from the analysis as well as proteins that were identified as potential contaminants or identified exclusively by one site modification.

Experimental Design and Statistics

A summary of the experimental design is enclosed in the Life Sciences Reporting Summary. To assess technical differences between BoxCar and shotgun methods, purified peptides were pooled before LC-MS analysis and the methods were run in alternating order (ABAB). Replicates refer to repeated LC-MS analyses (technical replicates). The exact N numbers for statistical tests are reported in the figure legends where applicable. To calculate coefficients of variation, missing values (displayed as 0 or *NaN* in the MaxQuant output) were removed prior to the analysis. Pearson correlation coefficients were used as a measure for linear correlation between logarithmized intensity values from two experiments. In Supplementary Fig. S5, separately for each acquisition method, the data was filtered for protein groups with no missing values in both conditions and a one-sided two-sample t-test was performed (*N*=3 replicates per group). To correct for multiple hypothesis testing, t-test significant hits were truncated at a permutation-based false discovery rate threshold of 0.05 (250 randomizations).

Data availability

The datasets generated during and analyzed during the current study have been deposited to the ProteomeXchange Consortium via the PRIDE[46] partner repository with the dataset identifier PXD006109.

Methods-only references

41. Ong, S.-E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1,** 2650–2660 (2007).

42. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513-2526 (2014).

43. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–40 (2016).

44. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2008)

45. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A 'proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* **13,** 3497–506 (2014).

46. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447-56 (2016).