Cell cycle state proteomics and classification using in-cell protease digests and mass spectrometry

Authors:

Van Kelly, Aymen al-Rawi, David Lewis and Tony Ly*

Addresses:

1. Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Michael Swann Building, Max Born Crescent, Edinburgh EH9 3BF, UK

* Correspondence should be addressed to Tony Ly (tly@ed.ac.uk)

Running title: Cell state proteomics and classification
Abbreviations: PRIMMUS, AMPL

Abstract (175 words max)

Proteomic analysis of rare cell states is a major challenge. We report an advance to our *PR*oteomics of *I*ntracellular i*MMU*nostained cell *S*ubsets (PRIMMUS) workflow whereby fixed cells are directly digested by proteases *in cellulo* for mass spectrometry-based proteomics. This decreased the cell number requirement by two orders of magnitude to <2,000 human lymphoblasts. We quantitatively measured the proteomes of 8 interphase and 8 mitotic states, avoiding synchronization. From 8 replicate pseudo-timecourses, we identify a core set of 119 cell cycle-regulated proteins that segregated into five clusters. These clusters varied in mitotic abundance patterns and regulatory short linear sequence motifs controlling their localisation and interaction with E3 ubiquitin ligases. We identified protein signatures that allowed accurate cell cycle state classification. We use this classification to stage an unexpected cell population as similar in proteome to early G0/G1 and telophase cells. Our data indicate DNA damage responses and premature APC/C activation in these cells, consistent with a DNA damage-induced senescent state. The advanced PRIMMUS approach is readily and broadly applicable to characterise rare and abundant cell states.

Introduction

The proteome is a functional readout of cellular phenotype, which includes dynamic and persistent molecular features that reflect cell state and cell type, respectively. Rare cell phenotypes play key physiological roles. Quiescent stem cells, while often rare relative to differentiated cell types in a tissue, are essential for tissue homeostasis. Similarly, mitosis is critical for the accurate propagation of genetic material and a phase during which cellular commitment to proliferation is made [1][2] . Mitotic states are generally short-lived and thus rare in an asynchronous population. Proteomic analysis of these critically important cell phenotypes is a major challenge because typical proteomic workflows require $>10^5$ cells as input.

Recent advances have been made in methods for low cell number proteome analysis. For example, ~2,700 proteins were identified from 6,250 CD34+ hematopoietic progenitor cells using optimized in-solution digests combined with data-independent acquisition (DIA) [3]. ~3,000 proteins were identified from 10 HeLa cells using 'nanodroplet processing in one pot for trace samples' (nanoPOTS) [4]. Single cell proteomic analysis using nanoPOTS with tandem mass tag (TMT) booster channels has been recently described [5]. NanoPOTS requires microfabricated glass chips, robotics that can handle picoliter volumes and sample storage in prepacked nano-LC columns. These requirements are challenging to satisfy in most labs and limit widespread adoption of the technique.

We previously developed an approach called 'PRIMMUS' or '*P*roteomics of *I*ntracellular I*mmu*nostained *Subsets*" to analyse abundant and rare cell cycle states [6]. Formaldehyde-fixed cells are fractionated into specific cell states by staining cells for intracellular markers and separating them using Fluorescence-Activated Cell Sorting (FACS). Cells grown in asynchronous culture are immediately fixed, thereby minimizing perturbation to physiological processes. This step is critical, as small molecule-based synchronsation can lead to effects on the proteome that are associated with stress responses arising from arrest rather than cell cycle regulation per se [7]. The application of PRIMMUS was limited to abundant subpopulations where $>10^5$ cells can be collected by FACS within a reasonable time [6]. A more sensitive PRIMMUS approach would enable high resolution mapping of proteomic changes during an unperturbed cell cycle,

including the analysis of mitotic states, where major changes in organellar structures, protein abundances and protein post-translational modifications are highly dynamic on the minute timescale.

Here, we report a major advance in the PRIMMUS method that increases sensitivity and enables detailed proteomic analysis of rare cell populations. The new approach makes use of the discovery that formaldehyde-fixed cells are suitable substrates for tryptic digestion without prior crosslinking removal, which greatly simplifies sample preparation. We combined this streamlined workflow with MS1-based library matching and an MS acquisition method that prioritizes quantitative MS1 quality, called AMPL (Averaged MS1 Precursor with Library matching). These improvements together enabled reproducible quantitation of ~4,500 proteins from 2,000 human lymphoblastoid cells with a low data dropout frequency.

We applied the advanced PRIMMUS workflow to analyze the proteomes of 16 cell cycle subsets, including 8 interphase and 8 mitotic subphases. We identified a core set of 119 cell cycle regulated proteins. Many of these proteins are well-characterized as having key functions in cell cycle regulation. We now provide detailed resolution on their variations in protein abundance across an unperturbed cell cycle. Novel cell cycle regulated proteins include FAM111B, which by sequence similarity to FAM111A, has putative roles in regulating DNA replication. We showed that the cell cycle regulated proteome is predictive of cell cycle state. PCA analysis correctly assigns the expected order of the subsets according to their temporal relationships in the cell cycle. We use this classification system to group an unexpected, rare subpopulation with 4N DNA content as more closely resembling G0/early G1 cells, but with an additional DNA damage response signature. Our data suggest that these cells are G2 cells entering senescence after DNA damage, consistent with previous reports [8]. The enhanced resolution in mitosis allowed us to identify two groups of proteins that are characterized by early versus late decreases in abundance. These two groups likely reflect a switch from early mitotic E3 ligases, including the APC/C-Cdc20 and SCF(Cyclin-F), to late mitotic E3 ligases, including APC/C-Cdh1.

Experimental Procedures

Reagents and antibodies

A description of reagents, including cell lines and antibodies, can be found in the Key Resources Table (Supplementary Table 1).

Cell culture

TK6 human lymphoblasts [9] were obtained from the Earnshaw laboratory (University of Edinburgh). Cells were cultured at 37 °C in the presence of 5% CO_2 as a suspension in RPMI-1640 + GlutaMAX (Thermo Scientific) supplemented with 10% v/v fetal bovine serum (FBS, Thermo Scientific). Cell cultures were maintained at densities no higher than 2 x 10^6 cells per ml. MCF10A cells (ATCC) were cultured in phenol red-free F12/DMEM media (Thermo Scientific) supplemented with 5% horse serum, 10 µg/ml insulin (Sigma), 100 ng/ml cholera toxin (Sigma), 20 ng/ml EGF (Sigma), 0.5 µg/ml hydrocortisone (Sigma), 100 units/ml penicillin and 100 µg/ml streptomycin (Thermo Scientific) at 37 °C in the presence of 5% CO_2. Cells were maintained at less than 100% confluency and were discarded when passage number exceed 20 passages. U2OS cells (ATCC) were cultured in DMEM media high glucose + GlutaMAX (Thermo Scientific) supplemented with 10% v/v FBS (Thermo Scientific). Cells were checked for mycoplasma at the point of cryo-storage using a luminescence-based assay (Lonza).

Cell fixation and immunostaining

Cells were washed with Dulbecco's phosphate-buffered saline (DPBS, Lonza) and resuspended in freshly prepared 1% formaldehyde solution (w/v) from a 16% stock (w/v, Thermo Scientific) in DPBS, fixed for 10 min at room temperature with gentle rotation, pelleted, washed with DPBS and permeabilized with cold 90% methanol. Cells were stored at -20°C prior to staining.

Cells stored in methanol were washed with DPBS and resuspended in blocking buffer, which is composed of 5% bovine serum albumin (BSA) in 0.1 M Tris-buffered saline, pH 7.4 (TBS). Cells were blocked for 10 min at room temperature, pelleted, and resuspended in primary antibody solution (1:200 in blocking buffer). Cells were stained with primary antibody overnight at 4°C. Stained cells were then washed twice with wash buffer (DPBS + 0.5% BSA) and stained with dye-conjugated secondary antibodies (1:200 in blocking buffer) for 1 hour at room temperature. Stained cells were washed twice with DPBS, pelleted, and stained in DAPI solution (20 µg/ml in DPBS + 0.1% BSA) for at least 1 hour prior to FACS.

FACS and gating strategy

Cells were collected using a BD FACSAria Fusion Cell Sorter equipped with 355nm UV, 405nm Violet, 488nm Blue, 561nm YG and 640nm Red lasers, and controlled by BD FACS Diva V8.0.1 software. Cells were first gated into 'narrow' (P1 – P8) and 'wide' (P9 – P16) populations based on DAPI fluorescence signal width. The narrow population contains single cells either in interphase, or in mitosis up to late anaphase. These single cells were then separated based on cyclin B into 8 different stages of interphase. Population P1 has low to no cyclin B protein and 2N DNA content, consistent with low to no E2F activity and a G0/early G1 cell state. Cyclin B rises monotonically from P2 to P6 and then rises more steeply from P6 to P8. Like cyclin B, cyclin A also increases during interphase, but at a faster rate from P1 to P6 as compared to P6 to P8. P9 to P13 are positive for histone H3 phosphorylation at Ser28 (pH3+). Highest levels of pH3+ are present in prometaphase and metaphase. Rising and declining H3 phosphorylation in early and late mitosis, respectively, result in low to medium levels of pH3+. Cyclin A and cyclin B levels are used to further discriminate mitotic subphases, as they are degraded during prometaphase and the metaphase-to-anaphase transition, respectively.

Finally, late mitotic subphases are enriched in the wide population, but so too are doublets. We reasoned that most doublets will have cyclin B signal, as single cells with the exception of P1 are cyclin B positive. Thus, we can further enrich late mitotic stages by selecting wide, 4N, cyclin B negative cells (P14-P16). P14-P16 are then discriminated further by pH3+ levels, which decrease during mitotic exit. We note that P16 may contain doublets of G0/early G1 cells (P1), but P14 and P15 should not as P14 and P15 are pH3+ and G0/early G1 cells are negative for pH3.

5000 cells for each gated population were collected using 4-way purity using either a 85 or 100 µm nozzle, into 1.5 ml Eppendorf Protein Lo-Bind tubes. Four biological replicates were collected. An interphase library sample were collected by combining 300,000 cells of G0/G1, S, and G2 populations. A mitotic library sample was composed of 800,000 mitotic cells gated by high DNA content and high Histone H3 Ser28 phosphorylation. Samples were centrifuged and supernatant removed before storing at -20 ºC.

In-cell digest

Cell sorted library samples, and unstained unsorted TK6 cells, were resuspended in DPBS at 2 - 5 million cells per ml and incubated with 1 µl (25 - 29 U) benzonase (Millipore) at 37 ºC for a minimum of 1 hr. Trypsin was added to approximately 1:25 w/w and in-cell digested at 37 ºC for ~16 hrs. Digests were acidified with TFA and desalted over Sep-Pak C18 cartridges (Waters) and dried.

Individual populations of 5,000 cells were diluted with 40 µl PBS and incubated with 0.25 µl [6 – 7 U] benzonase at 37 ºC for a minimum of 1 hr, then digested with 50 ng trypsin (~1:10 w/w) at 37 ºC for ~16 hrs. Samples were acidified with TFA and desalted over self-made C18 columns with 3 Empore C18 disks [10] and eluted directly into Axygen™ 96-well PCR Microplates (Fisher Scientific) and dried.

High pH reverse phase fractionation

Approximately 100 µg interphase, mitotic, and unsorted TK6 cell digests were fractionated by high-pH reverse phase chromatography using an Ultimate 3000 HPLC (ThermoFisher Scientific) and a 1 x 100 mm 1.7 µm Acquity UPLC BEH C18 column (Waters). Peptides were separated using a constant 10 mM ammonium formate (pH 10) and a gradient of water and 100% acetonitrile. Peptides were loaded at 1% acetonitrile followed by separation by a 48 min multistep gradient of acetonitrile from 3% to 6%, 25%, 45% and 80% acetonitrile at 4, 34, 44, 45 minutes, respectively, followed by an 80% wash and re-equilibration. Fractions were collected at 30 sec intervals resulting in 96 fractions which were concatenated into 12, and 1 µg aliquots dried.

LC-MS/MS

Peptide samples were resuspended in 0.1% TFA. Approximately 0.5 µg of library fractions were injected for DDA LCMS analysis. A volume equal to half the cell population (equivalent to ~2,500 cells) was injected and analysed twice by AMPL to produce two technical replicates for each of the four biological replicates. An Ultimate 3000 RSLCnano HPLC (Dionex, Thermo Fisher Scientific) was coupled via electrospray ionisation to an Orbitrap Elite Hybrid Ion Trap-Orbitrap (Thermo Fisher Scientific). Peptides were loaded directly onto a 75 µm x 50 cm PepMap-C18 EASY-Spray LC Column (Thermo Fisher Scientific) and eluted at 250 nl/min using 0.1% formic acid (Solvent A) and 80% acetonitrile/0.1% formic acid (Solvent B). Samples were eluted

over 90 min stepped linear gradient from 1% to 30% B over 72 min, then to 45% B over 18 min. AMPL analyses included up to 5 MS1 microscans of 1E6 ions in the Orbitrap at 120k resolution and with a 250 ms maximum injection time. MS1 scans were acquired over 350-1700 m/z and a 'lock mass' of 445.120025 m/z was used. This was followed by 5 data-dependent MS2 CID events (5E3 target ion accumulation) in the ion trap at rapid resolution with a 2 Da isolation width, a normalised collision energy of 35, 50 ms maximum fill time, a requirement of a 10k precursor intensity, and a charge of 2+ or more. Precursors within 5 ppm were dynamically excluded for 40 sec. DDA analyses were as for AMPL but with a single MS1 microscan with a 75 ms maximum injection time, followed by 20 CID events in the ion trap.

Libraries were acquired as for DDA analyses or acquired with 10 data-dependent MS2 HCD events at 30 NCE of 5E4 ions in the Orbitrap at 15k resolution and a maximum fill time of 100 ms, with a precursor intensity required to be at least 50k. For the sample preparation comparisons shown in Fig. 2, a 240 min gradient was used (1% to 30% B for 210 min, then to 42% B over 30 min). MS data was acquired as for DDA analysis described above with the exception that MS1 spectra were acquired at 60k resolution and MS2 events were acquired only on 2+ and 3+ precursors.

MS/MS data analysis

Data was processed using MaxQuant version 1.6.2.6 [11]. LC-MS/MS data was searched against the "Human Ref Proteome _ALL_2017-10-23.fasta" database allowing for variable methionine oxidation and protein N-terminal acetylation. Carbamidomethyl cysteine modification was allowed only for samples that were alkylated by iodoacetemide. A target-decoy threshold of 1% was set for both PSM and protein false discovery rate. Match-between-runs was enabled with identification transfer within 0.5 mins and a retention time alignment within 20 min window. Matching was permitted from the library parameter group, and 'from and to' the unfractionated parameter group. "Require MS/MS for LFQ comparisons" was deselected, and second peptide search was enabled. Both modified and unmodified unique and razor peptides were used for quantification. 'Evidence' and 'proteinGroups' output files were used for subsequent analysis in R.

Match-between-runs FDR filtering

A reference sample was generated by lysing TK6 cells in DPBS with 2% SDS and cOMPLETE protease inhibitors without EDTA (Roche, 1x concentration) at 70 ºC, homogenised with a probe sonicator and treated with benzonase. Protein was reduced with 20 mM TCEP for 2 hr before alkylation with 20 mM iodoacetamide at ambient temperature in the dark for 1 hr. Protein was precipitated with 4 volumes cold acetone at -20 ºC overnight, washed with 100% cold acetone and 90% cold ethanol. Protein pellet was air dried before resuspending in DPBS and digesting with 1:50 w/w trypsin for ~16hrs. Peptides were acidified, desalted, aliquoted, and fractionated as previously described. For isopropylation, 50 µg peptides were resuspended in 200 µl 90% acetonitrile containing 0.1% formic acid before addition of 50 µl acetone containing 36 µg/µl $NaBH_3CN$. The reaction was conducted at ambient temperature for ~16 hrs before quenching with ammonium bicarbonate, drying off solvent and desalting peptides over C18. For dimethylation, 50 µg peptide was resuspended in 200 µl DPBS before addition of 0.32% formaldehyde and 50 mM $NaBH_3CN$. The reaction was conducted at ambient temperature for ~16 hrs before quenching with ammonium bicarbonate and desalting peptides over C18. 200 ng of unmodified, dimethylated, and isopropylated peptides were analysed by AMPL and DDA, and unmodified fractionated peptide samples were analysis by DDA, as previously described. LCMS data were searched using MaxQuant, as previously described. Note that dimethylation and isopropylation modifications were not specified in in the search parameters.

Cell cycle proteomic data analysis

All subsequent data analysis on the protein intensity table, including the analysis of pseudoperiodicity, was performed using R (v. 3.5.0) within the RStudio integrated development environment. The R script will be made available on http://dynamic-proteomes.squarespace.com. The list of validated APC/C substrates was obtained from the APC/C degron repository (http://slim.icr.ac.uk/apc/). Proteins that contain D box, KEN and ABBA SLIMs in the human proteome were found using SLiMsearch with default settings (Disorder score cut-off: 0.30, Flank length: 5). In order to remove slight variations in total protein amount in each sample, protein intensities were divided by total intensities per sample and multiplied by 10^6 to obtain intensities in parts per million (ppm). There are four biological replicates analysed in technical duplicate. As described

above, sample analysis was completely randomized in the second technical repeat. Each technical repeat (i.e. set of four biological replicates) are considered as one 'pseudotimecourse' with samples in each biological replicate arranged in order from P1 to P16. Each of the two pseudotimecourse was then independently subjected to a Fisher's test for periodicity, as implemented in the ptest R library (v. 1.0-8). Fisher's periodicity test p-values were corrected for multiple hypothesis testing using the q value method as implemented in the qvalue R library (2.15.0). Those proteins that showed q values <0.10 in both sets of biological replicates and oscillation frequencies of either 0.0625 (1/16) or 0.125 (1/8) were classified as pseudoperiodic.

For clustering, protein ppm values were averaged (mean) to produce a single pseudotimecourse for each protein. These average abundance profiles were scaled using the base R function scale and subjected to hierarchal clustering using the Ward minimum variance algorithm. The appropriate range for cluster number was identified as 3 - 6 clusters using the 'elbow method', which involves plotting within-cluster sum of squares versus number of clusters. Bifurcating leaves of the subsequent dendrogram were swapped in order to produce a heatmap that follows a logical, sequential order of peak abundance, i.e. cluster 1 with highest abundance in P0-P8 and cluster 5 with peak abundance in P3-P7, etc.

For PCA and cell cycle state classification, scaled pseudotimecourses were used. Cell cycle states were classified using the k-NN model as implemented in the class R library (v. 7.3-15) using k = 6, with k being the number of nearest neighbours for classification. Three biological replicates were used as the training set and the remaining replicate was used as a test set.

For the pairwise comparison of the proteomes of P17 with P1 and P16, t-tests were performed on ppm intensities. Uncorrected p-values were plotted against mean fold change in order to identify candidate proteins that were specifically changed in abundance in P17.

Results

Impact of formaldehyde crosslinking on whole proteome analysis

Heat treatment at 95 °C is sufficient to reverse most formaldehyde crosslinks, as shown previously [9]. However, a pool of crosslinked, multimeric species remained in a protein-

dependent manner. Therefore, we aimed to optimize the PRIMMUS approach by first focusing on improving the decrosslinking efficiency (Fig. 1A).

Previous reports have suggested that the reversal step is accelerated by co-treatment with a nucleophilic quenching agent [12]. We tested addition of Tris and hydroxylamine on crosslink removal (Fig. 1B). Banding patterns of extracts from fixed cells heat treated for 15 and 45 mins (lanes 4 & 5) were similar to cells heated at the same temperatures in the presence of either 0.5 M Tris (lanes 6 & 7), or 1.25% hydroxylamine (lanes 8 & 9). Addition of both Tris and hydroxylamine led to a significant reduction in high MW crosslinked species with 15 min heat treatment (lane 10). A 45 min incubation led to a diffuse banding pattern (lane 11), indicative of protein degradation as previously reported for hydroxylamine [13]. Indeed, although not as obvious as with 0.5M Tris and hydroxylamine combined (lane 11), several bands were less sharp or absent for the 45 min incubation with 1.25% hydroxylamine (lane 9) as compared with a shorter 15 min incubation (lane 8). We conclude that the combination of Tris and hydroxylamine treatment shows decreased crosslinked proteins relative to control, or to either treatment alone.

These samples were then subjected to MS-based proteome characterization. The extracts from all 11 samples shown in Fig. 1B were trypsin digested, C18 cleaned and analysed by single-shot LC-MS/MS (Orbitrap Elite). Formaldehyde treatment produces chemically modified and methylene-bridged peptides [14], which are not identified with typical MS database search parameters. We were thus surprised to observe no significant differences observed in protein and peptide coverage between fixed and fixed+decrosslinked samples (Supplementary Table 2). We then hypothesized that formaldehyde-induced modifications were present in exceptionally low stoichiometry and therefore any differences between the samples were masked by the relatively low peptide coverage in the single-shot analyses. We therefore chose three samples for HPLC pre-fractionation and deeper proteome analysis: control protein extract from non-fixed cells, protein extract from fixed cells, and fixed and heat-treated protein extract from fixed cells (95 °C for 45 min). For reference, these samples correspond to lanes 1, 2, and 5, respectively, in Fig. 1B. Fig. 1C shows that the numbers of peptides identified are similar among all three samples; in total, 73,885,

72,785, and 72,779 peptides for control, decrosslinked and fixed samples, respectively. The numbers of proteins detected are similarly comparable (Fig. 1D), indicating that formaldehyde fixation has no measurable impact on proteome coverage.

We next used error-tolerant MS searches (MSFragger [15] and Data-dependent search in MaxQuant) to seek for peptides chemically modified by formaldehyde. Previous reports on short peptides have shown that formaldehyde produces +30 and +12 mass shifts, corresponding to methyloyl and imine modifications, respectively. We saw no appreciable increase in these mass shifts, which is consistent with the instability of these modifications in acid. Indeed, the pattern and frequency of detected mass shifts are remarkably similar between control and fixed samples (results for MaxQuant are shown in Fig. 1E; MSFragger results are shown in Supplementary Figure 1).

From these observations, we concluded that under our reaction conditions, the stoichiometry of crosslinking and chemical modification by formaldehyde is sufficiently low such that the non-detection of modified and crosslinked peptides is not detrimental for characterization of proteomes to a depth of at least 8,000 proteins.

The 'in-cell digest': direct protease digestion of fixed cells

Our observation of little to no significant impact of formaldehyde crosslinking on the MS-based proteomic analysis of fixed cell extracts led us to test whether fixed cells themselves would make suitable substrates for direct protease digestion. Digestion of fixed cells would significantly simplify the sample processing workflow by making several steps, including detergent homogenization and heat treatment, unnecessary. We therefore treated fixed, permeabilized cells suspended in DPBS with either mock treatment (DPBS), or trypsin, and monitored cell morphology by brightfield microscopy. As shown in Fig. 2B, prominent structural features visible in control cells, such as plasma membranes, nuclei and nucleoli, are degraded in a time-dependent manner with trypsin treatment (see Supplementary Video 1). For LC-MS/MS analysis, fixed cells were also pre-incubated with benzonase to digest RNA and DNA oligonucleotides, which may interfere with downstream sample processing. The peptide-containing supernatant from the digest was then subjected to C18 purification prior to analysis by LC-MS/MS. As the digestion occurs within the fixed cells, we have called this approach an 'in-cell digest'. As shown in Fig. 2C, the proteome coverages are similar for fixed

cells processed by the in-cell digest method (~4,678 proteins, n = 3), fixed samples that were subjected to the previously published PRIMMUS protocol (~4,446 proteins, n = 3) and extracts from non-fixed cells processed by precipitation (see Methods, ~4,561 proteins, n = 3). We conclude that the proteome coverage from the in-cell digest is similar, or higher, than the other protocols tested.

We did not observe a broad bias in quantitation, as label free intensities measured in fixed cells prepared by the in-cell digest and by decrosslinking followed by an in-solution digest showed high correlation (Fig. 2D, $\rho = 0.96$). Similarly, a high correlation was observed between fixed cells prepared by the in-cell digest and non-fixed cells (Fig. 2E, $\rho = 0.97$). However, some points lie off-diagonal (Figs. 2D and 2E), suggesting that a small proportion of proteins show a difference in intensity between methods. We next asked whether the sample preparation method systematically affected the abundance of specific proteins, and if so, whether these proteins reflect particular protein classes. Volcano plots comparing in-cell versus in solution methods of preparing fixed cells are shown in Figs. 2F and 2G, which highlight proteins with reproducibly decreased and increased abundance, respectively. Interestingly, RNA-binding proteins, such as proteins involved in mRNA processing, are enriched amongst proteins showing decreased abundance with the in-cell digest. In contrast, membrane-associated proteins are enriched amongst proteins showing increased abundance.

We conclude that the measurements of protein abundance from the in-cell digest are quantitative, reproducible and broadly comparable to conventional sample preparation methods. We note that each sample preparation method will have its own specific biases. In the case of the in-cell digest, the increased abundance of membrane proteins may more accurately reflect the abundance of these proteins in cells, as will be detailed in the Discussion section.

Averaged MS1 Precursors with Library matching (AMPL) improves feature detection

To increase the sensitivity and detection speed of the Orbitrap Elite MS instrument (release date in 2011), we utilised MS1-based identification and quantitation using accurate mass and retention time matching, as proposed originally by the Smith lab [16]. This approach has been recently demonstrated to be highly sensitive in an implementation called BoxCar [17]. The BoxCar method increases the dynamic range of

trap-based MS by collecting ions using segmented, spaced windows. Peptide identification relies on MS1 feature matching to a reference library generated from a fractionated reference sample using the MaxQuant function 'Match-between-runs' (MBR). The library is analysed separately using data-dependent acquisition (DDA) and peptides are identified by MS2 and database searches. Using BoxCar enabled quantitation of ~7,775 on average in single shot analyses of 1 µg HeLa digest on column using the Orbitrap HF.

As the BoxCar method cannot be directly implemented on the Orbitrap Elite, we developed a different approach to increase the dynamic range of MS1 feature detection. MS1 spectral averaging is frequently performed in direct infusion MS, but rarely employed in LC-MS bottom-up proteomics. We surmised that averaging several MS1 scans would improve signal-to-noise (S/N) and would rapidly plateau as it is known that averaging improves S/N by a factor of sqrt(n) where n is the number of spectra averaged. Features would then be matched between the single shot analyses to a fractionated reference library (Fig. 3A). We call this method Averaged MS1 Precursors with Library matching (AMPL), or AMP if no library is used. As shown schematically in Fig. 3B, like BoxCar, AMP(L) prioritises MS1 scans over MS2 scans as compared with DDA and includes top-5 DDA MS/MS scans to ensure identification of features for accurate retention time alignment throughout the chromatographic separation.

We therefore tested AMPL by analysing high on-column loads of tryptic digests of MCF10A (1 µg peptide). Fig. 3C shows the number of chromatographic features ('peaks') detected using DDA (triangle) versus AMP with variable number of MS1 scans to average (circles). AMP with 1 MS1 scan already increases the number of features detected (228,724) versus DDA (188,928), with further gains up to 5 averages (285,049). We then generated a 12-fraction reference library using high pH reverse phase fractionation for MBR. Fig. 3D and E shows a comparison of AMPL and DDA on peptide and protein quantitation, respectively. Increasing the frequency of MS1 scans with no averaging (i.e. AMPL with 1 MS1) performs similarly to DDA with matching to a library (DDA-L) (Fig. 3D, 47,496 vs 46,455 peptides quantitated, respectively). As shown in Fig. 3D, Increasing the number of MS1 scans averaged to 3 and 4 increases the number of peptides quantitated to 51,812 and 54,169 peptides, respectively,

corresponding to ~17% increase in peptides relative to DDA. AMPL with 4 averages also increases proteins quantitated by ~12.5% (Fig. 3E, 5,970 DDA versus 6,724 AMPL).

We reasoned that the additional peptides detected by AMPL originate from low-abundance features detected by virtue of the S/N increase due to averaging. Fig. 3F compares the peptide intensity distributions between DDA-L and AMPL. The distributions are bimodal, with MS/MS-dependent identification biased towards higher intensity features (cyan). Consistent with the idea that AMPL improves S/N, AMPL detects a higher number of matched features (pink) in the low abundance regime.

MS1-based matching approaches have been previously shown to improve dataset completeness by reducing missing values. Ten technical replicates were analysed by DDA, DDA-L and AMPL. Fig. 3G shows a histogram of proteins binned by the number of replicates where an intensity was measured. AMPL shows the highest data completeness (4,500 proteins with intensities measured in all 10 replicates) as compared with DDA-L (3,500 proteins) and DDA (2,900 proteins). Compared to DDA-L, AMPL consistently quantitates more proteins (Supplementary Figure 2) and shows slightly improved reproducibility as indicated by higher pairwise Pearson's correlation scores (Supplementary Figure 2).

The improvements in detecting low abundance features suggested that AMPL may be well suited to analysis of low sample loads. AMP (i.e. no library) consistently detects more features than DDA (Fig. 4A), which leads to significant improvements in the peptide and protein coverage (Figs. 4B and C, respectively). For example, at 10 ng loading, 21,483 unique peptides are quantitated by AMPL versus 14,702 by DDA-L, representing a 46% increase in coverage. AMPL provides 150-535% improvement relative to conventional DDA with no library and 24-46% improvement relative to DDA-L for protein coverage at all tested column loads with greatest gains observed at low column load. At 10 ng, 832 proteins are detected by DDA versus 2,891 proteins by DDA-L and 3,629 proteins by AMPL.

MS1-based matching significantly increases the sensitivity, coverage and data completeness of MS-based proteomics, as previously reported and shown here [18]. However, the lack of MS2-based identification for these matched sequences could lead

to an increased false discovery rate (FDR). We estimated that the matching FDR is ~4.5% using an empirical target-decoy approach where decoy proteomes created by chemical modification (demethylation and isopropylation) are matched against an unmodified library (Supplementary Figure 2A). By applying more stringent thresholds for match time, match m/z and match m/z error (Supplementary Figure 2B-D), we reduced the estimated FDR to 2.2% while retaining 96% of the matches in the target dataset (Supplementary Figure 2E and Supplementary Table 3).

We conclude the library matching approach dramatically increases sensitivity, particularly for low column loads, with AMPL providing the highest peptide and protein coverages overall with relatively low estimated match FDR (<3%).

An improved PRIMMUS for proteomic analysis of low cell number populations

As shown in Fig. 4C, AMPL detects a slightly higher number of proteins in 10 ng on-column load as DDA with 1 µg load, demonstrating a 100x increase in sensitivity. A 10 ng on-column load is equivalent to the protein content of ~67 cells based on the protein per cell measured in bulk assays. However, the effective number of cells required for proteome analysis is usually much higher. This is due to losses during sample preparation. We reasoned that these losses are significantly reduced using the streamlined in-cell digest.

We combined in-cell digest with AMPL to analyse FACS collected TK6 cells, a human lymphoblastoid cell line with a stable near-diploid karyotype. Notably, TK6 cells are smaller than typical adherent human cell lines, such as HeLa and MCF10A. Being cultured in suspension, TK6 cells are amenable towards cell separation techniques, including fluorescence-activated cell sorting (FACS) and centrifugal elutriation, without requiring cell dissociation, which can induce physiological perturbations.

Fig. 4D shows the result of a cell titration analysis of S-phase cells performed in duplicate whereby two aliquots at each indicated cell number (2,000 cells to 0 cells) were collected by FACS from the same starting cell population (Supplementary Table 4). Over 4,500 proteins were quantitated with 2,000 cells, with 4,480 proteins reproducibly quantitated in two technical repeats. At the lower end of the cell titration (shown in Fig. 4E), over 300 proteins on average were quantitated from 10 cells with 259 reproducibly quantitated in two cell aliquots that were separately collected by FACS. While approx.

30 proteins were detected in single cells, with 17 reproducibly detected, nearly all of these proteins were also detected in the background samples ('0 cells').

We conclude that combining in-cell digest and AMPL enables characterization of proteomes of 2,000 cells to a protein depth comparable to conventional single shot DDA analysis of 1 µg on-column loads. The advanced PRIMMUS method presented here significantly reduces the number of cells required, i.e. ~10^3 versus ~10^5.

High temporal resolution analysis of an unperturbed cell cycle using PRIMMUS

Scheduled degradation during mitosis is a key regulatory mechanism to control mitotic progression. Here, we characterised the proteome variation across 16 cell cycle subpopulations, including 8 interphase and 8 mitotic phases, using the advanced PRIMMUS workflow that combines the in-cell digest and AMPL (Fig. 5A). TK6 cells were immunostained for DNA content, cyclin B, cyclin A and histone H3 phosphorylation (Ser28), which are all markers of cell cycle progression, and separated into cell cycle populations (P1 – P16). Fig. 5B summarises the populations collected by FACS with respect to the markers used (Supplementary Figure 3 shows the full gating strategy). The marker combinations were chosen to produce a pseudo-timecourse. Biochemical differences are used as a surrogate for time and cell cycle progression. Based on past literature [18][19] and our previous data [5], we have correlated these biochemical changes with specific cell cycle states (as illustrated in Fig. 5B, bottom). While a full description of the gating strategy can be found in the Experimental Procedures, as an example, cyclin A and cyclin B levels are used to discriminate mitotic subphases, as they are degraded during prometaphase and the metaphase-to-anaphase transition, respectively.

The rarest target cells, in late anaphase of mitosis, are 0.01% of a typical asynchronous TK6 culture. Proteome characterisation of these cells, previously challenging due to lack of sensitivity, is now possible with the latest developments to PRIMMUS. Four separate cultures of TK6 cells were independently FACS separated into 16 populations. For each population, 5,000 cells were collected and processed using the in-cell digest. Collection of 5,000 cells provided sufficient material for duplicate injections for LC-MS/MS analysis by AMPL with DDA feature libraries generated from

interphase, mitotic and asynchronous cells. The data were then processed by MaxQuant with MBR and filtered by match parameters as discussed above.

Of the 7,757 proteins quantitated overall (Supplementary Table 5), 4,918 proteins were quantitated in all 8 replicates (4 biological x 2 technical repeats) in at least one population (Fig. 5C). Next, to identify cell cycle regulated proteins, we treated each set of 16 populations as an ordered series of related biochemical states. We have called each set a pseudotimecourse. While these states can be projected onto a temporal axis (i.e. cell cycle progression), the link with time is indirect as the duration of each phase has been shown to vary substantially on a per-cell basis. We then performed a Fisher's periodicity test to identify proteins abundance patterns that showed periodic behavior. In order to increase robustness, the periodicity test was separately performed on each technical repeat. Only those proteins showing a p-value <= 0.10 and a periodic frequency of 0.0625 or 0.125 (i.e. one cycle every 8 or 16 pseudo-timepoints) in both tests were considered further as periodic. Fig. 5D shows the abundance profiles for heat shock protein HSP90AA1 and ATPase AAA domain-containing protein ATAD2 as example non-periodic and periodic proteins, respectively. ATAD2 shows highly reproducible abundance variation in all 8 pseudotimecourses, with peak abundance in S-phase populations (P5-P6). We note proteins meeting the significance cutoffs are highly enriched in cell cycle GO terms (Supplementary Table 6) and contain many proteins previously reported as cell cycle regulated [5][20][21][22][23], suggesting that the cutoff criteria achieved a final set of proteins with high functional specificity.

Amongst these 119 proteins are cyclins A2 and B1. The MS-measured abundance patterns (Fig. 5E) show similarity with those measured by immunostaining (Fig. 6B) with accumulation in interphase and decreased abundance in mitosis. We also detect cyclin B2, an isoform of cyclin B that is localized to the Golgi apparatus. Cyclins B2 and B1 show a nearly identical abundance pattern in interphase. However, at anaphase and late mitosis (P13 – P16), cyclin B2 abundance does not decrease to background levels, which suggests that unlike cyclin B1, there is a pool of cyclin B2 that is stable towards degradation (Fig. 5E, right).

Hierarchal clustering of the 119 proteins (Fig. 6A) identified five major classes of protein abundance patterns (Fig. 6B). Cluster 1 proteins show high abundance in

552 interphase, which decreases in early mitosis (P8-P10) and recovers slightly in late
553 mitotic populations (P15-P16), as illustrated by the example protein hepatoma-derived
554 growth factor, SRSF6 (Fig. 6C). Like SRSF6, most proteins in this cluster are either
555 RNA- or DNA-binding (26 / 33). For example, several mRNA splicing factors are in this
556 group, including serine/arginine-rich proteins (SRRM2, SRSF2, SRSF3, SRSF5,
557 SRSF6). These proteins decrease in abundance in mitosis with a small fold change (≤ 2)
558 compared with, for example, cyclin B1 (Fig. 5E). The remaining proteins with no known
559 or anticipated oligonucleotide-binding properties are enriched in cytoskeleton-binding
560 factors, e.g. the actin-binding proteins MARCKS and ZYX.
561 Cluster 2 contained proteins that had peak abundances in late G1/S populations.
562 Included in this cluster is the protein SLBP, a histone gene expression factor that peaks
563 in P4-P5 (Fig. 6D). Indeed, nearly all proteins in this cluster are directly involved in DNA
564 replication or associated with the DNA damage response, including members of the
565 MCM helicase (MCM2, MCM5, MCM6), DNA damage checkpoint factors (ATM, RBBP8)
566 and a replication-dependent histone chaperone (CHAF1B).
567 These protein clusters vary in their enrichment in short linear (sequence) motifs
568 (SLIMs). SLIMs mediate protein-protein interactions that lead to changes in post-
569 translational modification, stability and/or subcellular localization of a protein. Using the
570 eukaryotic linear motif (ELM) database [24], we identified SLIMs that are enriched in
571 each cluster ($p < 0.01$, Fisher's exact test, Supplementary Table 7). In particular,
572 enriched SLIMs that modulate protein stability (i.e. degrons) could help identify relevant
573 degradation pathways for each protein cluster. Clusters 1 and 2 show an enrichment in
574 Skp1-Cullin-F box (SCF)-Fbxw7 motifs (Fig. 6I). Targeted degradation by SCF-Fbxw7
575 generally requires priming phosphorylation [25], which links the stability of a protein with
576 kinase activity, e.g. CDK. Cluster 1 is most enriched in the T-P-X-X-[ST] motif (2' motif),
577 which requires two priming phosphorylations for recognition by SCF-Fbxw7 and
578 targeted degradation. Cluster 2 is most enriched in the T-P-X-X-E motif, which requires
579 only one phosphorylation for substrate recognition. Interestingly, Cluster 1 is also more
580 highly enriched in CDK consensus sites. We conclude that multisite phosphorylation by
581 CDK may play a role in directing these proteins for degradation by SCF-Fbxw7.

Cluster 3 shows peak abundance in G2 and early mitosis (P6 to P9). This cluster contains several proteins associated with DNA replication and DNA damage repair, including the dsDNA exonuclease EXO1, PCNA-associated factor (PAF/KIAA0101) and ribonucleotide reductase M2 (RRM2, Fig. 6E). The abundance pattern of RRM2 is consistent with previous proteomic studies and targeted degradation of RRM2 in late G2/early mitosis by the cyclin F-SCF complex [5][26].

Clusters 4 and 5 show peak abundance during mitosis and contain the largest proportion of proteins with either known roles in mitotic progression or targeted for degradation in mitosis (9/12 for cluster 4, 38/46 for cluster 5). The distinguishing feature that separates cluster 4 and 5 is the mitotic abundance pattern. Cluster 4 proteins show decreased abundance in earlier mitotic populations, particularly in P11 – P12, coincident with the onset of cyclin A2 and cyclin B1 degradation (c.f. Fig. 5B). The three mitotic cyclins detected (A2, B1 and B2), the spindle assembly checkpoint kinases BUB1 (Fig. 6F) and BUB1B (BubR1), the kinesin-8 family member KIF18B, securin (PTTG1) and shugoshin (SGO2) are in this cluster. Functionally, this cluster is characterized by proteins that maintain sister chromatid cohesion (securing, shugoshin) and constitute a checkpoint that prevents anaphase (cyclins, Bub kinases) while proper microtubule attachment and biorientation of chromosomes takes place. Consistent with the MS-based proteomics data for KIF18B, the fluorescence of GFP-KIF18B begins to decrease 10 minutes prior to anaphase onset and is at 20% relative fluorescence by 20 minutes after anaphase onset [27].

By contrast, cluster 5 proteins show a significant increase in abundance at the end of interphase (P7 – P8) with peak abundance throughout mitosis (P9 – P15) and a significant decrease only in the last population (P16), i.e. cells undergoing mitotic exit. Example proteins from this cluster include TPX2 and Aurora A kinase (Fig. 6G). TPX2 is the activator of Aurora A kinase whose activity is important in centrosome separation in prophase and mitotic progression. Other proteins in cluster 5 with regulatory roles in mitotic progression include the catalytic E2 subunits of the APC/C (UBE2C, UBE2S), the chromosome passenger complex (AURKB, INCENP, BIRC5 – Survivin, CDCA8 – Borealin) and the spindle-associated protein FAM83D.

SLIM analysis of these clusters identified differences in the enrichment in nuclear import and export signals. As shown in Fig. 6I, clusters 1 and 2 are enriched in nuclear localisation signals (mono- and bi-partite). By contrast, cluster 4 shows a strong enrichment for the Crm1-mediated nuclear export signal (NES). Eight proteins in cluster 4 matched the NES consensus. Some predicted NES located in globular domains will likely be constitutively inaccessible to Crm1 but may be recognized upon conformational change. Notably, cluster 4 includes cyclins B1 and B2, whose constitutive export from the nucleus is thought to be important in preventing premature mitotic entry. Crm1-binding and/or exclusion from the nucleus of the remaining six proteins (e.g. Bub1, BubR1, cyclin A2, CLEC16A, MVP, and ARMC1) may also be important in the proper timing of cell cycle events.

We identified strongly pseudoperiodic proteins that have no reported function in cell cycle control. These novel cell cycle regulated proteins may, like many of the other proteins identified in this manner, have significant roles in cell cycle progression. These candidates include EXO1, the DNA helicase PIF1, the guanine-exchange factor NET1 and the uncharacterized protein FAM111B (Fig. 6H). Potential functional roles for FAM111B in cell cycle regulation are discussed further below.

Analysis of mitotic protein abundance dynamics in unperturbed cells

A major regulator of protein abundance during the cell cycle is the anaphase promoting complex/cyclosome (APC/C). The APC/C is an E3 ubiquitin ligase and is active during the mitotic and G0/G1 phases of the cell cycle [28][29]. Its substrates include key regulators of the cell cycle, including cyclin A2 and cyclin B1 [18][19]. However, ubiquitination of APC/C substrates is tightly temporally controlled, with APC/C substrate specificity changing during the cell cycle. This is mediated through changes in the APC/C co-activators and substrate recognition factors, Cdc20 and Cdh1. While APC/C-Cdc20 is active in early mitosis, the substrate receptor changes to Cdh1 in late mitosis, thereby conferring a temporal order to substrate degradation. Cdc20 is itself a substrate of the APC/C-Cdh1, allowing for switch-like handover in substrate receptor control.

Interestingly, 25 of the 119 core pseudoperiodic proteins are experimentally validated APC/C substrates and the vast majority (24) are found in clusters 3, 4 and 5. Substrate recognition by APC/C-Cdc20 and APC/C-Cdh1 is mediated by the interaction

643 between WD40 domains on the APC/C-(Cdc20/Cdh1) and SLIMs found on substrates.
644 The KEN and D-box (RxxL) degrons are well documented SLIMs that bind both APC/C-
645 Cdc20 and APC/C-Cdh1, with APC/C-Cdh1 having a preference for the KEN degron.
646 More recently, a third SLIM called the ABBA motif was shown to be important in
647 substrate recognition by APC/C-Cdc20 [30]. Its name comes from the four proteins in
648 which it is found: Cyclin *A*, *B*ub1, *B*ubR1 and the yeast-specific protein *A*cm1.
649 Fig. 6J shows the enrichment profile of these SLIMs across the six clusters. The
650 KEN motif is comparably enriched in 4 out of the 5 clusters (Fig. 6J), with highest
651 enrichments for the mitotic phase-peaking clusters (clusters 4 and 5). The frequencies
652 range from 25% of the proteins in a cluster having the KEN motif (cluster 2) to 43%
653 (cluster 5), representing a 3-5-fold enrichment over the background frequency (8%). All
654 four clusters show low to non-detectable abundance in P16, P1 and P2, i.e. mitotic exit
655 and G0/early G1 when APC/C-Cdh1 is active. In total, 35 cell cycle regulated proteins
656 contain a KEN SLIM, approximately 50% (18 proteins) that have been experimentally
657 characterized as APC/C substrates. The remaining uncharacterized 17 proteins are
658 excellent candidates to be APC/C-Cdh1 substrates. Consistent with this prediction,
659 cluster 1, which is the only cluster showing no enrichment for the KEN motif, contains
660 proteins that have on average, higher abundance in G0/early G1.
661 Six out of 12 proteins that peak in mid-mitosis (cluster 4) contain the RxxL D-box
662 sequence. The 50% frequency is ~8-fold higher than the background frequency (6%).
663 By contrast, the fold-enrichment is considerably lower in the other clusters (Fig. 6I).
664 Similarly, 5 out of 12 proteins contain the ABBA motif (42%, Fig. 6I), representing a ~9-
665 fold enrichment over the background frequency (5%). D-box and ABBA motif-containing
666 proteins in this cluster are mostly mutually exclusive: only two proteins contain both
667 SLIMs: BubR1 (BUB1B) and shugoshin-2 (SGOL2). Of the D-box and ABBA motif
668 containing proteins, three have not been previously experimentally characterized as
669 APC/C substrates: SGOL2, MVP and CLEC16A.
670 This high-resolution analysis of mitotic proteomes identified three clusters that
671 differed significantly in their protein abundance patterns and enriched degrons. ~50% of
672 these APC/C-degron containing proteins are experimentally validated APC/C substrates.
673 This is now accompanied with measurement of their protein abundance variation in an

unperturbed mitosis in human cells without fluorescent protein tags. We identify ~20 protein candidates as novel APC/C substrates. As shugoshin-2 (SGOL2) plays key roles in protecting sister chromatid cohesion [31], direct degradation by APC/C-Cdc20 during prometaphase could have major implications on how cohesion is lost at chromosome arms during early mitosis.

Proteomic assignment of cell cycle states

MS-based single cell proteome analysis is an emerging area. Recent advances in sample preparation, including NanoPOTS [4][5] and the in-cell digest described here, suggest that routine proteome analysis of single somatic mammalian cells will be possible in the near future. In comparison, single cell transcriptomics as a mature field with commercial kits now available. In single cell RNA-seq (sc-RNAseq) analysis [32], the deconvolution of cell cycle state has been critical [33][34]. This is because cell cycle variation contributes significantly to the variation observed in a cell population. For example, to identify cell fate trajectories during differentiation, researchers relied on reference cell cycle regulated genes in order to identify the effect of cell cycle variation in the gene expression differences observed [35]. A highly validated reference set of cell cycle regulated proteins will be important for the biological interpretation of single cell proteomic datasets.

We tested whether the abundances of the core 119 cell cycle regulated proteins determined in this study were sufficient to assign specific cell cycle states to cellular proteomes. The abundance patterns for the 119 proteins for each sample (16 timepoints x 8 replicates = 128 samples) were subjected to principal component analysis (PCA). The two major principal components, PC1 and PC2, explain 53% and 20.5% of the variance, respectively, as shown in Fig. 7A. Interphase (circles) and mitotic (triangle) are separated predominantly along PC1. To a lesser extent, subphases within each (for example, see arrows indicating P1 and P2) are separated along both principal components. Moving counterclockwise, starting from the top right for P1, the samples clearly follow a trajectory that reflects the position of each sample in the cell cycle, starting from early G1 (P1 and P2) to mitosis (left side, triangles). Telophase/cytokinesis populations (P16, pink triangles) are situated between the other mitotic populations and P1. Detection of relevant features is essential as PCA analysis of the entire proteome

dataset does not result in cell cycle separation. Repeating the PCA analysis with cyclin A2 and cyclin B1 removed essentially produces identical results, which indicates that the relationships produced by using ~119 cell cycle marker proteins are robust towards the absence of individual proteins, including key proteins that drive cell cycle progression.

A simple kNN-model was used to classify cell cycle states using these data. Replicates 1 - 3 were used as a training set and replicate 4 was used as the test set. Fig. 7B shows the performance of the classification by plotting predicted versus actual populations. There is a linear correlation with some minor deviations. We then repeated this kNN analysis for each combination of the four replicates using 3 replicates as the training set and the remaining replicate as the test set. We treated the populations as a circular, progressive series of cell states whereby P1 is the next state after P16. We then calculated the distance between predicted and actual populations for each replicate combination (Supplementary Figure 4A). The average distance is 0 in all four cases with a standard deviation ~1. This indicates that the kNN models are broadly accurate in predicting the cell cycle state with a precision of ± 1 cell state.

We then asked whether the PCA classification could be used to identify uncharacterised populations. During the FACS separation to collect the 16 populations described above, we noticed the presence of an unexpected, rare population (P17) in two out of the four biological replicates. As shown in Fig. 7C, these cells are similar to G2-phase cells in DNA content (i.e. 4N DNA content) but significantly differ from most G2-phase cells by having low cyclin B1 levels. P17 cells were analysed alongside the other 16 populations using the in-cell digest and AMPL. As shown in the simplified PCA where replicate data were averaged (Fig. 7D), the PCA places P17 between P16 and P1. Indeed, using the kNN model described above classifies two replicates of P17 as P16 and two other replicates as P1. From these observations, we conclude that P17, while having DNA content consistent with a G2-phase cell, has a cell cycle protein profile that is more consistent with an early G1/G0-phase cell.

This population may reflect a senescent state consistent with previous reports where the APC/C is re-activated in G2 phase cells in response to DNA damage, leading to premature degradation of cyclin B1. Our data suggest that numerous APC/C targets

are decreased relative to a typical G2 state (Fig. 7E), effectively resetting the cell cycle state of these cells to an early G1/G0-like state. We then performed pairwise comparisons proteome-wide between P1, P16 and P17 to identify proteins showing reproducible changed abundance in P17 cells. Of the top candidates (Supplementary Figure 4B), three are regulators of the DNA damage response (Fig. 7F): BMI1, HINT1 and PPP6R2. BMI1 and HINT1 show increased abundance in P17 compared to P1 and P16. Both proteins are recruited to sites of DNA damage and loss of either protein leads to defective repair [36][37]. PPP6R2 is a regulatory subunit of the protein phosphatase PP6. PP6 is involved in silencing the DNA damage response by dephosphorylation of γ-H2AX [38]. These results support the idea that P17 is a DNA damage-induced senescent state.

We conclude that we have found a core set of 119 proteins that can be used to robustly assign cell cycle states with high resolution and to phenotypically characterise cell populations whose position in the cell cycle is unknown.

Discussion

A major challenge with the comprehensive analysis of proteomes from low cell number samples is sample preparation. An on-column load of 200 ng peptide, the equivalent to the protein content of approximately 2,000 TK6 cells, is sufficient material to obtain proteome coverage of >4,000 proteins with current instrumentation. Removal of detergents used to produce soluble cell extracts by use of membrane filters, organic precipitation (with or without the aid of magnetic beads) or SDS-PAGE gel extraction are protocols involving many steps and repeated exposure to new plastic surfaces that introduce opportunities for non-specific peptide and protein adsorption. Here, we have presented a minimalistic approach for preparing cells for proteomics called the ‘in-cell digest’. Cells are fixed with formaldehyde and methanol to effectively trap them in biochemical states, then directly digested with trypsin and desalted prior to LC-MS/MS analysis.

We show that the in-cell digest enables reproducible and quantitative analysis of proteomes from 2,000 TK6 and MCF10A cells using AMPL analysis. The AMPL approach overcomes the low duty cycle of the Orbitrap Elite to enable proteome analysis with a sensitivity comparable with current instruments. Newer instrumentation

with higher duty cycles, including the TIMS-TOF Pro and Exploris 480, is expected to enable conventional DDA analysis of proteomes at a similar depth with 2,000 TK6 cells, or alternatively, improve proteome depth further using MS1-based matching methods.

The in-cell digest is compatible with other approaches of low cell number sample preparation for MS-based proteomics. In-cell digested samples can be efficiently labelled by isobaric tags, e.g. TMT and iTRAQ, and therefore compatible with use of carrier channels to boost the signal of rare or single cell channels (e.g. iBASIL). The protocol requires no specialized humidified sample handling chambers or direct loading onto premade, single-use analytical nanoLC columns, such as those described in the nanoPOTS workflow. While the proteome coverages obtained by nanoPOTS is higher than reported here, it is possible that a new workflow combining aspects of the in-cell digest and nanoPOTS could improve both generalizability and performance compared to either method as originally described.

Each sample preparation method will have its unique advantages and potential biases, which we evaluated by quantitatively comparing the in-cell digest with a more conventional in-solution digest. This analysis revealed an overrepresentation of membrane proteins amongst those proteins with higher abundance measured in the in-cell digest samples. These proteins include mitochondrial membrane proteins (e.g. TOMM7) and proteins that are known to be localized to the cell surface (ADAM15). Membrane proteins have been shown to irreversibly aggregate in soluble extracts when heat-treated and precipitated. Delipidation by methanol, which is used to increase cell permeability, could also play an important role in increasing digestion efficiency by trypsin. We suggest that the higher abundances measured for membrane proteins is unlikely to be an artefact of the in-cell digest; in contrast, the measurements are likely to more accurately reflect the abundances of these proteins in cells.

By contrast, RNA-binding proteins, including snRNP proteins, were overrepresented amongst those proteins with lower abundance in the in-cell digest samples. The lower abundances measured could represent on the one hand, a specific loss of peptide-RNA crosslinks, or on the other hand, non-specific loss of the RNA-binding proteins into the supernatant. Studies are ongoing examining the RNA-binding protein bias observed in more detail, and we have preliminary evidence suggesting the

latter. Interestingly, proteins in cluster 2 (Fig. 6A), which show a robust, pseudoperiodic change in abundance are nearly all known to interact with either DNA or RNA. Few of these proteins have been shown to be cell cycle regulated previously. It may be the changes in MS-measured abundance reflect differences in RNA- and/or DNA-interactions by these proteins rather than a change in the protein abundance in cells.

We identify novel proteins whose cell cycle function has not been previously characterized. FAM111B is a pseudoperiodic protein in cluster 1 (Fig. 6B, right, Fisher's $p_1 < 0.001$, $p_2 = 0.06$), showing peak levels in S-phase populations (P4 – P6), followed by a decrease in G2 populations (P7 – P8). FAM111B is poorly characterized despite its expression being associated with poor prognosis in pancreatic and liver cancers (Human Protein Atlas [39]) and mutation causative for a rare inherited genetic syndrome (hereditary fibrosing poikiloderma with tendon contracture, myopathy, and pulmonary fibrosis) [40]. Interestingly, FAM111A, the only other member of the FAM111 gene family, localizes to newly synthesized chromatin during S-phase, interacts with PCNA via its PCNA-interacting protein (PIP) box and its depletion reduces base incorporation during DNA replication [41]. FAM111B also contains a PIP box (residues 607 – 616). Data from HeLa S3 cells also suggest that FAM111B is a cell cycle regulated protein with peak levels in S-phase [22]. Interestingly, the mRNA abundance and translation rate of FAM111B peaks in G1-phase [22], suggesting that the protein abundance is subject to significant post-translational control. Consistent with this idea, FAM111B contains D-box and KEN-box motifs that are recognized by the APC/C E3 ligase to ubiquitinate targets for proteasomal degradation. Due to the similarity with FAM111A in sequence, predicted interactions with PCNA and peak protein abundance in S-phase, we propose that FAM111B also is likely to play a key role in DNA replication.

We present an unbiased pseudotemporal analysis of protein abundance changes during 8 biochemically resolved mitotic states (P9 to P16 in Fig. 5B) with a resolution extremely challenging to obtain with high precision using arrest and release methodologies. The protein clusters are functionally related. For example, clusters 4 and 5 both contain proteins essential for mitotic progression but differ in when during mitosis the functions are required. Cluster 4 contains proteins directly involved in or directly downstream of the spindle assembly checkpoint that are degraded upon

checkpoint satisfaction. These regulatory pathways ensure that proper spindle microtubule-chromatid attachments are formed prior to loss of sister-chromatid cohesion and separation of the sister chromatids. By contrast, cluster 5 contains proteins that are functional throughout mitosis, such as chromosome passenger complex (CPC), or primarily in cytokinesis, such as ECT2, PRC1, RACGAP1 and ARHGAP11A. Interestingly, several core subunits of the APC/C E3 ligase are also present in cluster 4. Their degradation at the end of mitosis is expected to significantly decrease APC/C-mediated substrate degradation promote accumulation of substrates and facilitate rapid progression into the next cell cycle.

A high proportion of proteins in clusters 4 and 5 (24/69, 35%) are experimentally validated APC/C substrates, which represents a 70-fold overrepresentation in these two clusters compared to non-pseudoperiodic proteins (0.5%). Previous studies have identified APC/C-Cdh1 and APC/C-Cdc20 substrates by bioinformatic analysis of co-regulation, stabilization by siRNA depletion of Cdc20 or Cdh1, and immunoprecipitation of APC/C at different timepoints during mitosis. Interestingly, the high mitotic phase resolution and purity obtained in this study enabled unbiased identification and separation of APC/C substrates. As discussed above, clusters 4 and 5 differ in the representation of ABBA and D-box short linear motifs, key degrons that are recognized by APC/C-Cdc20. Note that there are an additional 44 proteins in these two clusters that have not been previously experimentally validated as APC/C substrates and are candidates for future follow-up analysis as novel, uncharacterized substrates.

High resolution classification of cell cycle state is an important prerequisite to obtaining meaningful biological insights into single cell ‘omics’ data. However, datasets on the cell cycle regulated transcriptome and proteome generally provide low time resolution, particularly in mitosis. This is more important with single cell proteomics. Whereas transcriptional and translational activity are dampened during mitosis, there are tremendous changes in protein phosphorylation and protein abundances, which will contribute towards single cell proteome variation.

Here we have identified a cell cycle signature composed of the abundances from 119 pseudoperiodic proteins that can be used to classify the cell cycle state of a cell population by virtue of the proteome. By using a split train/test strategy, we showed a

kNN model predicted cell cycle state with relatively high accuracy and can provide clues into the phenotype of uncharacterised, rare populations. We anticipate that the high-resolution cell cycle dataset here will be important to understand the biological implications of single cell proteomics data, particularly in systems where cell cycle phase differences are an underlying source of variation but not the primary biological feature of interest.

Formaldehyde fixation is used frequently as a precursor to intracellular immunostaining for cellular analysis and for inactivating cells that potentially harbor infectious agents, e.g. viruses. We have shown that mild formaldehyde treatment is compatible with comprehensive and quantitative proteomics with low cell numbers. We anticipate that the in-cell digest will be broadly applicable to characterise the proteomes of formaldehyde fixed cells. Recently published data suggest that formaldehyde crosslinks can be directly detected from MS data [42]. We anticipate the in-cell digest would enhance the sensitivity of crosslink detection and lead to an increase in identified protein-protein interactions.

Acknowledgements

This work was supported by a Sir Henry Dale Fellowship to TL (Wellcome Trust & Royal Society 206211/Z/17/Z), a Darwin Trust PhD Studentship to ASA, an EASTBIO PhD Studentship to DAL, the Wellcome Centre for Cell Biology (WCB) core facilities (Wellcome Trust 203149), and funding for instrumentation, including equipment grants to the WCB Proteomics Core (091020) and the flow facilities. We thank valuable feedback and discussions with colleagues in the WCB and the University of Edinburgh, including Fiona Rossi (Scottish Centre for Regenerative Medicine), Christos Spanos (WCB) and Shaun Webb (WCB).

Data Availability

Raw MS data and processed MaxQuant output files are available on ProteomeXchange/PRIDE. These data can be accessed using the project accession number PXD020006

References

1. Spencer SL, Cappell SD, Tsai F-C, Overton KW, Wang CL, Meyer T (2013) The proliferation-quiescence decision is controlled by a bifurcation in CDK2 activity at mitotic exit. *Cell* **155**: 369–383.

2. Barr, A. R., Cooper, S., Heldt, F. S., Butera, F., Stoy, H., Mansfeld, J., et al. (2017). DNA damage during S-phase mediates the proliferation-quiescence decision in the subsequent G1 via p21 expression. *Nature Communications*, *8*(1), 14728.
3. Amon S, Meier-Abt F, Gillet LC, Dimitrieva S, Theocharides APA, Manz MG, Aebersold R (2019) Sensitive Quantitative Proteomics of Human Hematopoietic Stem and Progenitor Cells by Data-independent Acquisition Mass Spectrometry. *Mol Cell Proteomics* **18**: 1454.
4. Zhu Y, Piehowski PD, Zhao R, Chen J, Shen Y, Moore RJ, Shukla AK, Petyuk VA, Campbell-Thompson M, Mathews CE, et al. (2018) Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat Comms* **9**: 882.
5. Tsai C-F, Zhao R, Williams SM, Moore RJ, Schultz K, Chrisler WB, Pasa-Tolic L, Rodland KD, Smith RD, Shi T, et al. (2020) An Improved Boosting to Amplify Signal with Isobaric Labeling (iBASIL) Strategy for Precise Quantitative Single-cell Proteomics. *Mol Cell Proteomics* **19**: 828.
6. Ly T, Whigham A, Clarke R, Brenes-Murillo AJ, Estes B, Wadsworth P, Lamond AI (2017) Proteomic analysis of cell cycle progression in asynchronous cultures, including mitotic subphases, using PRIMMUS. *bioRxiv* 125831.
7. Ly T, Endo A, Lamond AI (2015) Proteomic analysis of the response to cell cycle arrests in human myeloid leukemia cells. *eLife* **4**:.
8. Müllers E, Silva Cascales H, Burdova K, Macurek L, Lindqvist A (2017) Residual Cdk1/2 activity after DNA damage promotes senescence. *Aging Cell* **16**: 575–584.
9. Skopek TR, Liber HL, Penman BW, Thilly WG (1978) Isolation of a human lymphoblastoid line heterozygous at the thymidine kinase locus: Possibility for a rapid human cell mutation assay. *Biochemical and Biophysical Research Communications* **84**: 411–416.
10. Rappsilber J, Ishihama Y, Mann M (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**: 663–670.
11. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367–1372.
12. Hoffman EA, Frey BL, Smith LM, Auble DT (2015) Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes. *Journal of Biological Chemistry* **290**: 26404–26411.
13. Bornstein P, Balian G (1977) [14] Cleavage at AsnGly bonds with hydroxylamine. In, *Enzyme Structure Part E* pp 132–145. Academic Press.
14. Metz B, Kersten GFA, Hoogerhout P, Brugghe HF, Timmermans HAM, de Jong A, Meiring H, Hove ten J, Hennink WE, Crommelin DJA, et al. (2004) Identification of Formaldehyde-induced Modifications in Proteins REACTIONS WITH MODEL PEPTIDES. *Journal of Biological Chemistry* **279**: 6235–6243.
15. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat Meth* **14**: 513–520.

16. Pasa-Tolic L, Masselon C, Barry RC, Shen Y, Smith RD (2004) Proteomic analyses using an accurate mass and time tag strategy. *BioTechniques* **37**: 621–639.
17. Meier F, Geyer PE, Winter SV, Cox J, Mann M (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Meth* **15**: 440–448.
18. Pines J, Clute P (1999) Temporal and spatial control of cyclin B1 destruction in metaphase. *Nat Cell Biol* **1**: 82–87.
19. Elzen den N, Pines J (2001) Cyclin a Is Destroyed in Prometaphase and Can Delay Chromosome Alignment and Anaphase. *The Journal of Cell Biology* **153**: 121–136.
20. Ly T, Ahmad Y, Shlien A, Soroka D, Mills A, Emanuele MJ, Stratton MR, Lamond AI (2014) A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *eLife* **3**: e01630.
21. Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., et al. (2010). Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. *Science Signaling*, *3*(104), ra3–ra3. http://doi.org/10.1126/scisignal.2000475
22. Aviner R, Shenoy A, Elroy-Stein O, Geiger T (2015) Uncovering Hidden Layers of Cell Cycle Regulation through Integrative Multi-omic Analysis. *PLoS Genet* **11**: e1005554.
23. Herr, P., Boström, J., Rullman, E., Rudd, S. G., Vesterlund, M., Lehtiö, J., et al. (2020). Cell cycle profiling reveals protein oscillation, phosphorylation, and localization dynamics. *Molecular & Cellular Proteomics : MCP*, mcp.RA120.001938.
24. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, Diakogianni A, Valverde JA, Bukirova D, Čalyševa J, et al. (2019) ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res* **48**: D296–D306.
25. Hao B, Oehlmann S, Sowa ME, Harper JW, Pavletich NP (2007) Structure of a Fbw7-Skp1-Cyclin E Complex: Multisite-Phosphorylated Substrate Recognition by SCF Ubiquitin Ligases. *Molecular Cell* **26**: 131–143.
26. D'Angiolella V, Donato V, Forrester FM, Jeong Y-T, Pellacani C, Kudo Y, Saraf A, Florens L, Washburn MP, Pagano M (2012) Cyclin F-mediated degradation of ribonucleotide reductase M2 controls genome integrity and DNA repair. *Cell* **149**: 1023–1034.
27. Tanenbaum ME, Macurek L, van der Vaart B, Galli M, Akhmanova A, Medema RH (2011) A Complex of Kif18b and MCAK Promotes Microtubule Depolymerization and Is Negatively Regulated by Aurora Kinases. *Current Biology* **21**: 1356–1365.
28. Kernan J, Bonacci T, Emanuele MJ (2018) Who guards the guardian? Mechanisms that restrain APC/C during the cell cycle. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1865**: 1924–1933.
29. Chang L, Barford D (2014) Insights into the anaphase-promoting complex: a molecular machine that regulates mitosis. *Current Opinion in Structural Biology* **29**: 1–9.

30. Di Fiore B, Davey NE, Hagting A, Izawa D, Mansfeld J, Gibson TJ, Pines J (2015) The ABBA Motif Binds APC/C Activators and Is Shared by APC/C Substrates and Regulators. *Developmental Cell* **32**: 358–372.
31. Hellmuth S, Gómez-H L, Pendás AM, Stemmann O (2020) Securin-independent regulation of separase by checkpoint-induced shugoshin–MAD2. *Nature* **580**: 536–541.
32. Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**: 133–145.
33. Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC, Buettner F (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Human Pluripotent Stem Cells* **85**: 54–61.
34. Liu Z, Lou H, Xie K, Wang H, Chen N, Aparicio OM, Zhang MQ, Jiang R, Chen T (2017) Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat Comms* **8**: 22.
35. Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, Amatya S, Madrigal P, Isaacson A, Buettner F, et al. (2020) Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat Comms* **11**: 810.
36. Li H, Balajee AS, Su T, Cen B, Hei TK, Weinstein IB (2008) The HINT1 tumor suppressor regulates both γ-H2AX and ATM in response to DNA damage. *J Cell Biol* **183**: 253–265.
37. Ginjala V, Nacerddine K, Kulkarni A, Oza J, Hill SJ, Yao M, Citterio E, van Lohuizen M, Ganesan S (2011) BMI1 Is Recruited to DNA Breaks and Contributes to DNA Damage-Induced H2A Ubiquitination and Repair. *Mol Cell Biol* **31**: 1972.
38. Douglas P, Zhong J, Ye R, Moorhead GBG, Xu X, Lees-Miller SP (2010) Protein Phosphatase 6 Interacts with the DNA-Dependent Protein Kinase Catalytic Subunit and Dephosphorylates γ-H2AX. *Mol Cell Biol* **30**: 1368.
39. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, et al. (2017) A pathology atlas of the human cancer transcriptome. *Science* **357**: eaan2507.
40. Mercier S, Küry S, Shaboodien G, Houniet DT, Khumalo NP, Bou-Hanna C, Bodak N, Cormier-Daire V, David A, Faivre L, et al. (2013) Mutations in FAM111B Cause Hereditary Fibrosing Poikiloderma with Tendon Contracture, Myopathy, and Pulmonary Fibrosis. *The American Journal of Human Genetics* **93**: 1100–1107.
41. Alabert C, Bukowski-Wills J-C, Lee S-B, Kustatscher G, Nakamura K, de Lima Alves F, Menard P, Mejlvang J, Rappsilber J, Groth A (2014) Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nat Cell Biol* **16**: 281–291.
42. Tayri-Wilk T, Slavin M, Zamel J, Blass A, Cohen S, Motzik A, Sun X, Shalev DE, Ram O, Kalisman N (2020) Mass spectrometry reveals the chemistry of formaldehyde cross-linking in structured proteins. *Nat Comms* **11**: 3128.

Figures and Figure Legends

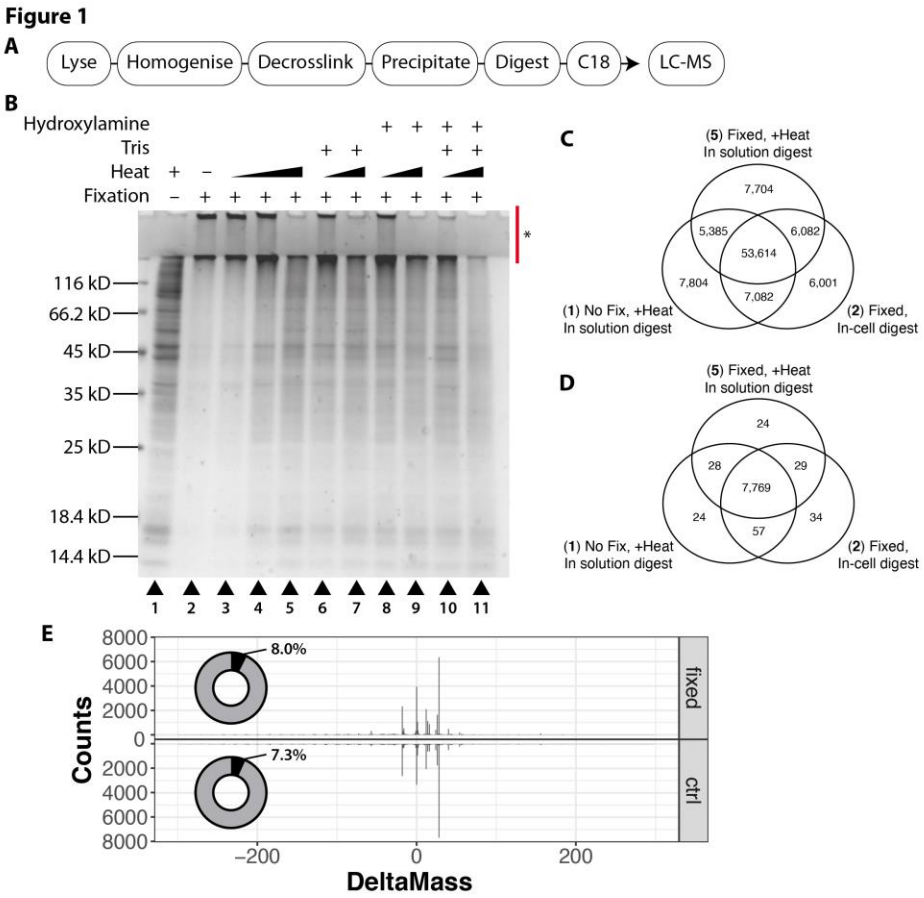


Figure 1. Low level formaldehyde crosslinking has negligible impact on proteome coverage and quantitation. A) Schematic of the sample processing workflow for PRIMMUS. B) Impact of heat, hydroxylamine and/or Tris on decrosslinking efficiency of cells fixed with 2% formaldehyde and 90% methanol as measured by total protein stain and SDS-PAGE. The red bar indicates bands corresponding to high molecular weight, crosslinked proteins. C, D) Comparison of peptides (C) and proteins (D) identified from in-solution and in-cell digests. E) Error-tolerant search for formaldehyde-induced chemical modifications to peptides using 'Data-dependent mode' in MaxQuant.

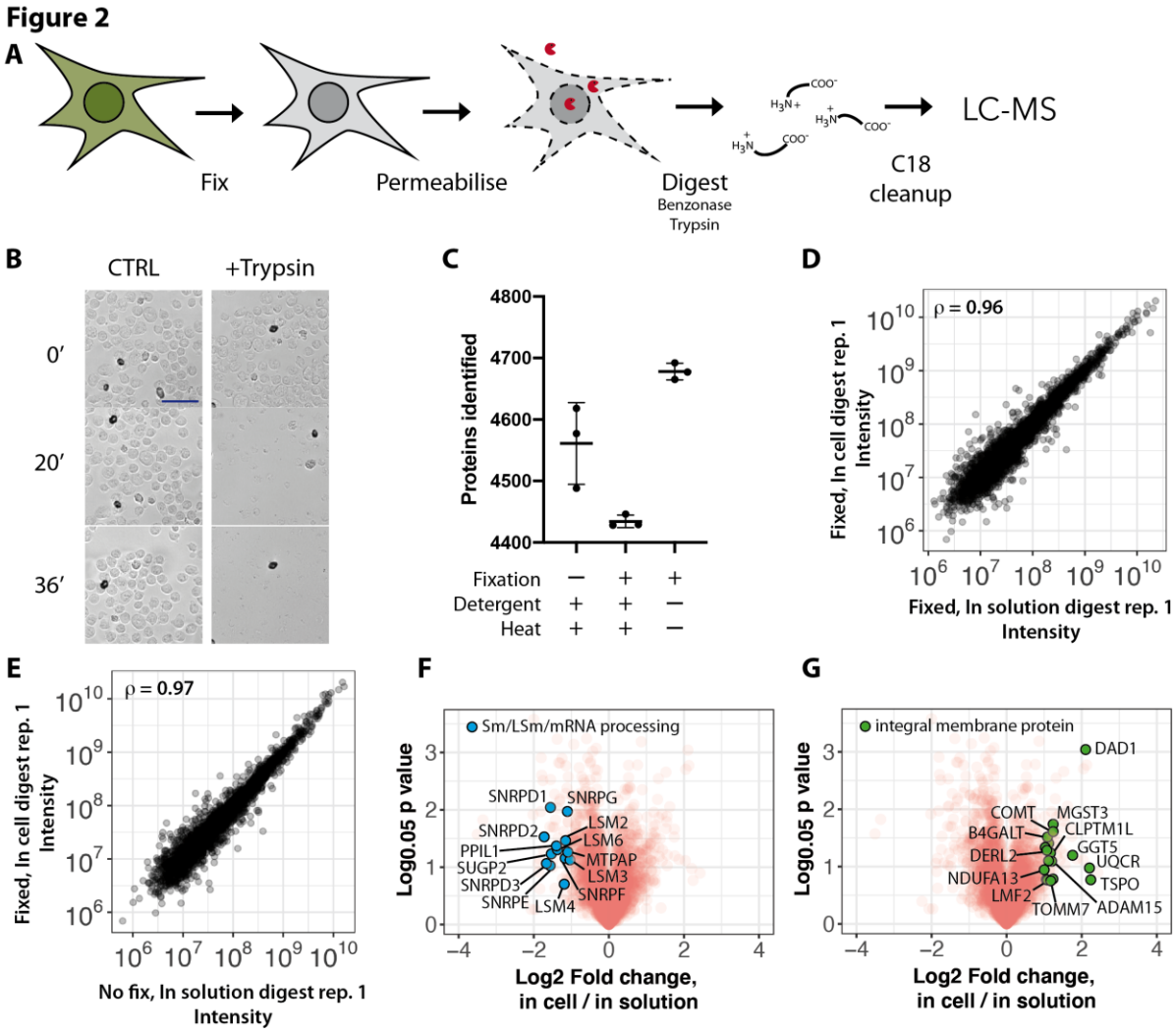


Figure 2. Direct tryptic digestion of fixed and permeabilized cells outperforms protein precipitation and in solution digest of cellular extracts. A) Schematic of the in-cell digest workflow. B) Fixed and permeabilized cells treated either with DPBS (left) or with trypsin (right) were imaged at the indicated times in minutes. Scale bar is 50 µm. C) Comparison of the proteome coverage reproducibility between in-solution and in-cell digests. D, E) Comparison of the intensities measured by in-cell digests and either (D) in-solution digest of fixed cells, or (E) in-solution digest of cells with no fixative. F, G) Volcano plots comparing an in-cell versus in solution digest of fixed cells. Two protein classes enriched amongst proteins reproducibly changing in abundance are RNA processing proteins (F) and integral membrane proteins (G).

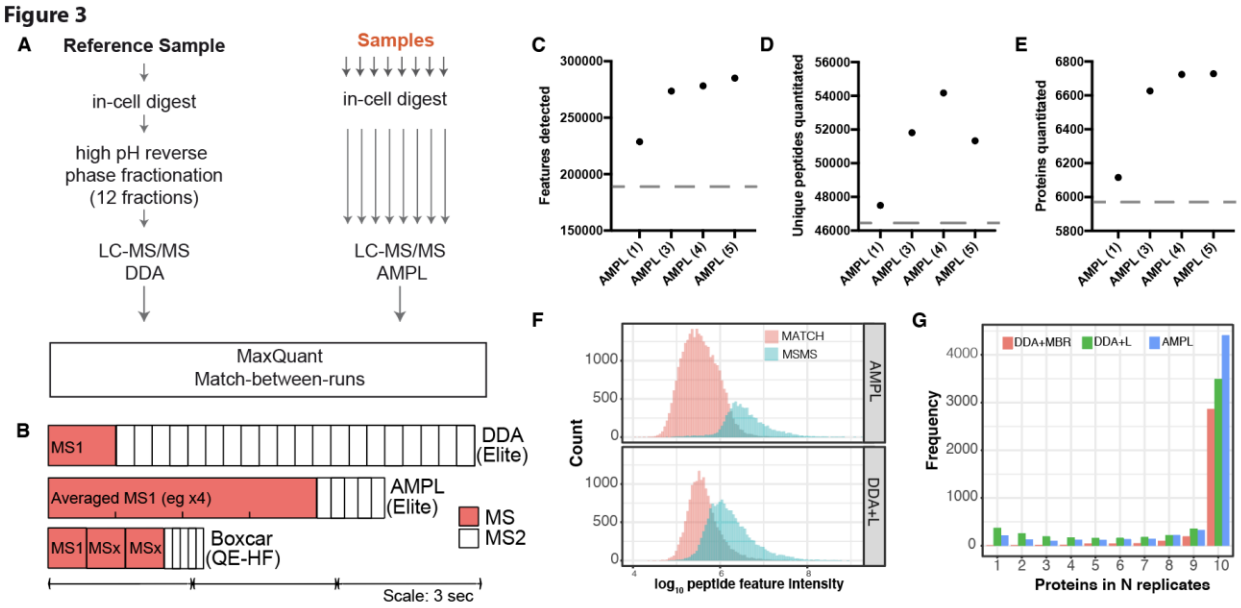


Figure 3. Averaged MS1 precursors with library matching (AMPL) increases peptide detection sensitivity. A) Schematic outlining the AMPL experimental design. B) Both the AMPL and BoxCar acquisition methods prioritise MS time to enhance MS1 scan quality. Schematic comparing duty cycles for DDA, AMPL amd Boxcar acquisition methods on the indicated MS instruments (Orbitrap Elite, Orbitrap HF). C-E) The effect of increasing MS1 averages on the number of features (C), the number of unique peptides (D) and protein groups quantitated (E). Results from DDA acquisition with a library (DDA+L) are shown as dashed line. DDA and AMPL select the top 20 and 5 precursors, respectively, for MS/MS. F) A comparison between AMPL and DDA+L, showing intensity distributions of peptide features identified by MS/MS (blue) and matching to identified library features (red). G) A bar plot indicating the data completeness across 10 replicates either by DDA matching only between replicates, DDA+L and AMPL.

Figure 4

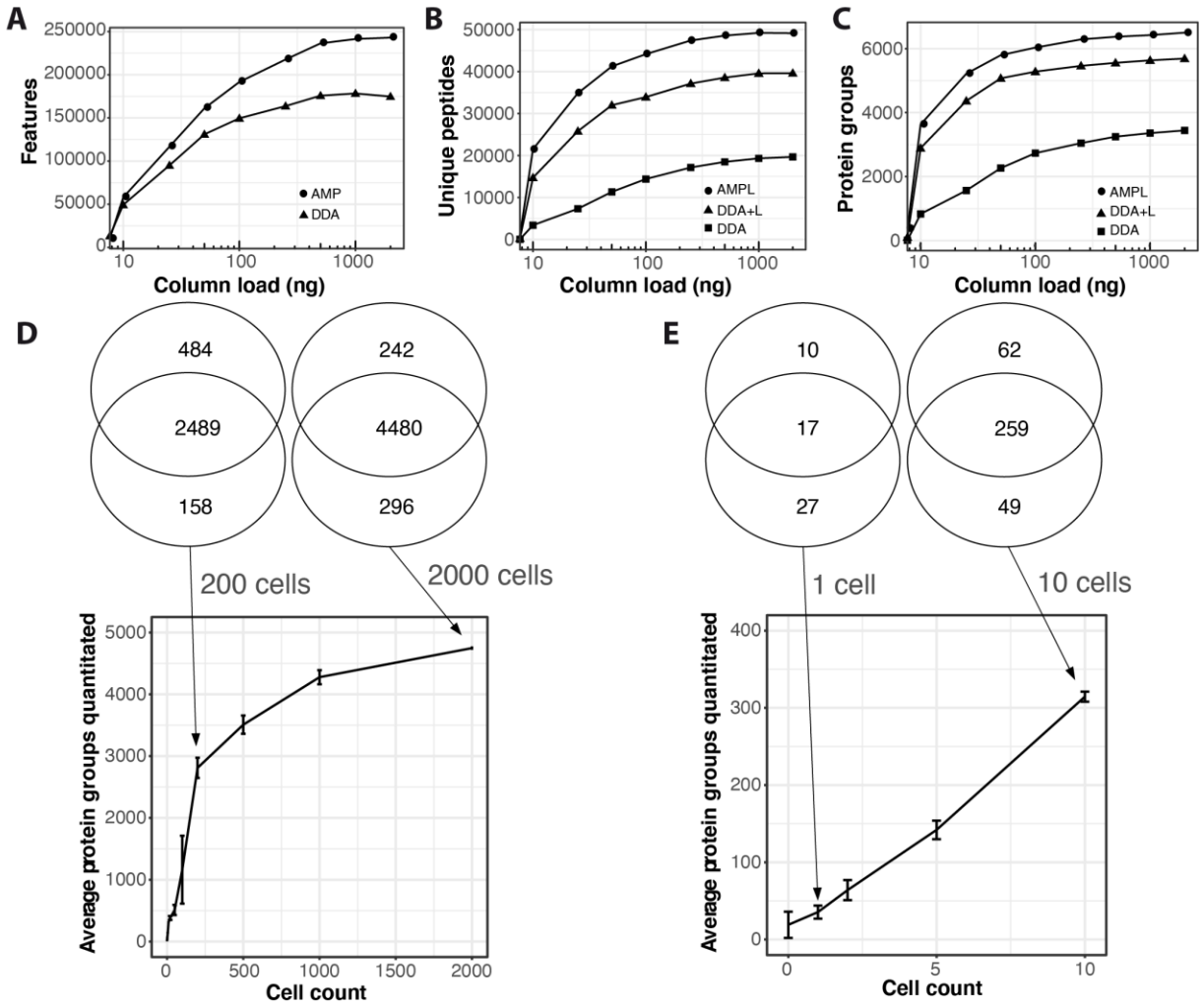


Figure 4. In-cell digest and AMPL enable proteomic analysis with 500-2000 cells. A-C) The sensitivity of AMPL was tested by measuring the number of features detected (A), unique peptides quantitated (B) and protein groups quantitated (C) at the indicated on-column peptide loads, ranging from ~2,000 ng (representing the protein content of ~ 12,000 MCF10A cells) to 10 ng (~ 60 cells). D) The combined in-cell digest and AMPL approach was tested by measuring proteins quantitated at the indicated number of FACS-isolated TK6 cells. The cell number ranged from 2,000 cells to 1 cell. The cell isolation was performed in duplicate. Venn diagrams above the plots show the overlap in proteins between the duplicate analyses. Error bars show the range. E) The same data as in (D) with the x-axis rescaled to show the low cell number measurements.

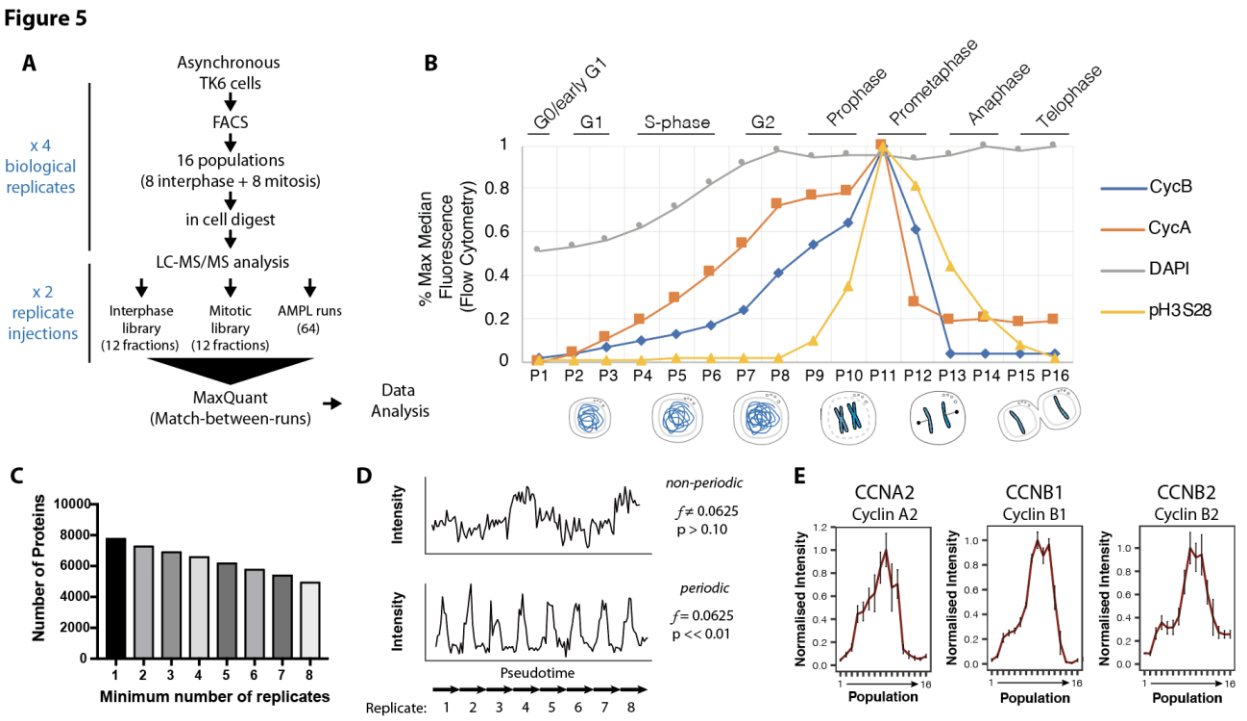


Figure 5. Identification of proteins regulated in abundance across sixteen cell cycle cell populations, comprising eight interphase and eight mitotic states. A) Schematic describing the experimental design and workflow. B) The normalized median fluorescence signal from DAPI (DNA content) and indirect immunofluorescence of cyclin A2, cyclin B1 an H3S28ph. These four markers and DAPI fluorescence width were used to identify and collect the 16 cell cycle populations by FACS. Representative cartoons of the cell cycle phases of these populations are shown below the graph. C) The number of proteins where that protein was detected in at least n number of replicates is shown. For 4,918 proteins, eight replicate intensities were measured in one or more cell cycle populations. D) Identification of cell cycle regulated proteins by pseudoperiodicity analysis. Each pseudotimecourse, representing one set of P1-P16 is arranged in sequence and intensities analysed using a Fisher's periodicity test. Example pseudotimecourses are shown for a non-periodic (HSP90AB1) and periodic (ATAD2) protein, respectively. E) Averaged intensities normalised to maximum are shown for cyclin A2, cyclin B1 and cyclin B2 across the sixteen cell cycle states. Error bars show s.e.m.

Figure 6

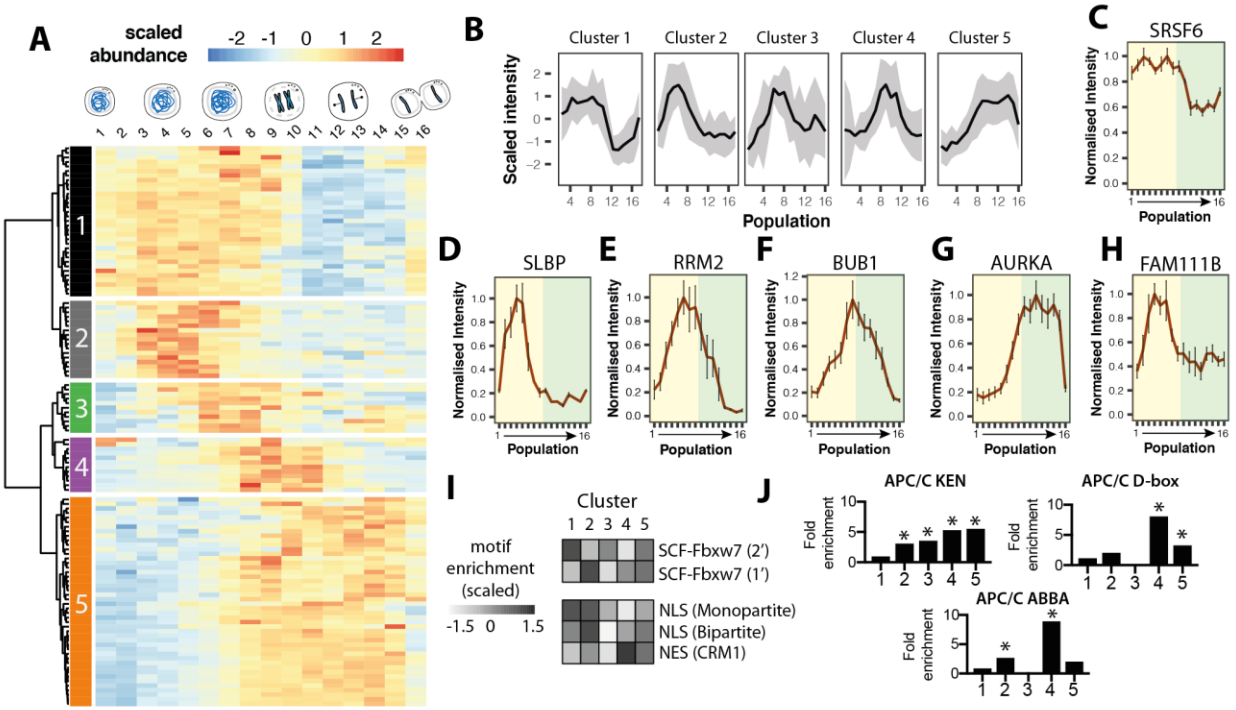


Figure 6. Hierarchal clustering of cell cycle regulated proteins shows classification by biological functions and differential degron enrichment. A) Heatmap of the 119 identified cell cycle regulated proteins organized by cluster. B) Average normalized intensity profiles for the five clusters. C-H) Normalized intensity profiles for example proteins from each cluster (C-G) and a poorly characterized protein, FAM111B (H). I) Enrichment analysis by cluster of SLIMs that mediate interaction with SCF-Fbxw7 (top) and are associated with nuclear import/export (bottom) J) Enrichment analysis by cluster of APC/C degrons. * indicate $p < 0.01$ (Fisher's exact test).

Figure 7

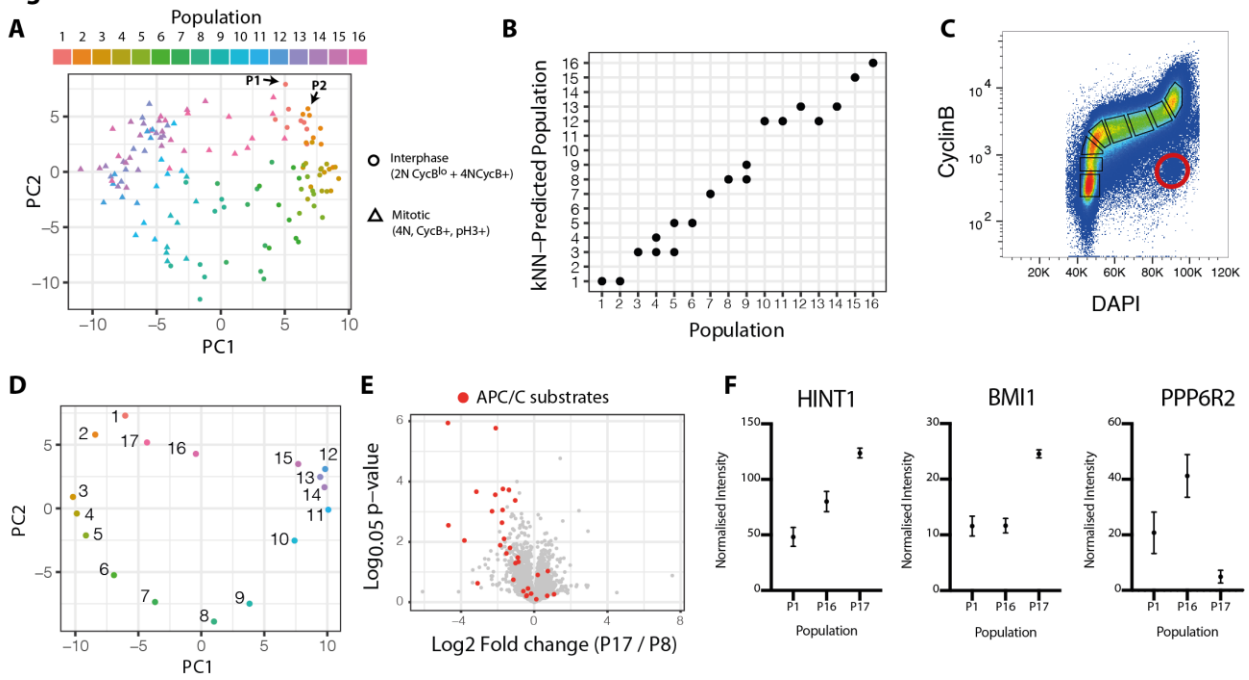


Figure 7. Unbiased classification of proteomes into cell cycle states. A) Principal component analysis of the cell cycle populations using the 119 cell cycle regulated proteins as features. P1 and P2 are highlighted by arrows. B) A kNN model was used to predict the cell population from the abundances of the 119 cell cycle regulated proteins. The performance of the kNN model was assessed using one replicate as the test set. Predicted versus actual cell populations are shown. C) A pseudocolour plot showing a population (P17) that contains 4N DNA content and low cyclin B1 staining. D) PCA analysis, as in (A), but using abundances averaged across the replicates (mean) and including P17. E) Volcano plot showing that characterized APC/C substrates generally show lower abundance in P17 relative to 4N DNA content cells with high cyclin B1 staining (P8). F) In addition to low abundance for APC/C substrates, P17 cells also show high levels of HINT1, BMI1 and low levels of PPP6R2. These proteins have been shown to be important in the DNA damage response. Error bars show s.e.m.

Supplementary Tables

Supplementary Table 1. Key resources

Supplementary Table 1. Impact of formaldehyde fixation on protein coverage.

Supplementary Table 2. Matching FDR empirical estimation.

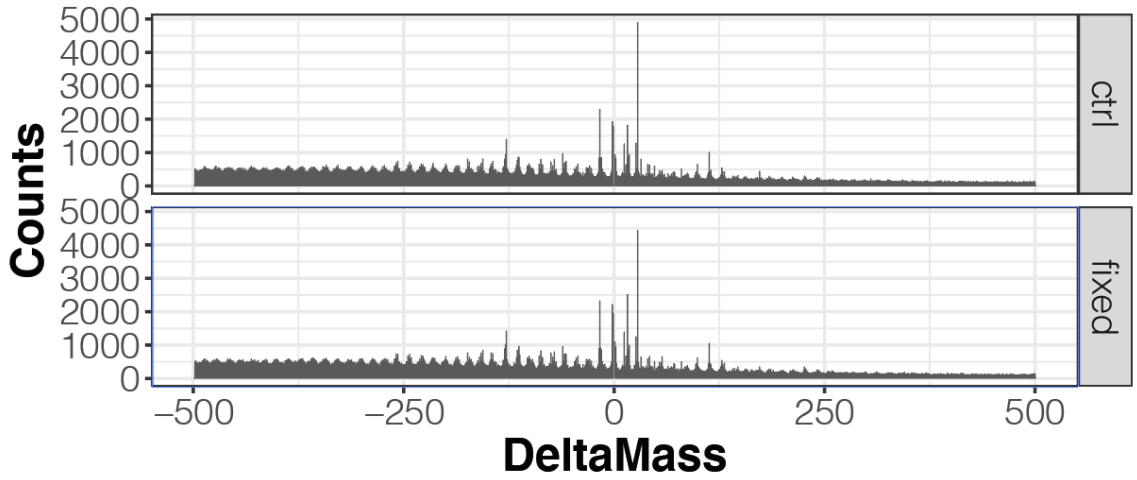
Supplementary Table 3. Summary of cell titration proteome analysis

Supplementary Table 4. Quantitative data and confidence metrics for protein groups identified in the 16 cell cycle populations

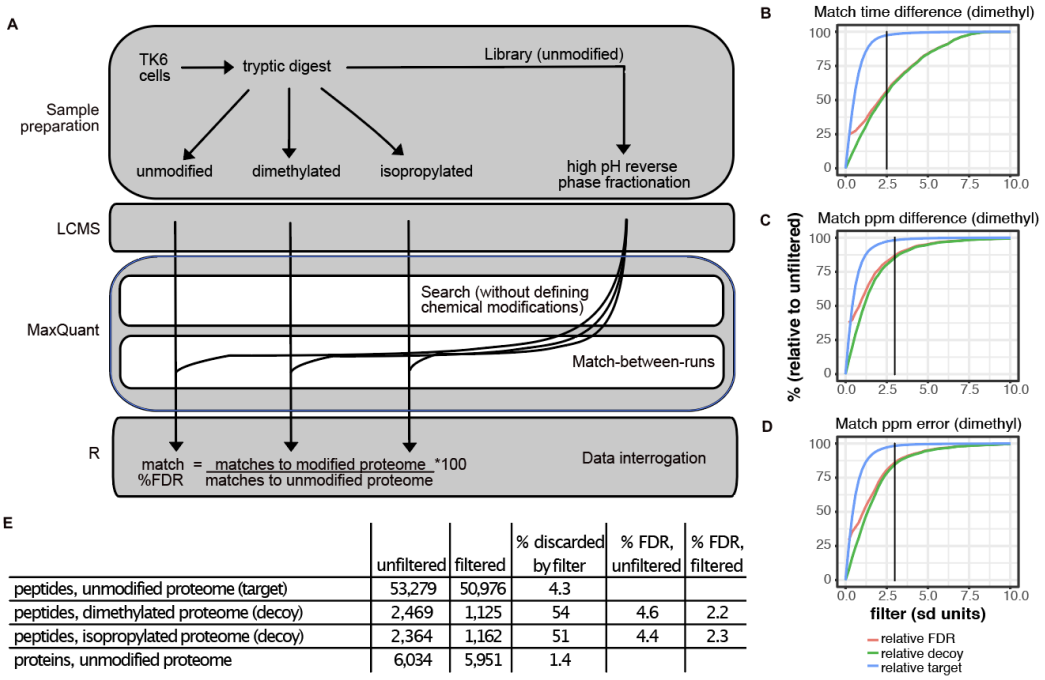
Supplementary Table 5. GO terms and keywords enriched in the group of cell cycle regulated proteins

Supplementary Table 6. Enrichment of short linear sequence motifs (SLIMs) by protein cluster

Supplementary Figures



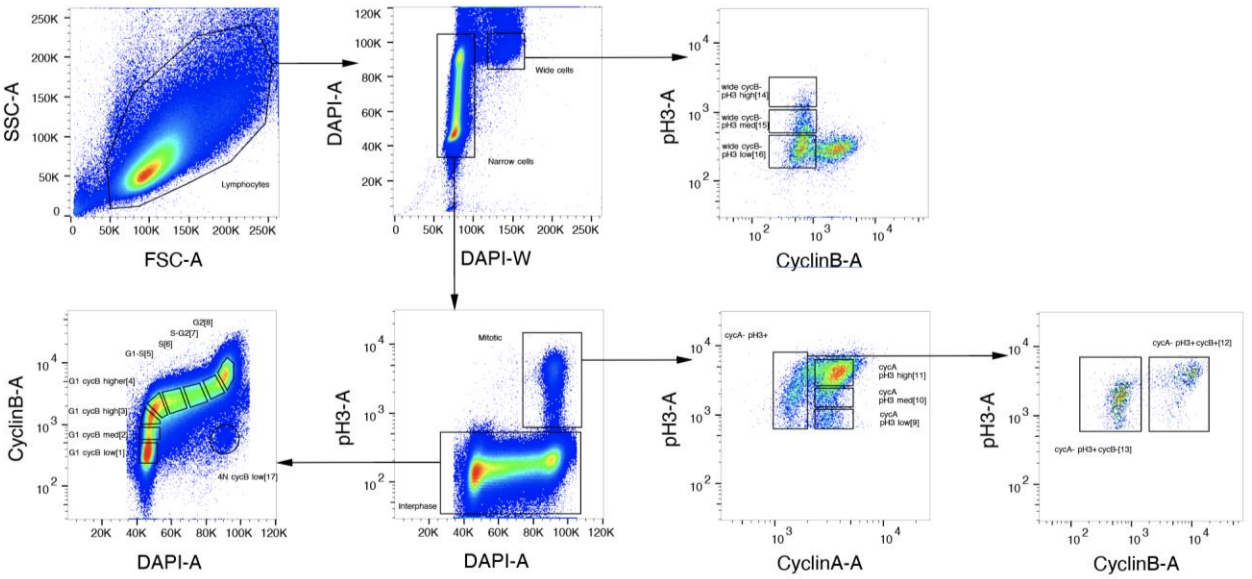
Supplementary Figure 1. Results from an error-tolerant MS-Fragger search comparing control (cells without fixative) and fixed cells.



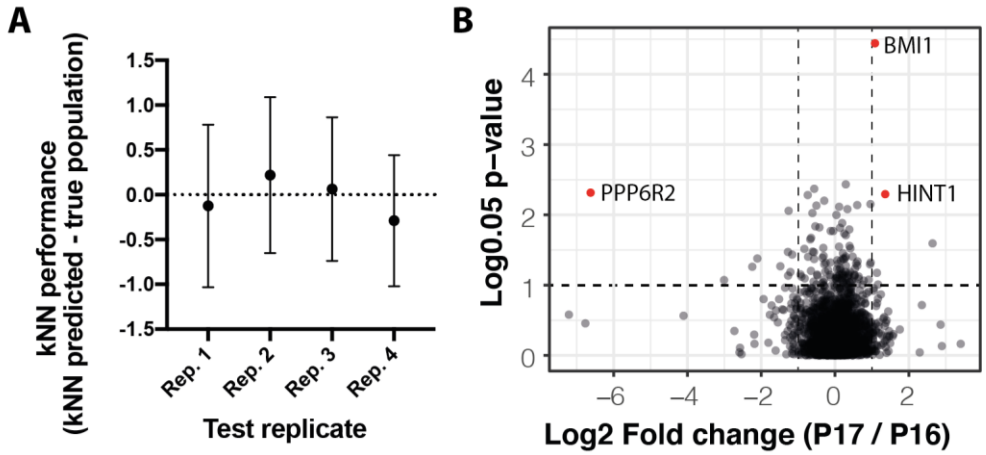
E

	unfiltered	filtered	% discarded by filter	% FDR, unfiltered	% FDR, filtered
peptides, unmodified proteome (target)	53,279	50,976	4.3		
peptides, dimethylated proteome (decoy)	2,469	1,125	54	4.6	2.2
peptides, isopropylated proteome (decoy)	2,364	1,162	51	4.4	2.3
proteins, unmodified proteome	6,034	5,951	1.4		

Supplementary Figure 2. Experimental estimation of match-between-runs FDR using decoy proteome samples. A) Schematic outlining experimental workflow for assessing match-between-runs FDR. For decoy proteomes, TK6 tryptic digests were chemically modified. Unmodified tryptic digests were fractionated for the library. Data was analyzed by MaxQuant with match-between-runs. Matching occurred from the library to unfractionated samples. Chemical modifications were not added to the database search. B-D) The impact of filtering based on match time difference (B), match ppm difference (C), match ppm error (D) at varying levels of stringency. Shown are relative cumulative frequency distributions of the matched peptide features retained in the unmodified proteome (blue), matched peptide features retained in the dimethylated proteome (green), and relative change in FDR (red) at the indicated filtering thresholds (x-axis). E) The number of unmodified peptide and protein quantitations, including both MS/MS and matched peptides, and the number of the match decoy matches, which contribute to estimated match FDR, in modified proteomes. Data are shown before and after filtering at the final thresholds chosen indicated by vertical lines in B-D. F) The analysis in (E) was repeated for a sample acquired by DDA, showing comparable estimated match FDRs between DDA and AMPL.



Supplementary Figure 3. Pseudocolour plots showing the gating strategy to isolate the 16 cell cycle populations by FACS. The red circle indicates P17.



Supplementary Figure 4. Cell state classification and proteomic characterization of P17, a G2-like cell with a DNA damage response protein signature. A) PCA of averaged normalized abundances from the 119 cell cycle regulated proteins as in Fig. 7D but with data from cyclin A and cyclin B2 removed. B) The performance of the kNN classification model was evaluated by using three replicates for the training set and one replicate for the test set. The difference between the predicted and true populations in integer values was calculated for each test set. Because the overall relationship between the populations is cyclic, i.e. P1 is the next assumed state after P16, the difference between P16 and P1 is considered to be 1. The mean and the standard deviation of this difference is shown for all four possible permutations of test and training sets. C) Volcano plot comparing P17 and P16. Proteins of interest are highlighted with red points (BMI1, HINT1, PPP6R2).