Robust and easy-to-use one pot workflow for label free single cell proteomics

Manuel Matzinger[1,*,§], Elisabeth Müller[1,*], Gerhard Dürnberger[3], Peter Pichler[1], Karl Mechtler[1,2,3,§]

[1]Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Vienna, Austria

[2]Institute of Molecular Biotechnology, Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna, Austria

[3]Gregor Mendel Institute of Molecular Plant Biology (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna, Austria.

* These authors contributed equally

§Correspondence to: manuel.matzinger@imp.ac.at, karl.mechtler@imp.ac.at

ABSTRACT:

The analysis of ultra-low input samples or individual cells is critical for answering a multitude of biomedical questions. But current proteomic workflows are limited in their sensitivity and reproducibility. Here we report a complete workflow that includes optimized strategies for cell lysis to data analysis. Thanks to the use of a convenient to handle volume of 1 µL and standardized 384 well plates, the workflow is easy to implement, even for newcomers to the field. At the same time, the workflow can be performed in a semi-automated fashion using the CellenONE® robot, which allows for high reproducibility. Ultrashort gradient lengths ranging to little as 5 min for highest possible throughput were tested with advanced µ-pillar columns. DDA, WWA and DIA as well as commonly used and advanced data analysis algorithms were benchmarked. From a single cell, 1790 proteins were identified using DDA and a dynamic range of 4 orders of magnitude was covered. Using DIA, the proteome coverage was increased to >2200 proteins from single cell level input within a 20 min active gradient. We applied our workflow to differentiate two cell lines showing its ability to investigate cellular heterogeneity as proof of principle.

Keywords: Single Cell Proteomics, one-pot workflow, label-free, WWA, DIA, µPAC, CellenONE®

INTRODUCTION

Investigating individual cells using advanced technologies for cell isolation and analysis of their nucleotide content have already revolutionized the depth of knowledge in cellular development and behavior.[1–

[4] Importantly, cellular identity is driven by its proteome, which is why proteome profiling has become increasingly popular over the last few years in complementing research on molecular cell biology. The importance of reaching single cell resolution is highlighted by the fact that already subtle changes in protein expression and turnover may lead to changes in cellular behavior and functionality. Even clonally identical cells differ from each other in dependence of their environment and age. Understanding cellular heterogeneity is therefore of significant importance to understand regulatory mechanisms or (stem-)cell development.

To provide a comprehensive and quantitative proteome profile by classical mass spectrometry techniques requires the bulk analysis from tens of thousands to millions of cells.[5] Supplemental biochemical approaches such as flow cytometry, immunoprecipitations or immunofluorescence-based techniques help to resolve intercellular networks but are limited in specificity as well as to a defined number of some dozen proteins.

Single cell proteomics aims to fill this gap by resolving cellular heterogeneity on a global and unbiased level.[6] There is no doubt, that mass spectrometry is a technique sensitive enough to perform such a task since detection limits go down to the attomolar sample level. The median number of protein copies in an individual mammalian cell is estimated to be 18,000[7] meaning that the identification of at least the more abundant fraction of proteins within a cell by mass spectrometry is a task promising to succeed. Reaching sufficient proteome depth is however still challenging as current workflows face several limitations:

- Sample losses: During sample preparation the manipulation of cellular content as well as its transfer into a mass spectrometer sample material is lost upon contact to surfaces or during purification steps, which leads to significant lowered sensitivity and reproducibility of the workflow. Recently developed techniques as SCoPE2[8], nanoPOTS[9,10] or the proteoCHIP[11] minimize sample losses by minimizing handling volume and automatizing sample handling but require specialized equipment.
- Sensitivity: Sample separation and detection of ions within the mass spectrometer has to be tightly finetuned but still lacks the required sensitivity.
- Throughput: To enable proper statistics hundreds to thousands of cells need to be analyzed. To boost sensitivity as well as reaching high throughput, multiplexed workflows and carrier proteomes are commonly used in the single cell proteomic field.[8,10–13] This comes with adverse effects such as lowered quantitative accuracy and dynamic range compared to label free approaches and potentially biases the type of peptides identified by the choice of the carrier proteome.[14–16]

Here we present a workflow that includes a complete bundle of optimized steps from sample preparation to data analysis. We aim to address the above-mentioned limitations by working in standardized 384 well plates which makes the workflow adoptable for a broader community without losing sensitivity. Further, we shortened the gradient time without losing separation power to increase throughput also for more robust label-free strategies. Finally, we combine clever data acquisition strategies with AI-based data analysis to boost identification numbers and dynamic range.

RESULTS

Sample workflow improvements and automatization.

Reproducibility and recovery

Aiming for full workflow automatization, ideal cell isolation parameters are to be selected. Using the cellenONE® robot two parameters, diameter, and elongation, are relevant. We optimized these parameters for selection of individual HeLa cells, which were used throughout this study. In order to get a first estimation regarding needed parameters, we first mapped distribution of all detected particles. In line with our expectation[17], the majority of cells had a diameter ranging from 15 to 30 µm. As a starting point, we chose the standard parameters used in our lab before for the isolation of HeLa cells. We isolated 384 single cells into an appropriate plate and distinguished four different cases when visually analyzing pictures of each isolated cell: Only one cell was isolated per well, more than one cell was isolated per well, the isolated cell appears apoptotic, or no cell was isolated at all (Figure 1A). Supplemental Figure 1shows representative examples for each of the four cases. Overall, only 88.8% of all wells confidently contained exactly one intact cell. Reducing the diameter range and elongation to a lower value ensured that multiple attached cells are not falsely recognized as single cells and that only intact living cells are isolated by a stricter elongation parameter. The fraction of correctly isolated individual and intact cells was eventually increased to 99.7%, which allows for automation of cell isolation without the need to manually validate pictures of each isolated particle. Cells were additionally Hoechst 33342 stained to ensure exclusive selection of viable cells.

For label free single cell sample preparation, a protocol based on the work of Ctortecka et. al.[11] was adopted. 1µL of master mix containing the detergent DDM for lysis, the enzyme trypsin for digestion and TEAB buffer was predispensed to each well of a 384 well plate followed by isolation of individual HeLa cells into the master mix (for details, see methods section). The 384 well plate was then placed inside the

94 cellenONE® for 2 hours at 50°C and 85% relative humidity for lysis and digestion. Since sample transfers
95 are prone to loss of hydrophobic peptides, transfer to different tube materials for storage was compared
96 to sample storage and injection directly from the same 384 well plate also used for sample preparation
97 (Figure 1B). Both, transfer to a PEG coated PCR tube, as described by Ctortecka et. al.[11], as well as
98 transfer to a glass vial lead to obvious losses compared to omitting any sample transfer. In conclusion, all
99 future single cell samples were processed in and injected from a 384 well plate. Next, the effect of the
100 used protease (Figure 1C) as well as of multiple protease addition (Figure 1D) were assessed. In detail,
101 Trypsin Platinum and Gold (both from Promega) were compared. According to the manufacturer, Trypsin
102 Platinum is free of any detectable nonspecific proteolytic activity which is why an improved overall di-
103 gestion efficiency was to be expected. Surprisingly Trypsin Gold delivered slightly better results in our
104 single cell setting in terms of protein identifications as well as in terms of observed missed cleavages. As
105 the tryptic activity is highest in the first 30 min of digestion, its activity was boosted by adding trypsin
106 again after 30 min, which significantly increased the number of identified proteins. Additionally, we tested
107 the protease enhancer ProteaseMAX™ from Promega. On average, we were able to achieve a 17% in-
108 crease in the proteins found in the single cell samples by adding the enhancer to our trypsin digestion
109 (Supplemental Figure 4). All future experiments were therefore performed by adding the enhancer Prote-
110 aseMAX™ in the master mix and Trypsin Gold twice.

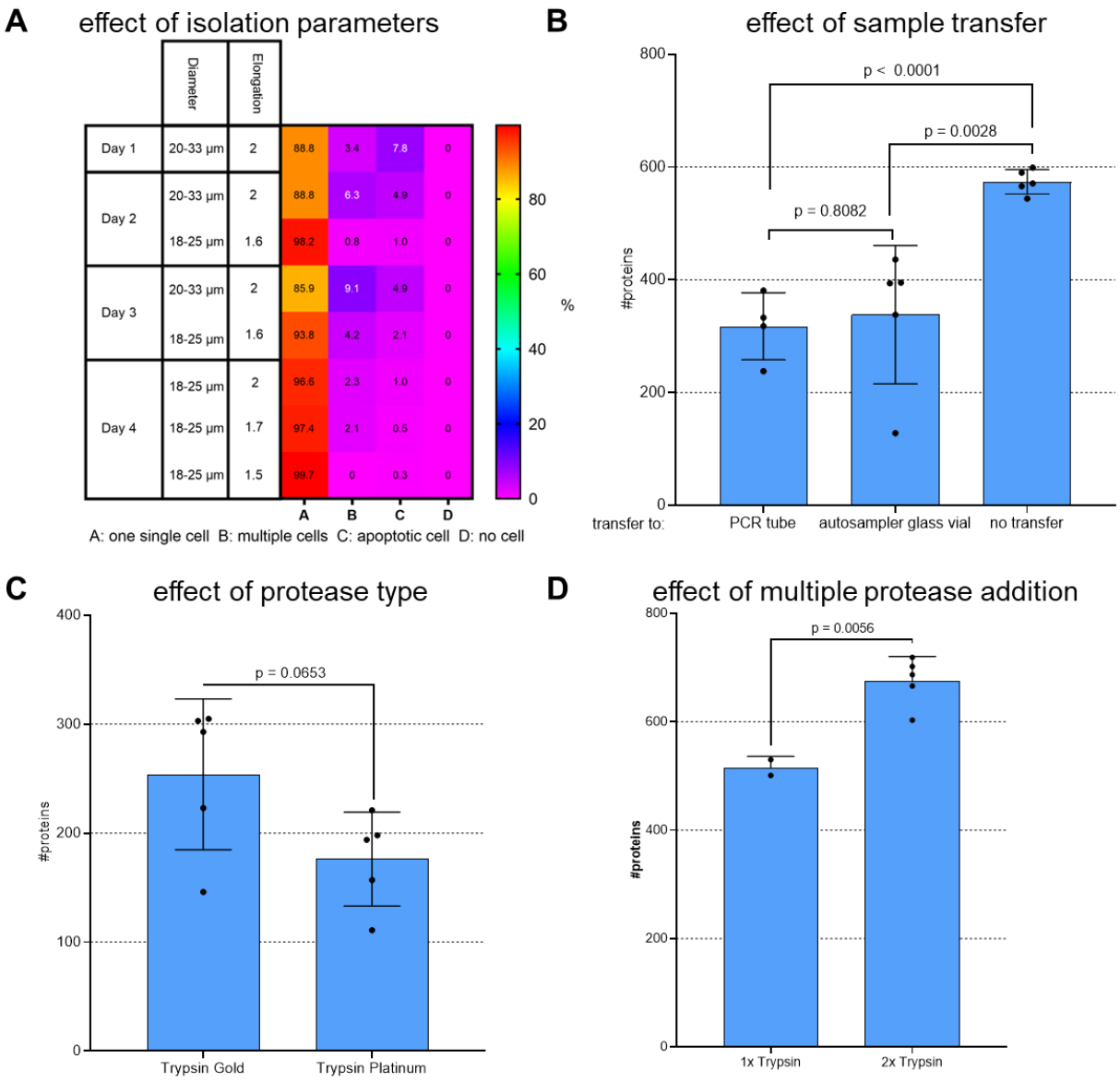


Figure 1: ***Improving recovery and reproducability by tuning workflow parameters****. Individual HeLa cells were isolated into predispensed mastermix using the CellenONE® (**A**): The diameter and elongation parameters were varied on the CellenONE®, numbers indicate the relative amount (%) of cells isolated in the category as indicated, n=384 cells on 4 individual days each. (**B**): Individual HeLa cells were lysed and digested in a 384 well plate and the sample was either injected from there to the LC-MS system or transferred as indicated. (**C**): Cells were digested with Trypsin Gold or Platinum as indicated. (**D**): trypsin was added either once within the mastermix or a second time by addition of 500 nL additional fresh 3 ng/µL trypsin after 30 min. (**B-D**): Bars represent protein identifications, error bars show standard deviations, an unpaired two tailed Students t-test was performed and resulting p values are depicted above each compared data set, n >4 replicates.*

Optimization of sample preparation to boost protein numbers

Using a sample handling volume of 1 µL rather than working in the nanoliter range makes sample handling much more convenient but it also increases the wetted surface area and in conclusion the chance

124 that peptides stick at the tube walls. With this in mind, we aimed to keep peptides solubilized by supple-
125 mentation of 5% DMSO. This was first tested on diluted bulk digests at a concentration of 250 pg / 3.5
126 µL, which is the same as expected under ideal conditions in our final real single cell samples. The presence
127 of DMSO reproducibly and significantly increased recovery as visualized by more protein IDs obtained
128 (Figure 2A). Those peptides additionally found in the DMSO containing sample are eluting predominantly
129 towards the end of our gradient (Figure 2B). This supports the ideas of an improved solubility of hydro-
130 phobic peptides thanks to DMSO supplementation. The vast majority of 80 % of all PSMs found in a
131 representative replicate without DMSO was rediscovered also in the presence of DMSO (Figure 2C),
132 while those not commonly found might be reasoned by run-to-run variability of the stochastic DDA
133 method used here.

134 To minimize adsorptive sample losses also during single cell sample preparation we further aimed to
135 hinder sample drying by automated addition of water during digestion at 50 °C within the CellenONE®.
136 By that the digestion mix was kept hydrated over a period from 0 - 120 min until the end of incubation.
137 In case of no hydration task performed, the sample was reproducibly dried after ~30 min, which is why
138 500 nL water were added every 15 min to keep the volume at the same level. Our results suggest, that an
139 elongated hydration allows for longer active digestion and lowered sample adsorptive sample loss as in-
140 dicated by more proteins identified when hydrating at least over the first 90 min of the digestion process
141 (Figure 2D). Analyzing single cells that were either kept hydrated during digestion or supplemented with
142 DMSO for storage lead to roughly 500 protein IDs. Combining both boosting strategies results in more
143 than 1000 average proteins identified from a single HeLa cell using DDA and analyzing data with
144 CHIMERYS™ on 1 % FDR at peptide and protein level (Figure 2E). This effect is even more distinct on
145 peptide level, where an average of 1883 peptides was found without DMSO of hydration compared to
146 4625 with hydration and DMSO, which corresponds to a boost by 146 %. Of note, the addition of DMSO
147 yields in identification of 550 additional proteins while 93 % of all proteins previously identified in the
148 absence of DMSO were rediscovered (Figure 2F).

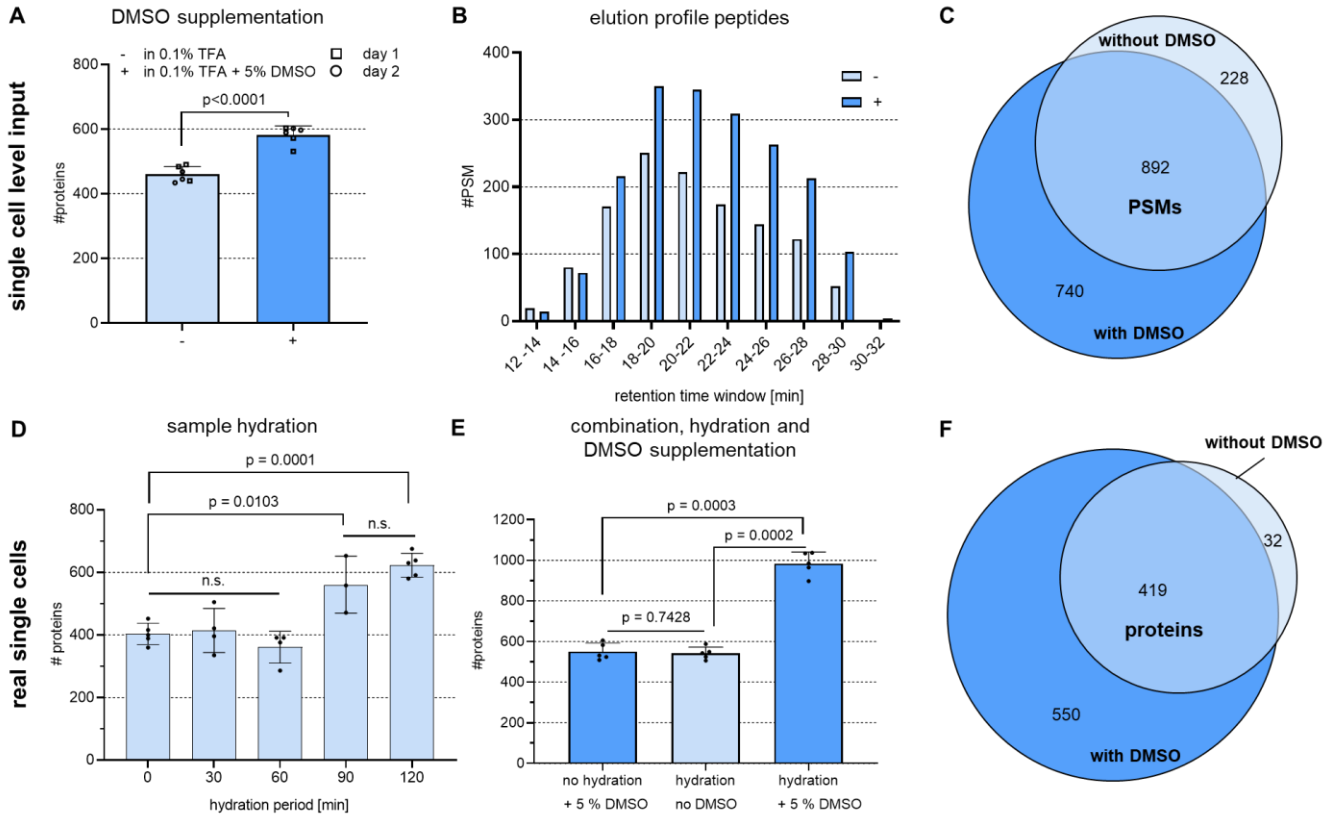


Figure 2: ***DMSO supplementation and keeping the sample hydrated improves recovery of hydrophobic peptides and boosts ID numbers****. All data was acquired using DDA and analyzed using CHIMERYS™ at 1 % FDR on PSM and protein level. (**A**):250 pg HeLa bulk digest dissolved at a concentration of 71 ng/µL in the presence or absence of 5 % DMSO as indicated was used for injection into LC-MS each, bars represent protein IDs, error bars indicate standard deviations. To estimate significance, an unpaired two tailed Students t-test was performed and resulting p values are depicted above each compared data set. n = 6 technical replicates measured on 2 different days. (**B**): Data from one representative replicate of A, bars indicate PSM matches in the indicated retention time windows. (**C**): Venn diagram showing common PSM matches from IDs shown in panel B. (**D**): Individual HeLa cells were lysed and digested at 50°C for 2h maintaining the sample hydrated by automated water supplementation for the indicated period. Bars represent protein IDs, error bars indicate standard deviations, a one-way ANOVA with Tukey's multiple comparison post-test was performed, n ≥ 3 replicates. (**E**): Individual HeLa cells were kept hydrated for 2h or dried upon evaporation during digestion in the presence or absence of DMSO in the final storage solution as indicated. Bars represent protein IDs, error bars indicate standard deviations, an unpaired two tailed Students t-test was performed and resulting p values are depicted above each compared data set, n = 5 replicates (**F**): Venn diagram showing common protein IDs from one representative replicate comparing the condition hydration, no DMSO with hydration + 5% DMSO of data from panel D.*

Workflow key points

To ensure a robust, reproducible, and automated single cell workflow, single cells isolation was performed using the picolitre dispensing robot cellenONE®. We aim that this workflow is implementable by labs without access to expensive specialized equipment for work with ultralow volumes. This is ensured by

working in a 384 well plate and within an easy to handle 1 µL volume. Constant hydration is automatically performed by the cellenONE® but can be replaced by manual pipetting as well. Our optimized workflow is sketched in Figure 3 and detailed within the methods section.

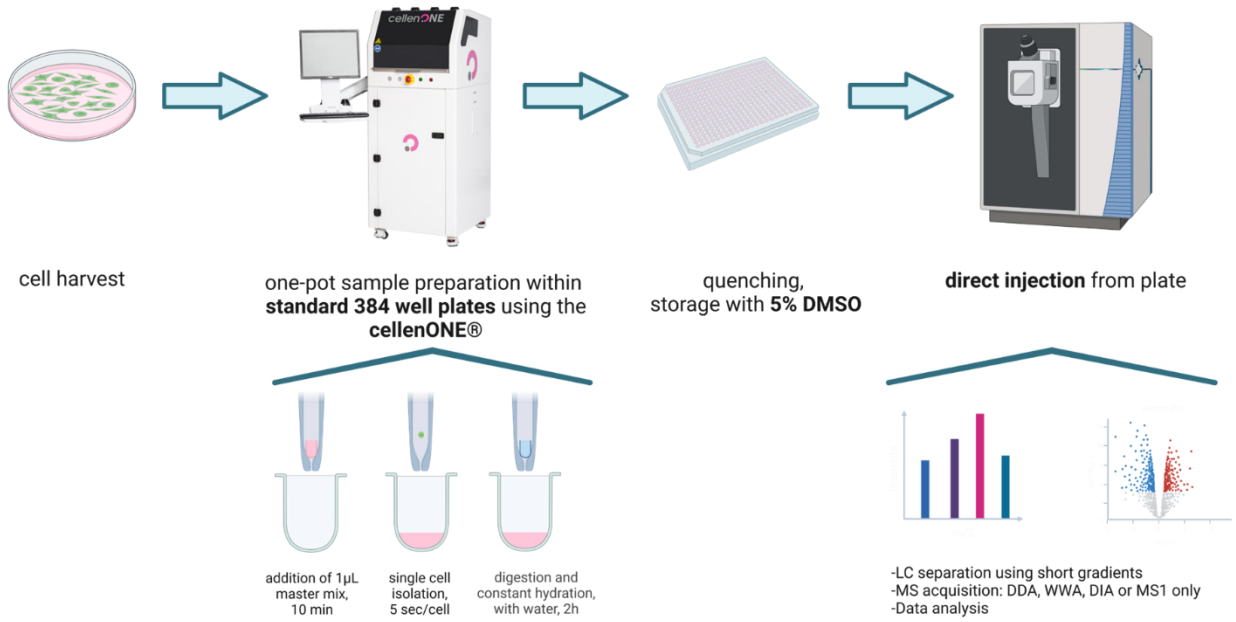


Figure 3: ***Overview label free, one-pot, single cell sample processing workflow.*** *Cells are harvested followed by lysis and digestion within the CellenONE® in a 384 well plate that is also used for direct placement into the autosampler of the HPLC-MS. Figure created with Biorender.*

Optimizing data acquisition and analysis to maximize ID numbers.

In parallel to parameter tuning for a loss free and robust single cell sample preparation, we aimed to improve proteome coverage by optimizing data acquisition and analysis as well. In a first step we compared two column types for their chromatographic performance: A classical packed bed column (CSH C18 Column, 130Å, 1.7 µm, 75 µm X 250 mm, waters) that was previously used in our lab and a µPAC™ column (brick shape pillars, 5.5 cm, prototype column, Thermo Fisher) which we recently successfully tested for low input amounts as well as short gradients.[18] Although the µPAC™ column is only 5.5 cm long, its internal flow part is roughly 50 cm leading to a sufficient surface area enabling powerful separation. The median FWHM peak width obtained using the packed bed column was 3.84 sec vs 4.77 sec for the µPAC™. Although peak withs seem slightly higher, the µPAC™ column clearly and reproducibly

189 outperformed its conventional counterpart in ID numbers (Figure 4C), which is why we used it for fol-
190 lowing optimizations as well.

191 Next, we minimized needed runtime as using a short gradient time has two potential advantages: Increased
192 throughput and improved sensitivity thanks to sharper peaks. Unfortunately, shortening an already fast
193 20 min active gradient down to 5 min significantly lowers resulting protein IDs by ~ 30% from 705 to
194 485 proteins (Figure 4A). Resulting peaks are sharp with a median FWHM of only 4.77 and 2.9 sec for
195 the 20 and 5 min gradient respectively which is a prerequisite for high sensitivity. However, especially
196 the 5 min method seems already too fast for the Orbitrap to allow recording enough MS2 spectra which
197 eventually lowers ID numbers. The utilization of short gradients is however important to improve through-
198 put. To combine the need for speed with sufficient proteome coverage, duty cycle time can be reduced by
199 omitting the requirement for any fragment spectra. A first successful attempt to do proteome wide anal-
200 yses from very short gradients of only 5 min was done by Ivanov and coworkers in 2020[19,20]. We adopted
201 their approach for the single cell input level. For these experiments we used the 5.5 cm µPAC™ column
202 that allows for fast loading and re-equilibration with 1 µL/min thanks to its low backpressure leading to
203 a total run time of 7.4 min. This MS1-only strategy significantly outperforms our standard DDA approach
204 for short gradients in identifying roughly the same number of proteins within a 5 min active gradient that
205 was previously identified in a 20 min active gradient (Figure 4B). Of those proteins found in the 20 min
206 DDA run, ~70 % were identified in the 5 min MS1 run as well, which is a comparable overlap as published
207 for higher input samples[19] (Supplemental Figure 2).

208 For the longer 20 min active gradient we furthermore benchmarked different acquisition strategies (Figure
209 4D). We showed that the proteome depth of our original DDA method, which uses a precursor isolation
210 window of 2 m/z, can be improved by widening up the isolation window to 12 m/z. Using such a wide
211 window acquisition (WWA[18]) on purpose generates chimeric spectra which can however be identified by
212 the AI based search engine CHIMERYS™ leading to identification of more than one peptide from a single
213 spectrum on average. With DIA we go even a step further and found that variable windows of up to 25
214 m/z can significantly increase identification numbers to >1,100 proteins from 250 pg HeLa.

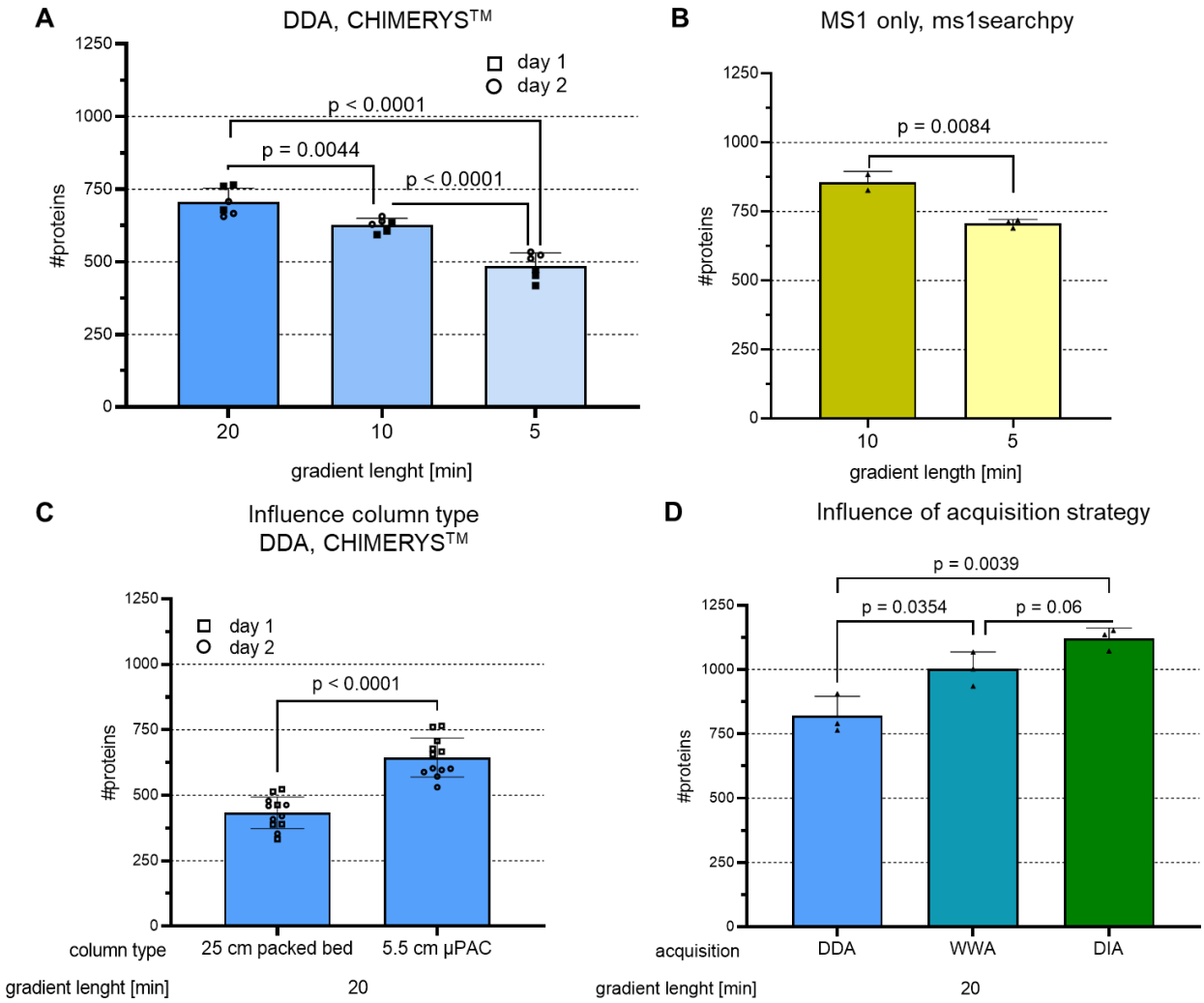
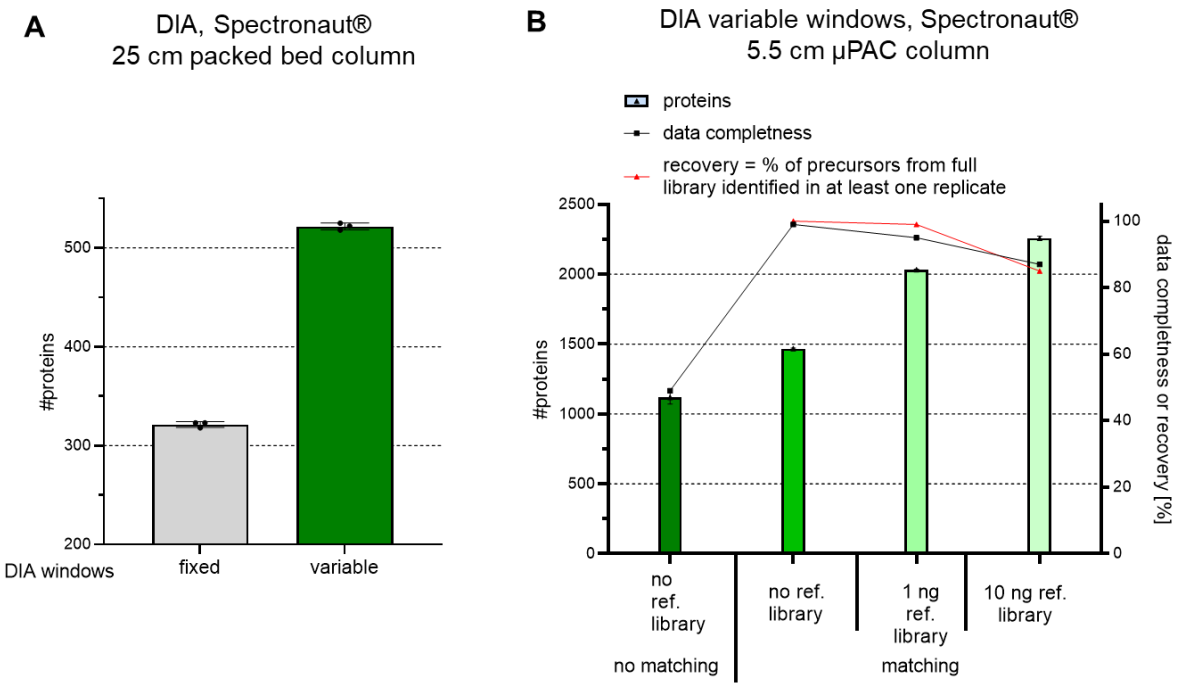


Figure 4: ***Influence of gradient length, column type and data acquisition strategy on protein IDs of single cell level inputs****. 250 pg HeLa bulk digest was used for injection into LC-MS each, bars represent protein IDs on 1 % FDR level, error bars indicate standard deviations. To estimate significance, an unpaired two tailed Students t-test was performed and resulting p values are depicted above each compared data set. (**A**): Peptides were separated on a 5.5 cm µPAC™ column, data was acquired using DDA and analyzed using CHIMERYS™, n = 6 technical replicates measured on 2 different days. (**B**): As A but data analysis was performed using ms1searchpy[19], n= 3 technical replicates. (**C**): Peptides were separated either on a 25 cm packed bed column or on a 5.5 cm µPAC™ column, data was acquired using DDA and analyzed using CHIMERYS™, n = 12 technical replicates measured on 2 different days. (**D**): As A but using either DDA (2 m/z isolation width), WWA[18] (12 m/z isolation width) or DIA (variable windows 8 – 25 m/z isolation width9 as acquisition strategy. DDA and WWA data were analyzed using CHIMERYS, DIA data was analyzed using Spectronaut® in directDIA mode, n = 3 technical replicates.*

Due to the superiority of the DIA we investigated this strategy in more detail. Figure 5A shows the advantage of using variable windows over fixed ones. By this duty cycle time is reduced as broader windows are used in a m/z region were only few ions are to be expected. Method details are shown in Supplemental Table .1 In line to our results using a DDA approach, protein IDs were boosted clearly, here by ~ factor 2, when comparing results from the packed bed column (Figure 5A) to those of the µPAC™ column

231 (Figure 5B). In both cases the same method and the same gradient was used (dark green bar, 20 min
232 gradient, variable windows). For more comprehensive proteome coverage and improved data complete-
233 ness across replicates we further enabled matching between runs (Figure 5B, no ref. library). Across 3
234 replicates data completeness was close to 100 % and 1500 proteins from 250 pg were approached in the
235 direct DIA search. Enriching the directDIA search data with higher input samples of 1 or 10 ng allows for
236 much better matching and boosts protein IDs to >2000 and >2200 respectively. Enlarging the library size
237 however leads to identifications not possible in all replicates anymore (lowered recovery) and in conclu-
238 sion data completeness is lowered to ~80% for the largest library. Of note, these results are in line with
239 our observations made on a timsTOF with DIA-PASEF and using a library for enrichment using single
240 cells (Pichler et. al., manuscript in preparation).



241

242 *Figure 5:* ***Benchmarking DIA methods and enrichment of searches to boost protein IDs of single cell level inputs****. 250 pg*
243 *HeLa bulk digest was used for injection into LC-MS each, bars represent protein IDs on 1 % FDR level, error bars indicate*
244 *standard deviations.* ***(A)****: Peptides were separated on a 25 cm packed bed column, data was acquired using a fixed window*
245 *size of 8 m/z or variable windows ranging from 8 – 25 m/z and analyzed using Spectronaut®, n = 3 technical replicates.* ***(B)****:*
246 *Peptides were separated on a 5.5 cm µPAC™ column, data was acquired using variable windows ranging from 8 – 25 m/z*
247 *and analyzed using Spectronaut® either separately, (no matching), or together (match between runs) with or without enrich-*
248 *ment using higher input data as indicated, n = 3 technical replicates each.*

249 Similar effects using match between runs (MBR) can be observed in single cell DDA data as well. We
250 analyzed 5 replicates of one individual HeLa cell each using 4 commonly used data analysis algorithms
251 (Figure 6A). MS Amanda[21] and CHIMERYS™ were both combined with apQuant[22] using Proteome Dis-
252 coverer which yielded in >100 additional protein IDs when enabling MBR. A comparable effect was seen

when using MSFragger[23] and IonQuant[24]. In contrast the number of protein IDs was increased by only 5 % when using Spectromine™. To improve sensitivity, we analyzed 5 replicates of 40 HeLa cells together with our single cell runs to enable more potential matches across files (Figure 6B). This increased the size of the matching library to 7327 peptides and 1201 proteins for Amanda and to 16229 peptides and 2955 proteins for CHIMERYS™. The resulting protein identifications for single cells were boosted by roughly 70 % leading to 931 and 1791 average proteins from a single cell in the Amanda and CHIMERYS™ searches respectively. With and without MBR, CHIMERYS clearly outperforms all other tested search engines, which is why it was preferred for all DDA and WWA searches.

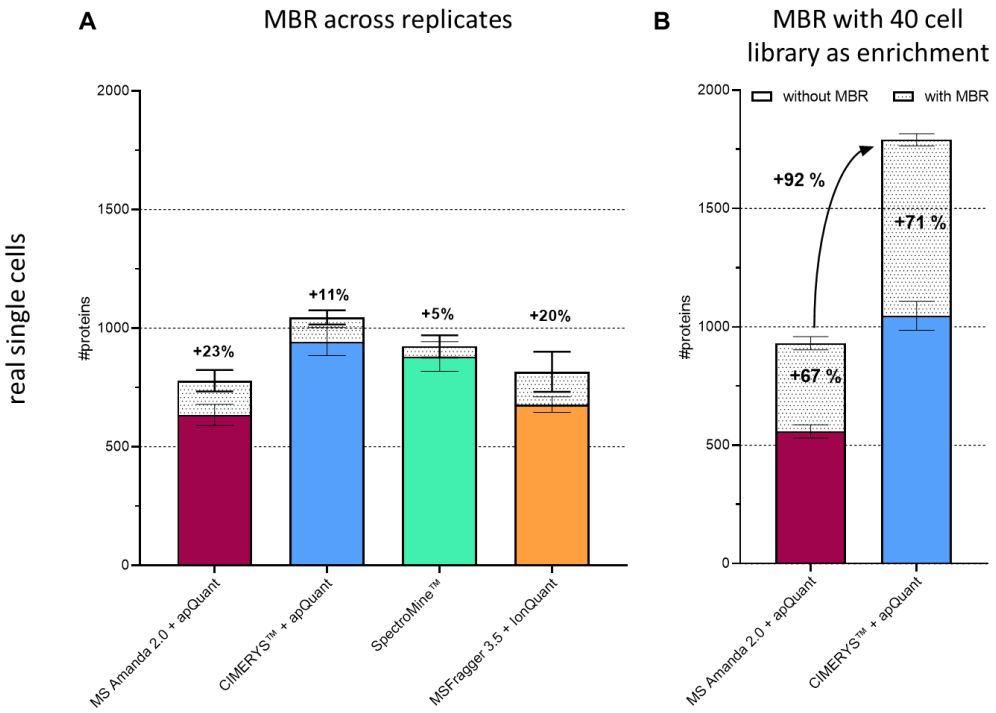


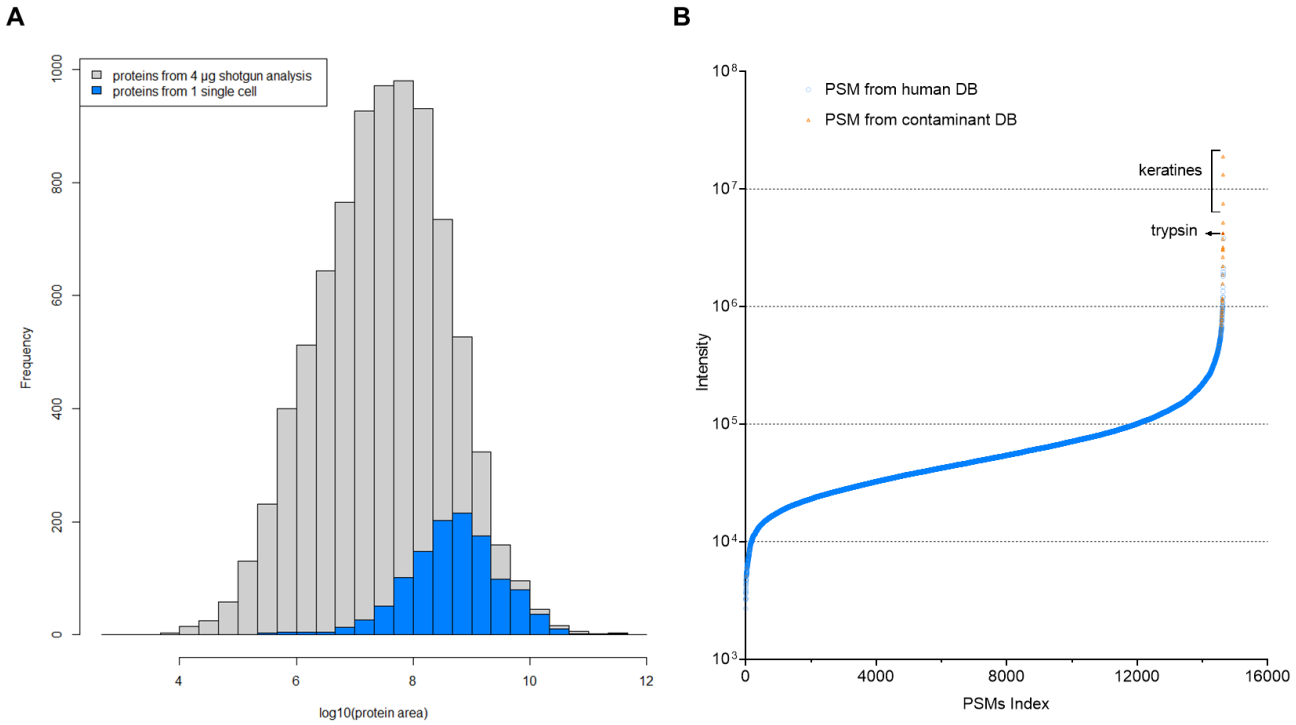
Figure 6: ***Benchmark of different data analysis tools and effect of enrichment for DDA runs****. Individual HeLa cells were processed using the optimized workflow. Peptides were separated using a 20 min active gradient and data was acquired using a DDA method with 2 m/z isolation width. (**A**): Data analysis was performed using the indicated software at 1 % FDR on protein level, either file by file or with MBR across all replicates, n = 5 replicates. (**B**): Data as shown in panel A but processed together with 5 runs of 40 HeLa cells each as enrichment for matching*

Proof of principle studies

Proteome depth

Reaching a sufficient proteomic depth is essential for every single cell workflow that aims to investigate cellular heterogeneity. To estimate the sensitivity of the here presented workflow we plotted all proteins

273 identified from a single cell run that yielded in >1000 proteins, into a histogram showing all protein IDs
274 from a high input sample with more than 8500 proteins quantified. In line with our expectations and as
275 visualized in Figure 7A, the single cell sample predominantly covers those proteins which are most abun-
276 dant within the proteome. Nevertheless, the identified proteins range over more than 4 orders of magnitude
277 within the proteome, which already seems sufficient to investigate abundance differences. We further
278 estimated that the covered dynamic range within this single cell run is in the range of 4 orders of magni-
279 tude by plotting all PSM intensities (Figure 7B). Of note, the most abundant hits are contaminants intro-
280 duced during the workflow. These are mostly keratins and trypsin which needs to be added in excess for
281 digestion.



282

283 *Figure 7:* ***Reachable proteome depth from one cell, label free and using DDA****. Data of a single HeLa cell generated using*
284 *the optimized workflow, acquired using a DDA method and analyzed using CHIMERYS™. (**A**): Proteins, excluding contami-*
285 *nants, identified from a single cell plotted on top of the abundance distribution of all proteins identified in a 4h shotgun run*
286 *from 4 µg HeLa digest. LFQ by apQuant. (**B**): PSMs identified from a single cell run, indexed based on their precursor inten-*
287 *sity. PSM matches originating from the contaminant database (DB) are highlighted in orange.*

288 To ensure that those IDs that are of low precursor signal intensity and of a spectrum quality that is just
289 accepted by the 1% FDR threshold, we performed an additional analysis including an entrapment of the
290 yest proteome in the database search. We found, that 1.5 % of all peptides and 1.3 % of all proteins in the
291 result file were originating from the yeast proteome, which indicates a proper overall FDR calculation.
292 Furthermore, those PSMs identified from yeast are not accumulated in the low abundant fraction but
293 evenly distributed across the whole intensity range. This indicates that not only crap is reported in the low

abundant fraction and the achieved dynamic range is truly in the range of 4 orders of magnitude (Supplemental Figure 3).

Application of optimal workflow parameters on 2 human cell lines – proof of principle study

To evaluate the final workflow, individual HeLa and K652 cells were isolated, processed and analyzed using those conditions which worked best in previous optimizations: Cells were kept hydrated, trypsin was added twice and DMSO was supplemented for storage. Peptides were separated using a 20 min active gradient on the 5.5 cm µPAC™ column and data was acquired using a DIA method with variable windows. Our data shows that a clear separation of both cell types is possible although the number of replicates is relatively small (Figure 8A). With this we prove that our workflow is sensitive enough to investigate cellular heterogeneity in principle. On average 988 and 859 proteins were identified from a single HeLa or K562 cell respectively (Figure 8B). Matching to the previously recorded 40 HeLa cell data or 10 ng HeLa bulk data did not yield in significant changes or only slightly boosted of the resulting protein IDs respectively. This might be reasoned by an already good matching across all replicates or by different type of peptides available in a bulk vs real single cell sample. For K562 such a matching did not influence the result which might be reasoned by matching to a different cell line (HeLa) is not optimal.

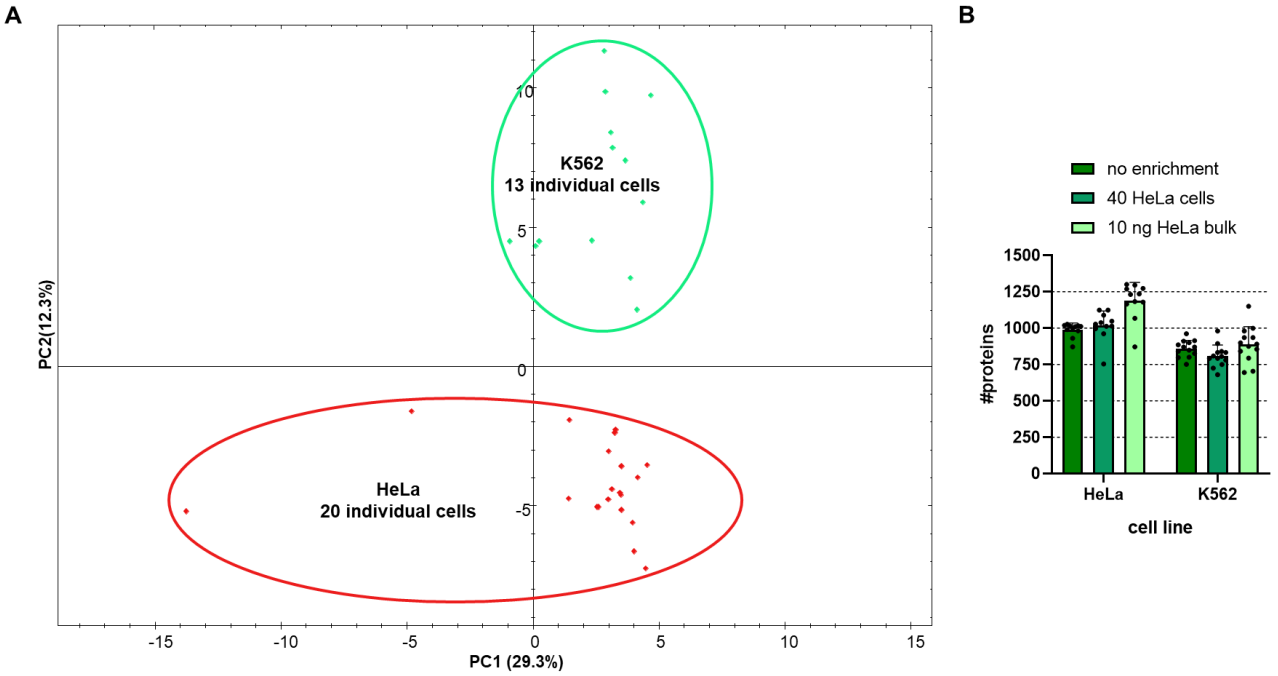


Figure 8: ***Applying optimal parameters for sample preparation, data acquisition and data analysis on two human cell lines****. Individual HeLa or K562 cells were processed using the optimized workflow and resulting peptides were separated using a 5.5 cm µPAC™ column and a 20 min active gradient. Data acquisition was performed by employing a DIA method with variable window sizes and data analysis was done by Spectronaut® at 1% FDR with match between runs. (**A**): Principal component analysis as given by Spectronaut® without using any library for enrichment, each dot represents one cell with HeLa cells highlighted in red and K562 highlighted in green. (**B**): Bars represent the average number of identified proteins*

316 *when analyzing data in directDIA mode and adding data from 40 HeLa cells of 10 ng HeLa bulk digest as enrichment, as*
317 *indicated.*

318

319 **CONCLUSIONS**

320 Here we present an optimized sample preparation workflow for single cell proteomics. The most im-
321 portant parameter to minimize losses of such ultra-low input samples seems to perform all steps in a single
322 pot. By omitting any transfer step but inject directly from a 384 well plate, not only proteome coverage
323 but also reproducibility was improved. The wells of a 384 plate thereby serve for all workflow steps from
324 cell isolation to injection into the LC-MS system. Furthermore, preventing drying of samples and the
325 addition of fresh protease after 30 min clearly boosted our results. We hypothesize that this not only
326 elongates the total time of efficient digestion but also reduces adsorption of hydrophobic peptides on the
327 well walls. This effect was supported by supplementation with DMSO for sample storage keeping pep-
328 tides solubilized.

329 The CellenONE® robot allows effective and semi-automated cell isolation, serves as incubation chamber
330 with controlled humidity for lysis and digestion and can automatically perform hydration. Although this
331 study was performed using the CellenONE® all steps, except cell isolation, could be easily adopted to a
332 low-tech environment as well since this workflow is performed in a volume compatible with manual
333 pipetting (1µL) and within standard 384 well plates. Such plates are also compatible for cell isolation
334 using alternative instrumentation as a FACS device which is available in many labs.

335 Since a powerful workflow not only covers sample preparation but a complete bundle of steps including
336 LC separation, data acquisition, analysis, and interpretation, we further aimed to improve the latter steps.
337 We found that the use of a short µPAC™ column with highly ordered and superficially porous brick
338 shaped pillars allows for excellent peptide separation as well as fast loading and equilibration thanks to
339 an exceedingly low backpressure. While analysis employing a fast 5 min gradients is possible and leads
340 to more than 700 protein identifications on average with an MS1 only acquisition, best results were ob-
341 tained using a 20 min active gradient that has a reasonable run to run time of 39 min. This allows for 37
342 samples a day which is still a low throughput compared to a multiplexed workflow. However, a multi-
343 plexed workflow may suffer from ratio compression and precursor co-isolation hampering accurate quan-
344 tification.[14,25–27] A label free strategy allows for unbiased and accurate quantification and is fully compat-
345 ible with data independent acquisition, which is why it can be the preferred way to go in case very high
346 throughputs are not the limiting factor.

To conclude, the here presented workflow allows for identification of more than 1000 proteins using DDA without MBR and more than 1790 with MBR over 4 orders of magnitude in abundance from a single cell. Our results further show that proteomic depth is even improved using WWA or DIA as acquisition strategy yielding in up to 2250 proteins from a single cell level input, as well as a successful differentiation of cell types based on alterations in protein abundance.

We believe that the workflow presented is easy to implement, robust and sensitive. It will be the basis for future improvements enabling deeper studies into the proteome of individual cells without the need for carrier or labelling. In addition, the presented techniques are not only suitable for single cell applications but also for all proteomic studies that are limited in sample amount and will therefore be highly valuable for a broad scientific community.

METHODS

Sample preparation

HeLa cells were cultured at 37°C and 5% CO_2 in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% FBS (10270, Fisher Scientific™, USA), 1x penicillin-streptomycin (P0781-100ML, Sigma Aldrich, Israel) and 100X L-Glut 200 mM (250030-024, Thermo Scientific™, Germany). After trypsinization with 0.05% Trypsin-EDTA 1x (25300-054, Thermo Scientific™, USA) cells were washed 3x with phosphate-buffered-saline (PBS). K562- WT cells were cultured at 37°C and 5% CO_2 in Gibco Roswell Park Memorial Institute (RPMI -1640) medium (Fisher Scientific™, USA). For harves K562 were pelleted by centrifugation (117g, 4°C) and washed 3x with phosphate-buffered-saline (PBS). All cells were resuspended n PBS at 200 cells/µL for isolation within the CellenONE®.

Benchmarking experiments mimicking single cell digests were prepared from dilutions of the Pierce™ HeLa (#1862824, Thermo Scientific™) protein digest standard in 0.1% TFA with or without 5 % DMSO supplemented as indicated.

Sample lysis and digestion were performed within a clean 384 well plate inside the CellenONE®. 1 µL of a master mix containing 0.2% DDM (D4641-500MG, Sigma Aldrich, Germany), 100mM TEAB (17902-500ML, Fluka Analytical, Switzerland), 3 ng/µL trypsin (Trypsin Gold, V5280, Promega, USA), 0.01 % enhancer (ProteaseMAX™, V2071, Promega, USA) were predispensed into the wells. To limit evaporation humidity was set to the 85%. Individual cells at 18 - 25 µm diameter and a max elongation of 1.5 are sorted by the CellenONE® into the respective wells. This was followed by incubation for 2 h at 50 °C at 85% relative humidity inside the instrument. Samples were kept hydrated every 15 min by automated addition of 500 nL water to each well. After 30 min of incubation, additional 500nL of 3ng/µL

trypsin are added which replaces one hydration step. After lysis and digestion, 3.5 µL of 0.1% TFA with 5% DMSO were added to the respective wells for quenching and storage. Samples were either directly injected from the 384 well plate or transferred to a PEG coated PCR tubes (PCR-02-L-C, Axygen®) or to the autosampler glass vials (601801655, Thermo Scientific).

LC-MS/MS analysis

Samples were analyzed using the UltiMate 3000 RSLCnano system (Thermo Scientific™) or the Vanquish™ Neo UHPLC system (VN-S10-A01, Thermo Scientific™). Peptides were separated on a 5.5 cm brick shape pillar column prototype (µPACTM, Thermo Fisher) or on a nanoEase M/Z Peptide CSH C18 column (130Å, 1.7 µm, 75 µm x250 mm, 18600810, Waters, Germany) as indicated in a trap and elute setting using an Acclaim™ PepMap™ 100 C18 trapping column (5 µM, 0.3 mm x 5 mm, Thermo Scientific™). All columns were operated at 50°C and connected to an EASY-Spray™ bullet emitter (10 µm ID, ES993; Thermo Fisher Scientific) An electrospray voltage of 2.4 kV was applied at the integrated liquid junction of EASY-Spray™ emitter. Mass spectrometry measurement was performed by an Orbitrap Exploris™ 480 Mass Spectrometer (Thermo Scientific™) equipped with a FAIMS Pro interface (Thermo Scientific™).

If not noted other, peptide separation, was performed at a constant flow rate of 250nL/min and the following 20 min active gradient: 1% buffer B (acetonitrile (ACN) with 0.08% formic acid) and 99% buffer A (0.1% formic acid (FA) from minute 0 to 2.5, rising to 2.5% B from minute 2.5 to 2.75, continues to rise to 25% until minute 17, followed by an increase from 25 to 40% in the last 5 minutes. The column gets washed by increasing buffer B to 97.5% until minute 22 for 10 minutes before buffer B concentration is decreased back to 1% for re-equilibration. This gradient was shortened to 10 and 5 min accordingly for benchmarking studies. Column re-equilibration was done for 18 min at 250nL/min when using the packed bed column. The µPACTM column was used on the Vanquish Neo with fast loading and equilibration at a maximum pressure of 350 bar and a maximum flowrate of 1µL/min enabled. For MS1only acquisition a linear gradient was applied instead, ranging from 1 -40 % buffer B but with no other changes compared to the above-described gradient.

MS/MS method settings are detailed in Supplemental Table 1 and 2.

Data analysis

If not indicated other, data analysis was performed using CHIMERYS™ within Proteome Discoverer™ (v3.00.757, Thermo Fisher). Data was searched against the human reference database from Uniprot (version: 20.08.2021 UP000005640, 20300 sequences, 11359367 residues) and a customized contaminants database (364 sequences; 142459 residues). For evaluation of the false- discovery rate a Yeast database was used (version 22.12.2021 UP000002311, 6050 sequences). Trypsin was specified as proteolytic enzyme, cleaving after lysine (K) and arginine (R) except when followed by proline (P) and up to two missed cleavages were allowed. Fragment mass tolerance was limited to 20 ppm and carbamidomethylation of cysteine (C) was set as a fixed modification and oxidation of methionine (M) as a variable modification. Identified spectra were rescored using Percolator[28] and results were filtered for 1% FDR on peptide and protein level. Abundance of identified peptides was determined by label-free quantification (LFQ) using IMP-apQuant[22].

Besides that, for benchmarking, data was also analyzed using MSAmanda (v2.0.0.19742)[21], SpectroMine™ (Version 3.2.220222.52329, Biognosys AG), MSFragger[23] and IonQuant[24] within FragPipe v18 and for DIA results Spectronaut™ (Version 16.1.220730.53000, Biognosys AG) was used in the directDIA mode. For MS1 only data ms1searchpy was used[19,20]. Detailed search settings can be found in Supplemental Table 3.

ASSOCIATED CONTENT

Supporting Information

Supporting information is available to download free of charge.

Supplemental Figure 1: Single cell isolation by visual detection using the cellenONE®.

Supplemental Figure 2: Venn diagram showing proteins commonly found in a representative replicate from 250 pg HeLa

Supplemental Figure 3: Reachable dynamic range from one cell, label free and using DDA

Supplemental Figure 4: Improving recovery and reproducibility by tuning workflow parameters

Supplemental Table 1: Label free MS/MS Method for DDA, DIA and MS1only

Supplemental Table 2: Summary of the variable isolation windows used for DIA

Supplemental Table 3: Summary of data analysis settings for different software and search algorithms

AUTHOR INFORMATION

Author Contributions

EM and MM performed experiments, data analysis and wrote the manuscript. EM and MM contributed equally to this work. MM and KM conceptualized the study. GD helped with data analysis and R scripting.

ACKNOWLEDGMENT

This work supported by the EPIC-XS, Project Number 823839, funded by the Horizon 2020 Program of the European Union, by the project LS20-079 of the Vienna Science and Technology Fund and the by the ERA-CAPS I 3686, P35045-B, P32054 (FB) and P33380 (FB) project of the Austrian Science Fund. We thank the IMP for general funding and access to infrastructure and especially Karel Stejskal and Gabriela Krššáková, for continuous laboratory support, Fränze Müller for help in setting up DIA methods and Rupert Mayer for proofreading the manuscript. Special gratitude goes to Ulla Schellhaas for providing us with K562 cells.

REFERENCES

1. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677 (2015).
2. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
3. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281.e24 (2019).
4. Popescu, D.-M. *et al.* Decoding human fetal liver haematopoiesis. *Nature* **574**, 365–371 (2019).
5. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24**, 11–17 (2015).
6. Kelly, R. T. Single-cell Proteomics: Progress and Prospects. *Mol. Cell. Proteomics MCP* **19**, 1739–1748 (2020).
7. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548 (2011).
8. Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* **22**, 50 (2021).
9. Williams, S. M. *et al.* Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography–Mass Spectrometry for High-Throughput Single-Cell Proteomics. *Anal. Chem.* **92**, 10588–10596 (2020).
10. Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun.* **9**, 882 (2018).
11. Ctortecka, C. *et al.* An automated workflow for multiplexed single-cell proteomics sample preparation at unprecedented sensitivity. 2021.04.14.439828 Preprint at https://doi.org/10.1101/2021.04.14.439828 (2022).
12. Tsai, C.-F. *et al.* An Improved Boosting to Amplify Signal with Isobaric Labeling (iBASIL) Strategy for Precise Quantitative Single-cell Proteomics *. *Mol. Cell. Proteomics* **19**, 828–838 (2020).
13. Dou, M. *et al.* High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform. *Anal. Chem.* **91**, 13119–13127 (2019).
14. Cheung, T. K. *et al.* Defining the carrier proteome limit for single-cell proteomics. *Nat. Methods* **18**, 76–83 (2021).
15. Ctortecka, C., Stejskal, K., Krššáková, G., Mendjan, S. & Mechtler, K. Quantitative Accuracy and Precision in Multiplexed Single-Cell Proteomics. *Anal. Chem.* **94**, 2434–2443 (2022).
16. Ye, Z., Batth, T. S., Rüther, P. & Olsen, J. V. A deeper look at carrier proteome effects for single-cell proteomics. *Commun. Biol.* **5**, 1–8 (2022).
17. James, M. B. & Giorgio, T. D. Nuclear-Associated Plasmid, but Not Cell-Associated Plasmid, Is Correlated with Transgene Expression in Cultured Mammalian Cells. *Mol. Ther.* **1**, 339–346 (2000).
18. Mayer, R. L. *et al.* Wide Window Acquisition and AI-based data analysis to reach deep proteome coverage for a wide sample range, including single cell proteomic inputs. 2022.09.01.506203 Preprint at https://doi.org/10.1101/2022.09.01.506203 (2022).
19. Ivanov, M. V. *et al.* DirectMS1: MS/MS-Free Identification of 1000 Proteins of Cellular Proteomes in 5 Minutes. *Anal. Chem.* **92**, 4326–4333 (2020).

20. Ivanov, M. V. *et al.* Boosting MS1-only Proteomics with Machine Learning Allows 2000 Protein Identifications in Single-Shot Human Proteome Analysis Using 5 min HPLC Gradient. *J. Proteome Res.* **20**, 1864–1873 (2021).

21. Dorfer, V. *et al.* MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. *J. Proteome Res.* **13**, 3679–3684 (2014).

22. Doblmann, J. *et al.* apQuant: Accurate Label-Free Quantification by Quality Filtering. *J. Proteome Res.* **18**, 535–541 (2019).

23. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520 (2017).

24. Yu, F., Haynes, S. E. & Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell. Proteomics* **20**, 100077 (2021).

25. Kelstrup, C. D. *et al.* Limits for Resolving Isobaric Tandem Mass Tag Reporter Ions Using Phase-Constrained Spectrum Deconvolution. *J. Proteome Res.* **17**, 4008–4016 (2018).

26. Savitski, M. M. *et al.* Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. *J. Proteome Res.* **12**, 3586–3598 (2013).

27. Ctortecka, C. & Mechtler, K. The rise of single-cell proteomics. *Anal. Sci. Adv.* **2**, 84–94 (2021).

28. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).

29. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).