Deep Visual Proteomics defines single-cell identity and heterogeneity

Andreas Mund1,19, Fabian Coscia1,2,19, András Kriston3,4, Réka Hollandi3, Ferenc Kovács3,4, Andreas-David Brunner5, Ede Migh5, Lisa Schweizer5, Alberto Santos1,6,7, Michael Bzorek8, Soraya Naimy8, Lise Mette Rahbek-Gjerdrum6,8,9, Beatrice Dyring-Andersen1,10,11, Jutta Bulkescher12, Claudia Lukas12,13, Mark Adam Eckert14, Ernst Lengyel14, Christian Gnann15, Emma Lundberg15,16,17, Peter Horvath3,4,18 and Matthias Mann1,5

Despite the availability of imaging-based and mass-spectrometry-based methods for spatial proteomics, a key challenge remains connecting images with single-cell-resolution protein abundance measurements. Here, we introduce Deep Visual Proteomics (DVP), which combines artificial-intelligence-driven image analysis of cellular phenotypes with automated single-cell or single-nucleus laser microdissection and ultra-high-sensitivity mass spectrometry. DVP links protein abundance to complex cellular or subcellular phenotypes while preserving spatial context. By individually excising nuclei from cell culture, we classified distinct cell states with proteomic profiles defined by known and uncharacterized proteins. In an archived primary melanoma tissue, DVP identified spatially resolved proteome changes as normal melanocytes transition to fully invasive melanoma, revealing pathways that change in a spatial manner as cancer progresses, such as mRNA splicing dysregulation in metastatic vertical growth that coincides with reduced interferon signaling and antigen presentation. The ability of DVP to retain precise spatial proteomic information in the tissue context has implications for the molecular profiling of clinical samples.

Modern microscopy’s versatility, resolution and multi-modal nature delivers increasingly detailed images of single-cell heterogeneity and tissue organization. Currently, a predefined subset of proteins is usually targeted, far short of the actual complexity of the proteome. Taking advantage of substantially increased sensitivity in technology based on mass spectrometry (MS), we set out to enable the analysis of proteomes within their native, subcellular context to explore their contribution to health and disease. We combined sub-micron-resolution imaging, image analysis for single-cell phenotyping based on artificial intelligence (AI) and isolation with an ultra-sensitive proteomics workflow to accurately define single-cell boundaries and cell classes as well as the transfer of the automatically defined features into proteomic samples, ready for analysis. To this end, we introduce the software BIAS (Biological Image Analysis Software), which coordinates scanning and laser microdissection (LMD) microscopes. This seamlessly combines data-rich imaging of cell cultures or archived biobank tissues (formalin-fixed and paraffin-embedded (FFPE)) with deep-learning-based cell segmentation and machine-learning-based identification of cell types and states. Cellular or subcellular objects of interest are selected by the AI alone or after instruction before being subjected to automated LMD and proteomic profiling. Data generated by DVP can be mined to discover protein signatures providing molecular insights into proteome variation at the phenotypic level while retaining complete spatial information.

Results

Image-guided single-cell isolation for cell-type-resolved proteomics. The microscopy-related aspects of the DVP workflow build on high-resolution whole-slide imaging, machine learning (ML) and deep learning (DL) for image analysis. First, we used scanning microscopy to obtain high-resolution whole-slide images and developed a software suite for integrative image analysis termed ‘BIAS’ (Methods). BIAS processes multiple two-dimensional (2D) and three-dimensional (3D) microscopy image file formats, supporting major microscope vendors and data formats. It combines image pre-processing, DL-based image

---

Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. 1Spatial Proteomics Group, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. 2Synthetic and Systems Biology Unit, Biological Research Centre, Eötvös Loránd Research Network, Szeged, Hungary. 3Single-Cell Technologies Ltd., Szeged, Hungary. 4Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. 5Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark. 6Big Data Institute, Li-Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. 7Department of Pathology, Zealand University Hospital, Roskilde, Denmark. 8Institute for Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. 9Department of Dermatology and Allergy, Herlev and Gentofte Hospital, University of Copenhagen, Hellerup, Denmark. 10Leo Foundation Skin Immunology Research Center, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. 11Biological Research Centre, Eötvös Loránd Research Network, Szeged, Hungary. 12Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. 13Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark. 14Protein Imaging Platform, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. 15Protein Signaling Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. 16Department of Obstetrics and Gynecology/Section of Gynecologic Oncology, University of Chicago, Chicago, IL, USA. 17Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden. 18Department of Bioengineering, Stanford University, Stanford, CA, USA. 19Chan Zuckerberg Biohub, San Francisco, CA, USA. 20Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. 21These authors contributed equally: Andreas Mund, Fabian Coscia. e-mail: andreas.mund@cpr.ku.dk; horvath.peter@brc.hu; mmann@biochem.mpg.de
segmentation, feature extraction and ML-based phenotype classification. Building on a recent DL-based algorithm for cytoplasm and nucleus segmentation, we undertook several optimizations to implement pre-processing algorithms to maintain high-quality images across large image datasets. DL methods require large training datasets, which is a considerable challenge due to the limited size of high-quality training data. To address this challenge, we used nucleAIzer and applied project-specific image style transfer to synthesize artificial microscopy images resembling real images. This approach is inherently adaptable to different biological scenarios, such as new cell and tissue types or staining techniques. We trained a deep neural network with these synthetic images for specific segmentation of the cellular compartment of interest (for example, nucleus or cytoplasm). We benchmarked it against two leading DL approaches—unet4nuclei and Cellpose—and a widely used adaptive threshold-based and object-splitting-based method. Our cell and nucleus segmentation algorithms of cell cultures and tissues showed the highest accuracy (Fig. 2b, Extended Data Fig. 1a, Table 1 and Supplementary Table 1). Our current benchmarking results are supported by a previous study where we performed an extensive comparison to additional methods and software (for example, ilastik, on a large heterogeneous microscopy image set). For interactive cellular phenotype discovery, BIAS performs phenotypic feature extraction, taking into account morphology and neighborhood features based on supervised and unsupervised ML (Extended Data Fig. 1b and Methods). Feature-based phenotypic classification is readily combined with biomarker expression level from antibody staining for precise cell classification. ML has previously been used for image analysis and cell selection but not combined with unbiased proteomics. Furthermore, we extended BIAS with a Python interface; thus, data access and manipulation is also possible using standard Python functions in a generic way, including the integration of open-source packages and custom algorithms.

To physically extract the cellular features discovered with BIAS, we developed an interface between scanning and LMD microscopes (currently Zeiss PALM MicroBeam and Leica LMD6 and LMD7) (Fig. 2c). BIAS transfers cell contours between the microscopes, preserving full accuracy. LMD has a theoretical accuracy of 70 nm using a ×150 objective, but, in practice, we reached 200 nm. After optimization, the LMD7 can autonomously excise 1,250 high-resolution contours per hour, equivalent to 50 to 100 cells per sample (Methods). To prevent potential laser-induced damage to cell membranes, we excise contours with an offset (Fig. 2c, Supplementary Videos 1 and 2).

Current LMD methods preserve the spatial context but are mostly limited to human-eye-observable phenotypes and require manual selection of cells, often resulting in admixing of different cell types, which constrains throughput and de novo discovery.
**Fig. 2 | BIAS for integrative image analysis and automated LMD single-cell isolation.**

**a.** AI-driven nucleus and cytoplasm segmentation of normal-appearing and cancer cells and tissue using BIAS. **b.** We benchmarked the accuracy of its segmentation approach using the F1 metric and compared results to three additional methods—M1 is unet4nuclei\(^6\), M2 is CellProfiler\(^8\) and M3 is Cellpose\(^7\)—while OUR refers to nuclceAlizer\(^3\). Bars show mean F1 scores with s.e.m.; n = 10 independent images for melanoma tissue and (U2OS) cells, and n = 20 for salivary gland tissue. Visual representation of the segmentation results: green areas correspond to true positive, blue to false positive and red to false negative. **c.** Practical illustration of the functions in the upper panel. **d.** Immunofluorescence staining of the human fallopian tube epithelium with FOXJ1 and EpCAM antibodies, detecting ciliated and epithelial cells, respectively. Left panel: Ciliated (FOXJ1-positive) and secretory (FOXJ1-negative) cells. Right panel: Cell classification based on FOXJ1 intensity. Class 1 (FOXJ1-positive) and class 2 (FOXJ1-negative); magnification factor = x387.

**e.** PCA of FOXJ1-positive and FOXJ1-negative cell proteomes. **f.** Heat map of known protein markers for secretory and ciliated cells. Protein levels are z-scored. Asterisks represent imputed data. The marker list was derived from the Human Protein Atlas\(^20\) project and based on literature mining. **g.** Volcano plot of the pairwise proteomic comparison between FOXJ1-positive and FOXJ1-negative cells. Cell-type-specific marker proteins are highlighted in green and turquoise, and black represents potential novel marker proteins. Significant enriched cell-type-specific proteins are displayed above the black lines (two-sided t-test, FDR < 0.05, s0 = 0.1, n = 4 biological replicates).
Table 1 | Mean F1 scores of the compared segmentation methods on our samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>OUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>U2OS cyto</td>
<td>0.0667 ± 0.0075</td>
<td>0.5994 ± 0.0262</td>
<td>0.7205 ± 0.0152</td>
<td>0.7336 ± 0.0218</td>
</tr>
<tr>
<td>Melanoma nuc</td>
<td>0.1126 ± 0.0151</td>
<td>0.4386 ± 0.0157</td>
<td>0.1801 ± 0.0504</td>
<td>0.5498 ± 0.0231</td>
</tr>
<tr>
<td>Melanoma cyto</td>
<td>0.0058 ± 0.0021</td>
<td>0.0549 ± 0.0083</td>
<td>0.4859 ± 0.0354</td>
<td>0.5536 ± 0.0625</td>
</tr>
<tr>
<td>Salivary gland nuc</td>
<td>0.0797 ± 0.0138</td>
<td>0.6488 ± 0.0430</td>
<td>0.0338 ± 0.0145</td>
<td>0.7684 ± 0.0316</td>
</tr>
<tr>
<td>Salivary gland cyto</td>
<td>0.0714 ± 0.0151</td>
<td>0.0793 ± 0.0167</td>
<td>0.3174 ± 0.0588</td>
<td>0.5051 ± 0.0386</td>
</tr>
<tr>
<td>Melanoma (pink) nuc</td>
<td>0.0682 ± 0.0183</td>
<td>0.2999 ± 0.0599</td>
<td>0.0364 ± 0.0238</td>
<td>0.5079 ± 0.0392</td>
</tr>
<tr>
<td>Melanoma (pink) cyto</td>
<td>0.0261 ± 0.0070</td>
<td>0.0865 ± 0.0213</td>
<td>0.2659 ± 0.0429</td>
<td>0.2839 ± 0.0229</td>
</tr>
<tr>
<td>Fallopian tube nuc</td>
<td>0.0006 ± 0.0009</td>
<td>0.3121 ± 0.0501</td>
<td>0.3160 ± 0.0631</td>
<td>0.4724 ± 0.0683</td>
</tr>
<tr>
<td>Fallopian tube cyto</td>
<td>0.0016 ± 0.0023</td>
<td>0.0671 ± 0.0208</td>
<td>0.4566 ± 0.0530</td>
<td>0.3455 ± 0.0473</td>
</tr>
</tbody>
</table>

The methods are as follows: M1 is unet4nuclei, M2 is CellProfiler, M3 is Cellpose and OUR refers to nucleAIzer (implemented in BIAS). High scores are highlighted in bold. Asterisks (*) mark that M1 is intended for nucleus segmentation but was applied to segment cytoplasm. s.e.m. is displayed with ± after the mean F1 scores in each cell.

To explore the sensitivity, specificity and robustness of our DVP workflow, we obtained normal human fallopian tube tissue and separated ciliated from secretory cells—the two major cell types of the fallopian tube epithelium—using the cell-lineage-specific transcription factor FOXJ1, a master regulator of cilia function, and measured their proteomes (Fig. 2e–h, Extended Data Fig. 1c–f and Supplementary Table 2). We solely detected FOXJ1 (ciliated cells) in FOXJ1-stained cells (Fig. 2e,g), along with more than 5,000 other quantified proteins with excellent correlations of biological replicates (Extended Data Fig. 1d,e). Bioinformatic analysis of differences in protein abundance mirrored the biologic features of the distinct cell types. (Fig. 2f–h and Extended Data Fig. 1c–f). This was driven by known protein markers of ciliated cells and expanded to proteins not yet functionally associated with these cell types. We used the fallopian tube epithelium as an example to highlight the importance of the combination of antibody-based tissue staining and unbiased, quantitative proteomics. Such in vivo cell type comparisons will allow the discovery of cell type and cell state markers and provide unbiased information to understand disease states at the global proteome level. Of note, high-grade serous ovarian cancer originates in the fallopian tube epithelium and, our method and our method can now be applied to study the early onset of the disease without admixing unrelated cell types.

DVP defines single-cell heterogeneity at the subcellular level. We applied our workflow to an unperturbed cancer cell line to determine if DVP can characterize functional heterogeneity between ostensibly similar cells (fluorescent ubiquitination-based cell cycle indicator (FUCCI) U2OS cells). After DL-based segmentation for nuclei and cell membrane detection, we isolated 80–100 single cells or 250–300 nuclei per phenotype (Figs. 2c,d and 3a,b). The analysis of small numbers of cells by MS has been a longstanding goal, held back by formidable analytical challenges in the transfer, processing and analysis of minute samples, which we addressed in turn. We processed samples using our recently developed workflow for ultra-low sample input, which omits any sample transfer steps and ensures de-crosslinking in very low volumes (Methods). We found that samples could be analyzed directly from 384 wells without any additional sample transfer or clean-up. For MS measurements, we employed a data-independent acquisition method using parallel accumulation–serial fragmentation with an additional ion mobility dimension and optimal fragment (diaPASEF) ion recovery on a newly developed mass spectrometer. Replicates of cell and nucleus proteomes demonstrated high quantitative reproducibility (Pearson r = 0.96), and proteomes of whole cells differed from those of nuclei alone, as expected from subcellular proteomics experiments based on biochemical separation (Extended Data Fig. 2a,b). In the bioinformatic enrichment analysis, terms like plasma membrane, mitochondrion, nucleosomes and transcription factor complexes were highly significant (false discovery rate (FDR) < 10⁻³) (Fig. 3c).

To address if morphological differences between nuclei are also reflected in their proteomes, we used an unsupervised phenotype finder model to identify groups of morphologically distinct nuclei.
Whole cell segmentation
Nuclei segmentation

Cells
Nuclei

Spliceosome
Transcription factor complex
Nucleosome
Nucleoplasm
ER
Ribosome
Plasma membrane
Mitochondrion
Cytoskeleton

Class1
Class2
Class3
Class4
Class5
Class6

Form

DNA
ERU

Class1
Class2
Class3
Class4
Class5
Class6

3,653 protein groups

515 significant protein groups (FDR < 0.05)

Catalytic activity, acting on RNA
Cancer-related genes, FDA approved drug targets
Switching of origins to a post-replicative state
Packaging of telomere ends
Condensation of prophase chromosomes
Unwinding of DNA
Cancer-related genes
Non-homologous end-joining (NHEJ)
Nucleoplasm
Cell cycle regulated
Transcriptional regulation by small RNAs
Condensation of prophase chromosomes

Human Protein Atlas
Reactome pathway

Enriched terms

0
2
4
6

Enrichment factor
based on nuclear area, perimeter, form factor, solidity and DNA staining intensity (Fig. 3d). ML found three primary nuclei classes (27–37% each) and also identified three rare ones (2–4% each) (Extended Data Fig. 2c). The resulting six distinct nuclei classes had visible differences in size and shape, with class 1 representing mitotic states and the remaining five classes representing interphase with varying feature weighting (Fig. 3e,f). We focused on those five nuclei classes of unknown origin for subsequent analysis. In principal component analysis (PCA), replicates of the respective proteomes clustered closely, and the more frequent classes (2, 3 and 5) grouped together (Fig. 3g). To verify and quantify this observation, we compared each cell class proteome to a proteome of all ‘mixed’ nuclei in a field of view. This revealed that the rarest cell classes had the highest numbers of differentially expressed proteins compared to unclassified ‘bulk’ proteomes (Extended Data Fig. 2d,e). We next asked if the proteomic differences across the five nuclei classes suggested any functional differences among the interphase states (Fig. 3d,f). The 515 significantly differentially expressed proteins across classes were enriched for nuclear and cell-cycle-related proteins (for example, ‘switching of origins to a post-replicative state’ and ‘condensation of prophase chromosomes’), suggesting the cell cycle as a functional driver of separation (Fig. 3h–j, Extended Data Fig. 2f and Supplementary Tables 3 and 4). Comparing our data to a single-cell imaging dataset of cell-cycle-regulated proteins, we found significant enrichment in our regulated proteomes (FDR < 10^{-6}). Nuclear area, one of the driving features among the different classes identified, increased during interphase from G1 to S/G2 cells (Fig. 3e and Extended Data Fig. 3a–c), further supporting the importance of the cell cycle in defining the nuclei classes.

Our single-cell-type proteomes discovered several uncharacterized proteins, presenting an opportunity to associate them with a potential cellular function. Focusing on C11orf98, C7orf50, C1orf112 and C19orf53, which remained after data filtering (ANOVA P < 0.05), showed class-specific expression patterns (Extended Data Fig. 3d). C7orf50 was most highly expressed in the nuclei of classes 2, 3 and 4 nuclei, which showed S/G2-specific characteristics (Fig. 3f and Extended Data Fig. 3d,e), suggesting that its expression is cell cycle regulated. Indeed, we confirmed higher levels of C7orf50 in G1/S and S/G2 compared to G1 phase cells (Extended Data Fig. 3e). As cell-cycle-regulated proteins may be associated with cancer prognosis, we investigated C7orf50 in the human pathology atlas where high expression was associated with favorable outcomes in pancreatic cancer (Extended Data Fig. 3g; P < 0.0011). Bioinformatic analysis revealed interaction, co-expression and co-localization with the protein LYAR (‘cell growth-regulating nucleolar protein’), suggesting a functional link to cell proliferation (Extended Data Fig. 3h,i).

Class 6 showed an intriguing proteomic signature independent of known cell cycle markers (Fig. 3i,j). These rare, bean-shaped nuclei showed upregulation of specific cytoskeletal and cell adhesion proteins (for example, VIM, TUBB, ACTB and ITGB1), suggesting that these signatures derived from migrating cells undergoing nuclear deformation, suggestive of cellular invasion. Note that we classified nuclei from 2D images, but LMD isolates them in 3D. Thus, samples also probe morphology-driven protein re-localization around the nucleus as exemplified by class 6 nuclei. Likewise, excising the nuclei captures the trafficking of proteins to and from the cytosol to some degree.

These cell culture experiments establish that DVP correlates cellular phenotypes, heterogeneity and dynamics with the proteome level in an unbiased way for common and rare phenotypes.

**DVP applied to cancer tissue heterogeneity.** Billions of patient samples are collected routinely during diagnostic workup and stored in the archives of pathology departments around the world. The precise proteomic characterization of single cells in their spatial and subcellular context from tissue slides could have a tremendous clinical effect, complementing the emerging field of digital pathology. We selected archived paraffin-embedded tissue of a salivary gland acinic cell carcinoma, a rare and understudied malignancy of epithelial secretory cells of the salivary gland. We developed an immunohistochemical (IHC) staining protocol on glass membrane slides for LMD and stained the tissue for EpCAM to outline the cellular boundaries for segmentation and feature extraction by BIAS (Methods). These histologically normal-appearing regions were mainly comprised of acinar, ductal and myoepithelial cells, whereas the carcinoma component had predominantly uniform tumor cells with round nuclei and abundant basophilic cytoplasm.

To identify disease-specific protein signatures, we aimed to compare the histologically normal-appearing acinar cells with the malignant cells rather than admixing with varying proportions of unrelated cells. To this end, we classified acinar and duct cells from normal parotid gland tissue based on their cell-type-specific morphological features and isolated single-cell classes for proteomic analysis (Fig. 4a and Extended Data Fig. 4a). Bioinformatics analysis of the measured proteome differences revealed significant biological differences between these neighboring cell types, reflecting their distinct physiological functions. Acinar cells, which produce and secrete saliva in secretory granules, showed high expression of proteins related to vesicle transport and glycosylation along with known acinar cell markers such as α-amylase (AMY1A), CA6 and PIP (Extended Data Fig. 4b). In contrast, ductal cells expressed high levels of mitochondria and metabolism-related proteins required to meet the high energy demand for saliva secretion (Extended Data Fig. 4c and Supplementary Table 5). For comparison, we exclusively excised malignant and benign acinar cells from the various regions within the same tissue section. The proteomes of acinar cells clustered together regardless of disease state, indicating a strong cell-of-origin signature (Extended Data Fig. 4d). Analyzing six normal-appearing replicates and nine neoplastic regions showed excellent within-group proteome correlation (Pearson r > 0.96). The lower correlation of normal cells and cancer cells reflected disease-specific and cell-type-specific proteome changes (Pearson r = 0.8; Fig. 4d,e and Supplementary Table 6). Acinar cell markers in the carcinoma were significantly downregulated, consistent with previous reports. DVP allowed us to discover upregulation of interferon response proteins (for example, MX1 and HLA-A; Supplementary Table 6) and the proto-oncogene SRC, both
actionable therapeutic targets\textsuperscript{26} (Fig. 4e). We validated the proteomic findings using IHC analysis of significantly enriched proteins in either normal-appearing or cancerous tissue. This resulted in the selection of CNN1, SRC, CK5 and FASN (Fig. 4f), which confirmed our proteomic results, demonstrated the absence of contamination and supported the specificity of our DVP approach.

Decoding the molecular alterations in melanoma development and progression is key to identifying therapeutic vulnerabilities in...
1 patient
7 regions
5 cell classes
27 samples
Facts
throughput
>30,000 contours/day
dimension
50–100 cells/sample

Melanocytes (mel) In situ Radial Vertical CD146-high CD146-low Stroma

Oxidative phosphorylation (hallmark, 151 proteins)

Interferon gamma response (hallmark, 53 proteins)

Increased in vertical growth
Degradation of the extracellular matrix
mRNA splicing – major pathway

Decreased in vertical growth
MHC class II antigen presentation
Interferon alpha/beta signaling

Isolate melanoma cells from different vertical regions and same predicted ML class
Compare proteomes (region 1 vs 2)
Analyze enriched pathways

Protein level (z-score)
Figure 5 | DVP applied to archived primary melanoma tissue. a, DVP sample isolation workflow to profile primary melanoma. b, DVP applied to primary melanoma immunohistochemically stained for the melanocyte marker SOX10 and the melanoma marker CD146. Left panel: stained melanoma tissue on a PEN glass membrane slide. Right panel: pathology-guided annotation of different tissue regions. Scale bar, 1 mm. c, Pathologist-guided and ML-based cell classification based on CD146 and SOX10 staining intensity and spatial localization: normal melanocytes, stromal cells, melanoma in situ, CD146-low melanoma, CD146-high melanoma, radial growth melanoma and vertical growth melanoma. Right lower panel: frequency of classes predicted by unsupervised ML (k-means clustering). d, Example pictures of the seven identified classes. Magnification factor ×4,400. e, Correlation matrix (Pearson r) of all 27 measured proteome samples. f, PCA of proteomes. g, PCA of all melanoma-specific proteomes from in situ to invasive (vertical growth) melanoma. h, Unsupervised hierarchical clustering based on all 1,910 ANOVA significant (FDR < 0.05) protein groups. Two clusters of upregulated (cluster A) or downregulated (cluster B) proteins in invasive melanoma are highlighted. i, Tissue heat map mapping the proteomics results onto the imaging data. Relative pathway levels of selected terms from the two clusters are highlighted in i. Median protein levels were calculated per annotation and plotted for each isolated cell class against their x and y coordinates, as defined by their segmented cellular contours. j, Box plots of z-scored protein levels for the differentially regulated pathways visualized in i above. The box plots define the range of the data (whiskers), 25th and 75th percentiles (box) and medians (solid line). Outliers are plotted as individual dots outside the whiskers. k, Comparing proteomic changes in CD146-high melanoma cells (class 4) of the vertical growth (region 2) with the radial growth (region 1). Blood vessels in proximity to melanoma cells of the vertical growth are highlighted in red. Scale bar, 1 mm. l, Gene set enrichment analysis plot of significantly enriched pathways for melanoma cells of the vertical and radial growth phase. Pathway enrichment analysis was based on the protein fold change between vertical and radial melanoma cells and performed with the ClusterProfiler R package. Enriched terms with an FDR < 0.05 are shown. MHC, major histocompatibility complex.

Discussion

DVP combines imaging technologies with unbiased proteomics to quantify the number of expressed proteins in a given cell, map tissue or cell-type-specific proteomes or to identify targets for future drugs and diagnostics. We showed how our analyses describe a rich ‘microcosm in a slide,’ uncovering key pathways dysregulated in cancer progression and effectively extending ‘digital pathology’ by a molecular dimension. It is broadly applicable to any biological system that can be microscopically imaged, from cell culture to pathology. As a single slide can encompass hundreds of thousands of cells, DVP can discover and characterize rare cell states and interactions. In contrast to single-cell transcriptomics, DVP can readily analyze the ECM’s subcellular structures and spatial dynamics. With further improvements in proteomics technology, DVP should also be suited to study proteoforms and post-translational modifications at a single-cell-type level.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01302-5.

Received: 8 March 2022; Accepted: 30 March 2022;
Published online: 19 May 2022

References


Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. © The Author(s) 2022.
Methods

Patient samples and ethics. We collected archival FFPE tissue samples of salivary gland acinic cell carcinoma and melanoma from the Department of Pathology, Zealand University Hospital, in Roskilde, Denmark. Melanoma tissue was from a 64-year-old female and was macroscopically and histologically normal appearing. All patients consented before surgery. Patient-derived tissues were obtained at full histological sections, and tissue specimens were assessed by a board-certified pathologist. 

The U2OS FUCCI cells were kindly provided by Atsushi Miyawaki14. These cells are endogenously tagged with two fluorescent proteins fused to the cell cycle regulators CDT1 (mKO2-hCdt1+) and p57 (mAG-hGem+). CDT1 accumulates during the G1 phase, whereas geminin accumulates in the S and G2 phases, allowing cell cycle monitoring. The cells were cultivated at 37 °C in a 5.0% CO2 humidified environment in McCoy's 5A (modified) medium GlutaMax supplement (Thermo Fisher Scientific, 36000021) supplemented with 10% FBS and penicillin–streptomycin (Thermo Fisher Scientific, C2500). 

Immunocytochemistry was carried out following an established protocol38. For validation experiments (Extended Data Fig. 3), 96-well glass-bottom plates (Greiner Bio-One) were coated with 12.5 µg/ml of human fibronectin (Sigma-Aldrich) for 1 hour at room temperature (RT) and loaded on the fully automated instrument Omnis (Dako). Immunocytochemistry was performed in a water bath employing EDTA retrieval buffer (1 mM EDTA, 0.05% Triton X-100) according to the manufacturer's instructions. Finally, slides were rinsed in water, counterstained with Mayer's hematoxylin and cover-slipped.

Immunofluorescence staining. Cells were first incubated with 5-ethyl-2’-deoxyuridine (EdU) for 20 minutes and then fixed for 5 minutes at room temperature with 4% parafomaldehyde (PFA) and washed three times with PBS. Cells were then permeabilized with PBS/0.2% Triton X-100 for 2 minutes on ice and washed three times with PBS. Cells were then stained with an EdU labeling kit (Life Technologies) and counterstained with Hoechst 33342 for 10 minutes. Slides were mounted with GBI (GBI Labs, E01-18).

For validation experiments (Extended Data Fig. 3), 96-well glass-bottom plates (Greiner Bio-One) were coated with 12.5 µg/ml of human fibronectin (Sigma-Aldrich) for 1 hour at room temperature (RT) and loaded on the fully automated instrument Omnis (Dako). Immunocytochemistry was carried out following an established protocol38. Then, 8,000 U2OS cells were seeded in each well and incubated in a 37 °C and 5% CO2 environment for 24 hours. Cells were washed with PBS, fixed with 40 µl of 4% ice-cold PFA and permeabilized with 40 µl of 0.1 Triton X-100 in PBS for 3x5 minutes. Rabbit polyclonal HPA antibodies targeting the proteins of interest were diluted in blocking buffer (PBS + 4% FBS) at 2–24 µg/ml along with primary marker antibodies (see below) and incubated overnight at 4 °C. Cells were washed with PBS for 4x10 minutes and incubated with secondary antibodies (goat anti-rabbit Alexa Fluor 488 (A11034, Thermo Fisher Scientific), goat anti-mouse Alexa Fluor 568 (A21424, Thermo Fisher Scientific), rabbit anti-chicken Alexa Fluor 647 (A21449, Thermo Fisher Scientific)) in blocking buffer at 1.25 µg/ml for 20 minutes at room temperature. Cells were counterstained in 0.05 µg/ml of DAPI for 15 minutes, washed with for 4x10 minutes and mounted in PBS.

Primary antibodies used were as follows:

- For Ctfr050 cell cycle validation: mouse anti-ANLN at 1.25 µg/ml (ab39966, AbCam). Mouse anti CCNB1 at 1 µg/ml (160220, BD Biosciences). Rabbit anti-Ctfr050 at 1 µg/ml (HPA052281, Atlas Antibodies).

Mouse anti ANLN at 1.25 µg/ml (ab39966, AbCam). Mouse anti CCNB1 at 1 µg/ml (160220, BD Biosciences). Rabbit anti-Ctfr050 at 1 µg/ml (HPA052281, Atlas Antibodies).

For human fallopian tube tissue, FFPE tissue sections (2.5 µm) were mounted and pre-processed as described above. Therefore, tissue was dewaxed by washing 2x2 minutes in 100% xylene, followed by a series of 100%, 95% and 70% ethanol for 1 minute, respectively, and 3x1 minute in ddH2O. Antigen retrieval was performed in a water bath employing EDTA retrieval buffer (1 mM EDTA, 0.05% Tween 20, pH 8.0) at 95 °C for 1 hour. Subsequent to a cooling phase of 1 hour at room temperature, blocking was conducted with 10% goat serum in TBST for 1 hour at room temperature. Primary antibodies targeting FOXJ1 (mouse, dilution 1:200, 14-9965-80), Invitrogen) and EpCAM (rabbit, dilution 1:1200, 14452, Cell Signaling Technology) were diluted in 10% goat serum and incubated overnight at 4 °C in a humidified chamber. Tissue specimens were washed 5x in TBST and secondary antibodies for the visualization of FOXJ1 (Alexa Fluor 647 goat anti-mouse, dilution 1:200, A21323, Invitrogen) and EpCAM (Alexa Fluor 555 goat anti-rabbit, dilution 1:200, A21428, Invitrogen), and SYTO 10 for nuclear visualization (10624243, Invitrogen) was applied for 1 hour at room temperature in darkness. Samples were washed 5x in TBST, followed by 2x in TBS and cover-slipped for high-content imaging.

High-resolution microscopy. Images of immunofluorescence-labeled cell cultures were acquired using an AxioImager Z.2 microscope (Zeiss), equipped with wide-field optics, a ×20, 0.8 NA dry objective and a quadruple-band filter set for Hoechst, FITC, Cy3 and Cy5 fluorescent dyes. Wide-field acquisition was performed using the Colibri 7 LED light source and an AxioCam 702 mono camera with 5.86 µm per pixel. Z-stacks with 19 z-slices were acquired at 3-mm increments to capture the optimal focus plane. Images were obtained automatically with Zeiss ZEN 2.6 (blue edition) at non-saturating conditions (12-bit dynamic range).

IHC images from salivary gland and melanoma tissue were obtained using the automated slide scanner Zeiss Axio Scan.Z1 for bright-field microscopy. Before slide acquisition was performed using the VLS 1260 machine (Leica) and an ECDF Hitachi HV-F202CLS camera. PEN slides were scanned with a ×20, 0.8 NA dry objective yielding a resolution of 0.22 mm per pixel. Z-stacks with eight z-slices were acquired at 2-mm increments to capture the optimal focus plane. Color images were obtained automatically with Zeiss ZEN 2.6 (blue edition) at non-saturating conditions (12-bit dynamic range).

Wide-field fluorescence microscopy for validation of cell-cycle-dependent Ctfr050 expression. Cells were imaged on a Leica Dmi8 wide-field microscope equipped with a 0.8 NA, ×40 air objective and a Hamamatsu Flash 4.0 V3 camera using LAS X software. The segmentation of each cell was performed using CellProfiler software using DAPI for nucleic segmentation. The mean intensity of the target protein and the cell cycle marker protein was measured in the nucleus. The cells were grouped into the G1 and G2 phases of the cell cycle by using the 0.2 and 0.8 quantile of ANLN or CCNB1 intensity levels in the nucleus, and cell-cycle-dependent expression of Ctfr050 was validated by comparing differences in expression levels between G1 and G2 cells.

Cell lines. The human osteosarcoma cell line U2OS was grown in DMEM (high glucose, GlutaMax) containing 10% FBS and penicillin–streptomycin (Thermo Fisher Scientific). The human osteosarcoma cell line U2OS was grown in DMEM (high glucose, GlutaMax) containing 10% FBS and penicillin–streptomycin (Thermo Fisher Scientific). Cells were imaged on a Leica Dmi8 wide-field microscope equipped with 0.8 NA, ×40 air objective and a Hamamatsu Flash 4.0 V3 camera using LAS X software. The segmentation of each cell was performed using CellProfiler software. The cell cycle marker protein was measured in the nucleus. The cells were grouped into the G1 and G2 phases of the cell cycle by using the 0.2 and 0.8 quantile of ANLN or CCNB1 intensity levels in the nucleus, and cell-cycle-dependent expression of Ctfr050 was validated by comparing differences in expression levels between G1 and G2 cells.
LMD. To excise cells or nuclei, we used the Leica LMD7 system, which we adapted for automated single-cell automation. High-cutting precision was achieved using an HC PL FLUOTAR L X63/0.70 (tissue) or X40/0.60 (cell cultures) CORR XT objective. We used the Leica Laser Microdissection V.8.2.3.7603 software (adapted for this project) for full automated excision and collection of contours. For FFPE tissue proteome analysis, we collected 50–100 cells per sample (total area collected × slide thickness / average mammalian cell volume of 2,000 µm³; BNID 100434), in agreement with estimations in spatial transcription analysis8.

Leica LMD7 cutting accuracy (Leica R&D, patent EP1276586)
For ×150 objective: \( \text{LMD} = 0.07 \mu m \)

Segmentation methods and accuracy evaluation. nucleAIzer \(^1\) models were integrated into BIAS and customized for these experiments by retraining and refining the nucleus and cytoplasm segmentation models. First, style transfer learning was performed as follows. Given a new experimental scenario such as our melanoma or salivary gland tissue sections stained immunohistochemically, the approach of which produces such an image that type that cannot yet exist, preventing efficient segmentation with even powerful DL methods.

With an initial segmentation or manual contouring by experts (referred to as annotation), a small mask dataset is acquired (masks represent, for example, nuclei), which is used to generate new (synthetic) mask images such that the spatial distribution, density and morphological properties of the generated objects (for example, nuclei) are similar to those measured on the annotated images. The initial masks and their corresponding microscopy images are used to train an image style transfer model that learns how to generate the texture of the microscopy images on the masks, marking objects using GANs (generative adversarial networks): foreground to mimic, for example, nuclei, and background for surrounding, for example, tissue structure. Parallel to that, either nuclei or cytoplasm objects were created and input to the image style transfer learning network that generated realistic-looking synthetic microscopy images with the visual appearance of the original experiment. Hence, with this artificially created training data (synthetic microscopy images and their corresponding, also synthetic, masks), their applied segmentation model, Mask R-CNN, is prepared for the new image type and can accurately predict the target compartment.

We benchmarked the accuracy of the segmentation approach on a fluorescent Lck-U2OS cell line as well as tissue samples of melanoma, salivary gland and fallopian tube and compared results to three additional methods, including two DL approaches—unet4nuclei (denoted as M1 in Fig. 2a and S1) \(^6\) and Cellpose (M2) \(^7\)—alongside a widely used, conventional adaptive threshold-based and object splitting-based software broadly used in the bioimage analysis community. We note that M1 is not intended for cytoplasm segmentation on various tissue proteome analysis, we collected 50–100 cells per sample (total area collected \( 63/0.70 (tissue) \) or \( 10–20 \text{ random images sampled from visually distinct regions for each sample type (U2OS cells and melanoma, salivary gland and fallopian tube tissues)} \) to show robustness, compared to ground truth annotations drawn by experienced imaging annotators \(^6\). We included images for each sample—for example, duct cells, acini cells, cells without any membrane staining and lymphocytes—in the salivary gland tissue, and similarly for the other samples as well, to ensure robustness. Outlines or contours of all visible objects (nucleus or cytoplasm) were drawn individually and then exported to mask images in the same format that the segmentation yielded (instance segmentation masks with increasing gray intensities by objects). The ground truth masks were solely used in evaluation; the aforementioned image style transfer learning was trained on automatically fetched masks of the new experiments. Considering the mean F1 scores measured, we conclude that the applied DL-based segmentation method \(^7\), available in BIAS produced segmentations on both nucleus and cytoplasm level in a higher quality than the compared methods (see results in Fig. 2a,b and Extended Data Fig. 1a).

\[
\text{Jaccard index} = \frac{|x \cap y|}{|x \cup y|} = \frac{|x \cap y|}{|x| + |y| - |x \cap y|} \\
\text{precision} = \frac{TP}{TP + FP} \\
\text{recall} = \frac{TP}{TP + FN} \\
F1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

Our evaluation results of nucleus and cell body segmentation on melanoma, salivary gland and fallopian tube epithelium tissues and U2OS cells is presented in Table 1.

These results correlate with our previous study \(^7\) that showed superior performance of nucleAIzer on various microscopy image data modalities (fluorescent cell culture, hematoxylin and eosin tissue and further experimental scenarios) compared to multiple segmentation approaches, including, for example, M1 and instakt \(^7\).

We also note that previous methods, such as CellProfiler or instakt, can perform accurate segmentation of cells; moreover, the performance of M1 on tissue nucleus segmentation is remarkable. On the other hand, robust methods (for example, DL-based) offer the convenience of not needing to reset parameters when working on images from a different sample or type.

Sample preparation for MS. Cell culture (nuclei or whole cells) and tissue samples were collected by automated LMD into 384-well plates (Eppendorf, 003012947). For the collection of different U2OS nuclei classes (Fig. 3 and Extended Data Figs. 2 and 3), we normalized nuclear size differences (resulting in different total protein amounts) by the number of collected objects per class. On average, we collected 267 nuclei per sample. For FFPE tissue samples of salivary gland and melanoma (2.5-µm-thick sections cut with a microtome), an area of \( 80,000–160,000 \mu m^2 \) per sample was collected for an estimated number of 100–200 cells based on the average HeLa cell volume of 2,000 µm³ (BNID 100434).

Next, 20 µl of ammonium bicarbonate (ABC) was added to each sample well, and the plate was closed with sealing tape (Corning, CLS6569-100EA). After vortexing for 10 seconds, plates were centrifuged for 10 minutes at 2,000g and heated at 95 °C for 30 minutes (cell culture) or 60 minutes (tissue) in a thermal cycler (Bio-Rad S1000 with 384-well reaction module) at a constant lid temperature of 110 °C. Then, 5 µl of 5x digestion buffer (60% acetonitrile in 100 mM ABC) was added, and samples were heated at 75 °C for another 30 minutes. Samples were shortly cooled down, and 1 µl of LysoG was added (pre-diluted in ultra-pure water to 4 ng/µl) and digested for 4 hours at 37 °C in the thermal cycler. Subsequently, 1.5 µl of trypsin was added (pre-diluted in ultra-pure water to 4 ng/µl) and incubated overnight at 37 °C in the thermal cycler. The next day, digestion was stopped by adding trifluoroacetic acid (TFA, final concentration 1% v/v), and samples were vacuum dried (approximately 1.5 hours at 60 °C). Then, 4 µl of MS loading buffer (3% acetonitrile in 0.2% TFA) was added, and the plate was vortexed for 10 seconds and centrifuged for 5 minutes at 2,000g. Samples were stored at −20 °C until liquid chromatography–mass spectrometry (LC–MS) analysis.

High-pH reversed-phase fractionation. We used high-pH reversed-phase fractionation to generate a deep U2OS cell precursor library for data-independent MS analysis (below). Peptides were fractionated at pH 10 with the spider-fractionator \(^7\). Next, 30 µg of purified peptides was separated on a 30-µm C18 column over 100 minutes and concatenated into 12 fractions with 90-second exit valve switches. Peptide fractions were vacuum dried and reconstituted in MS loading buffer for LC–MS analysis.

LC–MS analysis. LC–MS analysis was performed with an EASY-nLC-1200 system (Thermo Fisher Scientific) connected to a modified trapped ion mobility spectrometry quadrupole time–of–flight mass spectrometer with about five-fold-higher ion current (timsTOF Pro, Bruker Daltonik) with a nano-electrospray ion source (CaptiveSpray, Bruker Daltonik). The autosampler was configured for sample pick-up from 384-well plates.
Peptides were loaded on a 50-cm in-house-packed HPLC column (75-µm inner diameter packed with 1.9-µm ReproSil-Pur C18-AQ silica beads, Dr. Maisch). Peptides were separated using a linear gradient from 5–30% buffer B (0.1% formic acid and 80% ACN in LC-MS-grade water) in 55-Mutes followed by an increase to 60% for 5 minutes and a 10-minute wash in 95% buffer B at 300 nL/min. Buffer A consisted of 0.1% formic acid in LC-MS-grade water. The total gradient length was 70 minutes. We used an in-house-made column oven to keep the column temperature constant at 60°C.

Mass spectrometric analysis was performed as described in Brunner et al., either in data-dependent (ddaPASEF) (Fig. 1) or data-independent (diPASEF) mode (Figs. 2, 3 and 5). For ddaPASEF, one MS1 survey TIMS-MS and ten PASEF MS/MS scans were acquired per acquisition cycle. Ion accumulation and ramp time in the dual TIMS analyzer was set to 100 ms each, and we analyzed the ion mobility range from 1/K₀ = 1.6 V cm⁻² to 0.6 V cm⁻². Precursor ions for MS/MS analysis were isolated with a 2-T window for m/z 700 and 3-7 T for a total m/z range of 100-1,700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at 1/K₀ = 1.6 V cm⁻² to 20 eV at 1/K₀ = 0.6 V cm⁻². Single charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonik). Precursors for MS/MS were picked at an intensity threshold of 1,000 arbitrary units (a.u.) and re-sequenced until reaching a ‘target value’ of 20,000 a.u., taking into account a dynamic exclusion of 40-second elution. For data-independent analysis, we made use of the correlation of ion mobility with m/z and synchronized the elution of precursors from each ion mobility scan with the quadrupole isolation window. The collision energy was ramped linearly as a function of the ion mobility (from 59 a.u. at 1/K₀ = 1.6 V cm⁻² to 20 a.u. at 1/K₀ = 0.6 V cm⁻²). We used the diPASEF method for library generation.

Data analysis of proteomic raw files. Mass spectrometric raw files acquired in ddaPASEF mode (Figs. 2, 3 and 5) were analyzed with MaxQuant (version 1.6.7.0)43,44. The UniProt database (2019 release, UP000005640_9606) was searched with the precursor elution profile from the TIMS device. The collision energy in a total m/z range of 100-1,700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at 1/K₀ = 1.6 V cm⁻² to 20 eV at 1/K₀ = 0.6 V cm⁻². We used the diPASEF method for library generation.

Bioinformatic analysis. Proteomics data analysis was performed with Perseus45 and within the R environment (https://www.r-project.org/). MaxQuant output tables were filtered for ‘Reverse’, ‘Only identified by site modification’ and ‘Potential contaminants’ before data analysis. Data were stringently filtered to keep proteins with only 30% or less missing values (those displayed as 0 in MaxQuant output). Missing values were imputed based on a normal distribution (width = 0.3; downshift = 1.8) before statistical testing. PCA was performed in R. For multi-sample (ANOVA) or pairwise proteomic comparisons (two-sided unpaired t-test), we applied a permutation-based FDR of 5% to correct for multiple hypothesis testing. An s value of 0.1 was used for the pairwise proteomic comparisons. Pathway enrichment analysis was performed in Perseus (Supplementary Tables 2, 3, 5 and 9; Fisher’s exact test with Benjamini–Hochberg FDR of 0.05) or ClusterProfiler (Supplementary Tables 7 and 10), the ReactomePA package46 and the WebGestalt gene set analysis toolkit (WebGestaltR)47, with an FDR filter of 0.05, respectively. Minimum category size was set to 20 and maximum size to 500.

Microscopy and proteomics data integration. To visualize combined microscopy and MS-based proteomics results, we exported the spatial data files for each predicted class from the BIAS software. This export generates .xml output files with the geometry and location of cells within a class. We used Python to extract this information and aggregate it into a data frame. We then plotted the centroid (x, y) coordinates of each cell in a scatterplot and overlapped proteomics data. To visualize protein functional results in spatial context, we performed a REACTOME pathway enrichment analysis on the generated proteomics results and used normalized enrichment scores (z-scores) as a color gradient reflecting overrepresentation of a given pathway.

Data availability
The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository42 with the dataset identifier PXD023904. BIAS raw data, image raw data, a demo dataset and online material of how to install BIAS and reproduce our work can be accessed at the European Bioinformatics Institute BioStudies database43 (https://www.ebi.ac.uk/biostudies/) with the accession number S-BST820. We used the UniProt database (2019 release, UP000005640_9606, https://www.uniprot.org) for all mass spectrometric raw file searches.

Code availability
A free compiled version of BIAS with limited high-throughput capabilities is available at the BioStudies Archive (accession number S-BST820), containing all features applied in the described workflows. Several major components of our work are available in open-source repositories (Supplementary Table 11).

References

Acknowledgements
The authors thank M. Rykær, J. Madsen (NNF CPR Mass Spectrometry Platform, University of Copenhagen) and L. Drici (NNF CPR Proteomics Program) as well as J. Mueller (MPIB Munich) for technical assistance. We acknowledge F. Hoffmann, C. Greb and S. Schlaudraff from Leica for technical support; T. Danka and M. Kovacs
for fruitful scientific discussions; and T. Hartig Braunstein, P. Hernandez-Varas and C. Prats from the Core Facility of Integrated Microscopy for microscopy support. We thank J. Lukas for scientific support and guidance and J. Percival for the scientific illustrations (Illustration Ltd.). This work was supported by grants from the Novo Nordisk Foundation (grant agreements NNF14CC0001 and NNF15CC0001) and the Max Planck Society for the Advancement of Science and by the Chan Zuckerberg Initiative for partial funding of the cell cycle work (grant CZF2019-002448) to E. Lundberg, M.M. and P.H.. F.C. acknowledges the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement #67595 (Marie Skłodowska-Curie grant) and the German Ministry of Education and Research (BMBF), as part of the National Research Node ‘Mass Spectrometry in Systems Medicine’ (MSCoreSys), under grant agreement 16110222. B.D.A. acknowledges support from the Lundbeck Foundation (R252-2017-1414) and the Novo Nordisk Foundation (NNF20OC0065720). P.H., R.H., F.K., E.M. and A.K. acknowledge support from the LENDULET-BIOMAG Grant (2018-342), European Regional Development Funds (GINOP-2.2.1-15-2017-00072), H2020 (ERAPEMED-COMPASS, ERAPEMED-SYMMETRY, DiscovAIR, FAIR-CHARM), OTKA-SNN, TKP2021-EGA09 and ELKH-Excellence grants. E. Lengyel is supported by NIH R35CA264619 and the Chan Zuckerberg Initiative (CZIF2019-002435). We acknowledge S. Ito and H. Masai (Tokyo Metropolitan Institute of Medical Science) for providing the stable U2OS FUCCI cell line. The LCK-GFP plasmid was a gift from S. Green (Addgene, plasmid 61099).

Author contributions

Funding
Open access funding provided by Max Planck Society.

Competing interests
P.H. is the founder and a shareholder of Single-Cell Technologies Ltd., a biodata analysis company that owns and develops the BIAS software. The remaining authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01302-5.

Supplementary information
The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01302-5.

Correspondence and requests for materials should be addressed to Andreas Mund, Peter Horvath or Matthias Mann.

Peer review information Nature Biotechnology thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Benchmarking of segmentation algorithm. a, Cell body and nuclei segmentation of melanoma, salivary gland and fallopian tube tissue using the Biological Image Analysis Software (BIAS). We benchmarked the accuracy of our segmentation approach using the F1 metric and compared results to three additional methods M1-M3: unet4nuclei (M1), CellProfiler (M2), CellPose (M3), while OUR refers to nucleAIzer. Bars show mean F1-scores with SEM (standard error of the mean). Visual representation of the segmentation results: green areas correspond to true positive, blue to false positive and red to false negative. Data provided in Table 1 and Supplementary Table 1. b, BIAS allows the processing of multiple 2D and 3D microscopy image file formats. Examples for image pre-processing, deep learning-based image segmentation, feature extraction and machine learning-based phenotype classification. c, Left: Contour alignment in the LMD7 software before laser microdissection of fallopian tube epithelial cells. Middle: Screenshot after laser microdissection. Right: 384-well inspection after laser microdissection in individual fallopian tube epithelial cells. d, Number of quantified proteins per replicate of FOXJ1 positive and negative epithelial cells. Samples were acquired in data-independent mode and analyzed with the DIA-NN software. e, Replicate correlations of proteome measurements. Correlation values show Pearson correlations. f, Pathway enrichment analysis for proteins significantly higher in ciliated cells compared to secretory fallopian tube epithelial cells.
Extended Data Fig. 2 | PCA and loadings of cell culture classes at sub-cellular level and number of significantly changed proteins vs. class abundance.

**a**, Quantitative proteomic results of whole cell and nuclei replicates, and comparison between whole cells and nuclei. **b**, Principal component analysis (PCA) of whole cell ($n = 3$) and nuclei proteomes ($n = 3$). Proteins with the strongest contribution to PC1 are highlighted. **c**, Relative proportions of the six nuclei classes. **d**, Number of differentially expressed proteins (two-sided t-test, $n = 3$ biological replicates) compared to unclassified nuclei (bulk). Proteins with an FDR less than 0.05 were considered significant. **e**, Correlation between number of significantly regulated proteins per nuclei class vs relative class proportion. A linear model was fitted to the data showing an inverse correlation with Pearson $r = -0.96$ ($p$-value = 0.01). **f**, Relative protein levels (z-score) of known cell cycle markers across the five nuclei classes. All bar graphs represent mean of data ($n = 3$ biological replicates) and error bars are s.d. ANOVA $p$-values are shown.
Extended Data Fig. 3 | DVP discovers uncharacterized proteins with potential clinical relevance. a, Violin plots showing nuclear area in pixels of the 6 nuclei classes identified by ML. b, Nuclear area in pixels of U2OS FUCCI cells in relation to the cell cycle pseudotime. Color code indicates point density. c, Nuclear area of three major cell cycle states G1, G1/S and S/G2 determined by fluorescently tagged CDT1 and GMNN intensities and Gaussian clustering. Box plots show the results of n = 238,675 cells in total (85,551 for G1, 83,121 for G1/S and 70,003 for S/G2). d, Relative protein levels of all identified ORF proteins in the dataset. C7orf50, C1orf112, C19orf53 and C11orf98 were differentially expressed (ANOVA p-value < 0.05) across the 5 nuclei classes (n = 3 biological replicates). e, Mean intensities of immunofluorescent stained C7orf50 and the cell cycle markers ANLN and CCNB1 in U2OS cells. C7orf50 levels were quantified in nuclei with low and high ANLN and CNNB1 intensities. Box plots show the results of n = 263 cells per condition (C7orf50-ANLN) and n = 412 per condition (C7orf50-CCNB1). f, Upper panel: Representative immunofluorescence images of C7orf50 and DNA (DAPI) stained U2OS cells. Scale bar is 20 µm. Note, C7orf50 is enriched in nucleoli. Lower panel: Immunohistochemistry of a C7orf50 stained pancreatic adenocarcinoma (https://bit.ly/2X4re05). Image credit: Human Protein Atlas. Scale bar is 40µm. g, Kaplan-Meier survival analysis of pancreatic adenocarcinoma (https://bit.ly/3BAXleW) based on relative C7orf50 RNA levels (FPKM, number of Fragments Per Kilobase of exon per Million reads). RNA-seq data is reported as median FPKM, generated by The Cancer Genome Atlas (https://bit.ly/3SG88d). Patients were divided into two groups based on C7orf50 levels with n=41 low and n=135 high patients. A log-rank test was calculated with p = 0.0001. h, String interactome analysis for C7orf50. A high confidence score of 0.7 was used with the five closest interactors highlighted by color. The box plots in c and e define the range of the data (whiskers), 25th and 75th percentiles (box), and medians (solid line). Outliers are plotted as individual dots outside the whiskers.
Extended Data Fig. 4 | DVP applied to archival tissue of a rare salivary gland carcinoma. a, Immunohistochemical staining of normal salivary gland stained for the cell adhesion protein EpCAM. Supervised (random forest) ML was trained to identify acinar (green) and duct cells (turquoise). Scale bar = 20µm. b, Quantitative proteomic comparison between acinar and duct cells from tissue in A with known cell type specific markers highlighted (https://bit.ly/3iOK8Q1). c, Relative protein levels of selected pathways that were significantly higher in acinar or duct cells. d, Unsupervised hierarchical clustering of acinar and duct cell proteomes from two different patients together with acinar cell carcinoma cells. Note that normal acinar cells of two different tissues clustered together. Duct cells clustered furthest away. Prior to clustering, protein levels from different sample groups (duct cell tissue #1, acinar cell tissue #1, acinar cell tissue #2, carcinoma tissue #2) were averaged and z-scored. Bar on the left shows differentially expressed pathways from panel b with acini and duct specific proteins in green and turquoise, respectively.
Extended Data Fig. 5 | See next page for caption.
Extended Data Fig. 5 | DVP applied to archival tissue of primary melanoma. a, Isolation of tumor adjacent SOX10 positive melanocytes from a cutaneous melanoma tissue. Left: Contour alignment before laser microdissection. Right: Inspection after laser microdissection. b, Number of protein quantifications per sample type with \( n = 4 \) (melanocytes), \( n = 5 \) (stroma), \( n = 5 \) (melanoma in situ) and \( n = 13 \) (melanoma) independent replicates. Bar graphs represent mean of data and error bars are s.d. Samples were acquired in data-independent mode and analyzed with the DIA-NN software. c, Upper panel: Heatmap from Fig. Sh shown with identified protein clusters (color bar). Unsupervised hierarchical clustering based on all 1,910 ANOVA significant (FDR < 0.05) protein groups. Protein levels were z-scored. Lower panel: Pathway enrichment analysis of different row clusters obtained by unsupervised hierarchical clustering. The ReactomePA package was used for enrichment analysis with an FDR cut-off of 0.05 for all enriched terms. d, Relative levels (z-score) of proteins related to the KEGG term ‘melanogenesis’. Note, melanocytes show highest protein levels. The box plots define the range of the data (whiskers), 25th and 75th percentiles (box), and medians (solid line). Outliers are plotted as individual dots outside the whiskers. e, Pathway enrichment analysis of proteins up or down-regulated in vertical versus radial growth melanoma cells. Enrichment results were obtained with the ClusterProfiler R package based on an FDR < 0.05.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a  Confirmed

☐   The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐   A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐   The statistical test(s) used AND whether they are one- or two-sided
   *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐   A description of all covariates tested
☐   A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐   A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐   For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
   *Give P values as exact values whenever suitable.*
☐   For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐   For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐   Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

*Our web collection on statistics for biologists contains articles on many of the points above.*

Software and code

Policy information about availability of computer code

Data collection

Proteomic mass spectrometry data were acquired using Compass Hystar software (Bruker Daltonik GmbH, Germany, version 6.0). AxiolImager Z2.2 microscope (Zeiss, Germany) for immunofluorescence microscopy. Zeiss Axio Scan.Z1 (Zeiss, Germany) for brightfield microscopy.

Segmentation evaluation was performed on 10 or 20 randomly selected images sampled from visually distinct regions for each sample type (U2OS cells, melanoma, fallopian tube, or salivary gland tissues) to show robustness, compared to ground truth annotations drawn by experts using Annotator (Holland, R., et al., Mol. Biol. Cell (2020)). The following R packages were used: ReactomePA (version 1.30.0), ggplot2 (version 3.3.1.9000)

Data analysis

Proteomics: MaxQuant (version 1.6.7.0) for dda-PASEF raw files, DIA-NN (version 1.8) for dia-PASEF raw files, statistical and bioinformatics analyses were done using Perseus (version 1.6.2.3) or R Studio (version 1.2.5033, R version 3.6.0).

Microscopy: ZEN blue acquisition software (ZEISS, 2.6.)

Image analysis, phenotyping and single cell isolation: BIAS, Biological Image Analysis Software (ver. 20/12/2020, Single-Cell Technologies Ltd.)

Segmentation evaluation: Matlab [9.4.0.813654 (R2018a)].

Visualization of microscopy and proteomics results was done using Python 3.7.9., Pandas 1.3.0, Geopandas 0.8.1, Plotly express 0.4.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data.
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD023904 (Username: reviewer_pxd023904@ebi.ac.uk, Password: vRAviOml).
Numerical source data underlying all graphical representations for Figs. 2b, d, f, g, h; 3c, e, g, h, i, j, 4d,e; 5e, f, g, h, i, j, l and Extended Data Figs. 1a, d, e, f; 2a, b, c, d, e, f; 3a, b, c, d, e, g; 4a, c, d, 5b, c, d, e have been co-submitted with the manuscript as freely accessible files. Primary BIAS imaging data in Figs. 2a,b,e; 3a, d, f; 4a, b, c; 5b, c, d, k and Extended Data Figs. 1a,c, 3f; 4a; 5a have been deposited at BioStudies Archive (accession number S-BSTS820). There are no restrictions on data availability.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
Sample size calculations were not performed. Experiments were repeated in minimum biological triplicates unless noted otherwise in the corresponding figure legend.

Data exclusions
No collected data were excluded.

Replication
All attempts at replication were successful. Experimental results were confirmed with biological triplicates unless otherwise noted.

Randomization
Randomization of sample groups was performed for mass spectrometric data acquisition. For all other data, no randomization was performed.

Blinding
Investigators were not blinded to sample group allocation during data collection.

Reporting for specific materials, systems, and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

<table>
<thead>
<tr>
<th>n/a</th>
<th>Involved in the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>Antibodies</td>
</tr>
<tr>
<td>☐</td>
<td>Eukaryotic cell lines</td>
</tr>
<tr>
<td>☑</td>
<td>Palaeontology and archaeology</td>
</tr>
<tr>
<td>☒</td>
<td>Animals and other organisms</td>
</tr>
<tr>
<td>☑</td>
<td>Human research participants</td>
</tr>
<tr>
<td>☐</td>
<td>Clinical data</td>
</tr>
<tr>
<td>☐</td>
<td>Dual use research of concern</td>
</tr>
</tbody>
</table>

Methods

<table>
<thead>
<tr>
<th>n/a</th>
<th>Involved in the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>☑</td>
<td>ChIP-seq</td>
</tr>
<tr>
<td>☐</td>
<td>Flow cytometry</td>
</tr>
<tr>
<td>☑</td>
<td>MRI-based neuroimaging</td>
</tr>
</tbody>
</table>

Antibodies

Antibodies used:
- Antibody for immunofluorescence (IF) or immunohistochemistry (IHC):
  - Rabbit anti-C7orf150 at 1 ug/ml (HPA052281, Atlas Antibodies)
  - Mouse anti ANLN at 1.25 ug/ml (amab00652, clone CL0303, Atlas Antibodies)
  - Mouse anti CCNB1 at 1 ug/ml (610220, BD Biosciences)
- EpCAM (Nordic BioSite, Copenhagen K, Denmark; clone B514, cat. #BHH-7402-1, dilution 1:400)
- SOX10; Nordic BioSite, clone 857 cat. #BHH-7959-1, dilution 1:200
- CD146; Cell Marque, Rocklin, CA, USA, clone EP54, cat. #AC-0052, dilution 1:400
- Src/ß-Src; Cell Signaling Technology, Danvers, Massachusetts, USA, clone 36010, # 2108, dilution 1:3200
Validation

Detailed description about IF antibody validation can be found here:

FOX11: https://bit.ly/3zw8Oab
EpcAM: https://bit.ly/3N1U2A

C7orf50: Species reactivity with human samples was verified. The antibody was validated for IF according to the workflow within the Human Protein Atlas project. See validation information here: https://www.proteinatlas.org/ENSG00000146540-C7orf50/antibody; PMID: 28495876.

ANLN: Species reactivity with human samples was verified. The antibody was validated for IF as part of the Human protein Atlas by staining similarity with an independent antibody. See validation information here: https://www.proteinatlas.org/ENSG00000114266-ANLN/antibody. The antibody has been successfully used in an IF assay in the following article. (PMID: 30103211)

CCNB3: PMID: 29052541

All IHC antibodies are assessed and validated by NovoQC (https://www.novoqc.org/about.php).

After washing and blocking of endogenous peroxidase activity, the reactions were detected and visualized using Envision FLEX+ High pH kit (Agilent/Dako, #G8000+G8090/G8211) according to the manufacturer's instructions.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)
The commercial source for the human osteosarcoma cell line U2OS is ATCC (American Type Culture Collection), cell line order number HTB-96. https://www.atcc.org/products/htb-96. U2OS Fucci cell line was provided by Dr. Sayuri Ito and Dr. Hisao Masai (Tokyo Metropolitan Institute of Medical Science). A clonal U2OS cell line with homogenous and moderate expression levels of Lck-EGFP at the plasma membrane was established from a single colony.

Authentication
All cell lines were authenticated by STR profiling (Identigene Molecular Diagnostics).

Mycoplasma contamination
All cell lines were tested negative for mycoplasma (MycoAlert, Lonza).

Commonly misidentified lines
(See CLAC register)
Cell lines used in this study were not listed in the commonly misidentified category.

Human research participants

Policy information about studies involving human research participants

Population characteristics
not relevant

Recruitment
No patient recruitment was done for the study. Biobanked (FFPE) patient material was used for retrospective studies.

Ethics oversight
We collected archival FFPE tissue samples of salivary gland acinic cell carcinoma and melanoma from the Department of Pathology, Zealand University Hospital, Roskilde, Denmark. Melanoma tissue was from a 51-year-old male and located at the left upper chest. TNM stage at diagnosis was T3aN1M0. Histological subtype was superficial spreading melanoma (SSM), Clark level was 4 and Breslow thickness measured 2.27 mm. Tumor immune infiltration was categorized as non-bright. The FFPE sample was 17 years old. The patient experienced recurrence at different locations 17 months after diagnosis and died after 71 months. The acinic cell carcinoma was removed from the right parotid gland on a 29-year-old male. There was no sign of mitosis, necrosis dedifferentiation, perineural or intravascular growth. The tumor cells were positive in EpcAM, CK7, DOG1 and SOX10. Mammaglobin was negative. The sample was four years old and the patient is currently disease free. The study was carried out in accordance with the institutional guidelines under approval by the local Medical Ethics Review Committee (S1-742), the Data Protection Agency (REF-066-2019) and in agreement with Danish law [Medical Research Involving Human Subjects Act]. The fallopian tube tissue shown in Fig. 2 is from a 64-years-old female and was macroscopically and histologically normal appearing. All patients have consented before surgery, and patient-derived tissues were obtained fresh or were paraffin embedded according to an approved Institutional Review Board (IRB, 133728) protocol at the University of Chicago hospital. In accordance with the Medical Ethics Review Committee approval, all FFPE human patient tissue samples were exempted from consent as these studies used existing archived pathological specimens. Human tissue specimens were assessed by a board-certified pathologist.

Note that full information on the approval of the study protocol must also be provided in the manuscript.