horvath.peter.2_10_22

Life beyond the pixels: single-cell analysis

Peter Horvath

Synthetic and Systems Biology Unit
Institute of Biochemistry
Biological Research Center
Eötvös Loránd Research Network

Szeged

2022

To my Family who has always been by my side and let me fly.

“Two roads diverged in a wood, and I –
I took the one less traveled by,
And that has made all the difference.”

“Az erdőben egy útelágazáshoz értem, s én –
Én a kevésbé jártat választottam,
S ez volt minden különbség.”

Robert FROST – The Road Not Taken (1920)

Acknowledgements

Firstly, the work presented in this thesis is not only my achievement. This is a joint scientific achievement involving my supervisors, the members of my groups, the BIOMAG (Biological Image Analysis and Machine Learning group) family in Szeged and Helsinki, and many collaborators in Zurich, Nice, Boston, Helsinki, Szeged, Stockholm, Munich, Copenhagen and many more.

I am really grateful for knowing, working with and learning from these people.

My Family has been by my side from early childhood to date, and has taught me the value of hard work and learning. I have always had the opportunity to be engaged with my profession, as my base camp has always been there to support me. Here I apologize to them for spending probably more time with science and career-building than with storytelling in the evenings.

Finally, I am especially grateful to the Biological Research Centre, where my colleagues have created a fantastic scientific environment and gave me the opportunity to progress.

Table of contents

CHAPTER 1.

Introduction

The journey from Fourier wave theory to single cell proteomics may not sound straightforward and cohesive, but I would like to invite you, my dear Reader, to take this journey with me. I will show you how image analysis methods, originally developed to detect trees on aerial images, can be modified to revolutionize single cell analysis, how deep learning methods will identify cancer and how artificial intelligence can control a needle to communicate with a single cell of a human brain.

Much of the current understanding of biology, including many models of cell networks and signaling, is based on population-level averaged measurements (Altschuler and Wu 2010). However, measurements that average the behaviour of large cell populations can lead to misleading conclusions if they mask the presence of rare but critical subpopulations (Pelkmans 2012). It is now well known that heterogeneities, even within small subpopulations, can have important consequences for the population as a whole. Genetic heterogeneity, for example, plays a crucial role in drug resistance and tumour survival (Heppner 1984). Even genetically homogeneous cell populations have a high degree of phenotypic intercellular variability due to individual gene expression patterns (Tay et al. 2010). To better understand heterogeneous biological systems, we increasingly rely on single-cell molecular and morphological analysis methods (Strack 2022). Cell separation and isolation techniques available to collect single cells for further molecular studies and, most importantly, methods to infer the behavior, structure and function of biological systems based on single-cell molecular and morphological properties are still challenging and lack a solution (Giladi and Amit 2017).

My research aims to develop techniques for single-cell microscopy, morphological and phenotypic analysis, single-cell isolation and molecular analysis for rapid, reliable and versatile characterization of cells of interest, even in the most challenging cellular microenvironments such as the human cerebral cortex, pediatric brain tumor spheroids in 3D, or pathological tissue sections.

1.1 Single cell image analysis techniques in microscopy (Thesis 1)

Image analysis transforms digital images into measurements that describe the state of every single cell in an experiment. This process makes use of various algorithms to compute measurements that can be organized in a matrix in which the rows are cells in the experiment, and the columns are extracted features.

Every image acquired by a microscope exhibits inhomogeneous illumination mainly because a nonuniform light source or optical path often yields shading around edges. This effect is often underestimated; however, intensities usually vary by 10–30%, thus corrupting accurate segmentation and intensity measurements. Illumination correction is a process to recover the true image from a distorted one (Smith et al. 2015). Three major types of approaches for illumination correction exist; (1) Prospective methods develop correction functions from reference images, such as dark and bright images with no sample in the foreground. This approach requires careful calibration at the time of image acquisition (Singh et al. 2014). (2) Retrospective single-image methods calculate the correction model for each image individually (Babaloukas et al. 2011). However, the result may change from image to image, and thus may alter relative intensity. (3) Retrospective multi-image methods establish the correction function by using the images acquired in the experiment. These methods are often based on energy-minimization models (Peng et al. 2017). Illumination correction is an important step towards high-throughput quantitative profiling; in most of the laboratories, the strategy of choice is retrospective multi-image correction function.

Typically, each cell in the image is identified and assessed individually; that is, its constituent pixels are grouped to distinguish the cell from other cells and from the background. This process is called ‘segmentation’, for which two main approaches exist. (1) In model-based approaches, the experimentalist chooses an appropriate algorithm and manually optimizes parameters based on the visual inspection of segmentation results. A common procedure includes identifying nuclei first, as it is often a straightforward task, and then the results are utilized as seeds for the identification of the cells’ contours. This approach requires *a priori* knowledge (i.e. a 'model') (for details see Section 2.1), such as the objects’ expected size and shape (Molnar et al. 2016). Model-based approaches typically involve histogram-based methods, such as thresholding, edge detection, and watershed transformation. (2) In case of machine learning-based approaches a classifier is trained to find the optimal segmentation solution by providing it with ground-truth data and manually indicating which pixels of an image belong to different classes of objects (Sommer et al. 2011)(Reka Hollandi et al. 2020). This approach typically involves applying various transformations to the image in order to capture different patterns in the local pixel neighborhood. Segmentation is ultimately achieved by applying the trained model to new images to classify pixels accordingly.

In a review by (Caicedo et al. 2017) we gave an insight into image and data analysis strategies used for profiling cells.

1.2 Single cell phenotyping (Thesis 2)

One of the greatest achievements of science was the complete sequencing of the human genome (Liesegang 2001). Today, many believe, including the Human Cell Atlas consortium, that the next great challenge in biology lies in phenomics, i.e. the quantification of the set of phenotypes that fully characterizes an organism (the phenome) (Houle, Govindaraju, and Omholt 2010). Phenotype is defined as the observable characteristics of an organism, including its morphology, biochemical properties, behavior, etc. defined as a totality. By collecting and analyzing rich phenotypic data, we hope to better understand how genetic and environmental factors cause changes in the organisms or in their behavior, and may become capable of better predicting important outcomes such as fitness, reproduction, disease, carcinogenesis, resistance or mortality. Unlike genome sequencing, a complete understanding of the phenome is impossible with current technologies. In the efforts to understand the phenom, it is critical to make intelligent choices about what to measure and what phenomics tools to use.

Imaging is a fast and flexible technology for studying phenomics. Spatial and temporal information can be recorded with high accuracy and on an extremely wide scale. It implicitly represents the morphological features of the cell, and labelling technologies such as fluorescent labels allow localization of subcellular structures, proteins, and other molecules. Images can be taken at low cost and quickly, enabling large-scale screening experiments. Recent advances in microscopy, automation, and computing have dramatically increased our ability to take images rich in phenotypic information. Also, images can now be generated by orders of magnitude faster than they could be examined manually. Consequently, for a phenotypic readout, we rely on phenotypic image analysis techniques, i.e. computational methods that transform raw image data into meaningful phenotypic information. We have presented a comprehensive review of this field in Smith *et al* (Smith et al. 2018).

My group has developed a number of tools and software packages that intelligently help to determine the phenotype of certain cells. Each of these research projects aims to address the following two questions.

- 'Have I entirely discovered my data? Or at least partially?'
- 'Is my analysis as accurate as possible?'

As previously discussed, modern computational technologies are now capable of producing an unimaginable amount of data in a short period of time. In many applications, this includes hundreds of thousands of images, billions of single cells and often a few hundred unique features of each cell. With such a large amount of data, answering the two main questions above is not trivial anymore.

In the early 2010's we have introduced the software Advanced Cell Classifier (ACC, available at www.cellclassifier.org) (Horvath et al. 2011) with the aim of providing easy access to machine learning methods for field experts performing single-cell phenotyping. This software was later extended (ACC2.0) to answer the above questions by utilizing intelligent phenotype finder and active learning methods (Piccinini et al. 2017). Over the past decade, different versions of ACC were exploited in the discovery of new drugs and genes, in resolving fundamental biology related issues, as well as in developing individualized therapies. Prestigious publications appeared in journals like Cell, Science and Nature journals (see webpage for a reference list).

As a significant milestone, my research groups have laid the foundations of the theoretical framework that emphasizes the importance of cellular microenvironment, i.e. the phenotypes of cells can be determined more precisely if we know their microenvironment. Even our early method, in which only simple features and neighborhood definitions were taken into account, resulted in remarkable improvement (Toth et al. 2018). The extension of this method, which combined fisheye transformation with deep learning, yielded a highly significant improvement in single-cell phenotyping, both in tissue and cell culture samples (Toth et al. 2022 in press).

Moreover, we have discovered that cells almost continuously undergo dynamic changes, which also entails a variation in their phenotype. In my thesis, I give a detailed presentation of our proposed solution to execute a more abstract quantification of the continuous variation of individual cells by utilizing artificial intelligence. For this task, we have introduced the regression plane concept (Szkalisity et al. 2021).

1.3 Single cell isolation (Thesis 3)

Numerous single-cell isolation methods can be used to characterize or collect cells based on certain characteristics. These include fluorescence-activated cell sorting (FACS), immunomagnetic cell sorting, microfluidics and limiting dilution. However, these collection techniques disrupt and dissociate cells from their microenvironment, and cannot target cells based on their location within the sample or based on their morphological profile. In contrast, micromanipulation (e.g. patch clamping), laser capture microdissection (LCM) (Espina et al. 2006), imaging mass spectrometry (IMS) (Bodzon-Kulakowska and Suder 2016) or Raman spectroscopy (Schie and Huser 2013) are microscopy-based alternatives that can directly analyze individual cells within solid tissue samples.

My research groups have combined submicron resolution imaging, single-cell phenotyping and isolation supported by artificial intelligence (AI) with electrophysiology, sequencing, and ultra-sensitive proteomics workflows. In this task, the most important challenge lies in precisely defining single-cell boundaries and phenotypic cell classes. Our software tools correlatively combine classical, laser microdissection (LMD) and patch clamp microscopy. It 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 artificial intelligence before being subjected to automated omics profiling. The data we generate this way can be mined to discover transcriptomic and protein signatures that provide molecular insights into diversities at the phenotypic level of the genome, proteome or other *omes*, while preserving full spatial information.

We have recently introduced four families of single cell isolation methods, namely (1) computer aided microscopy imaging (CAMI) (Brasko et al. 2018), which is a machine learning-based single-cell isolation technique that uses laser microdissection from a 2D microenvironment; (2) the AutoPatcher system, which is able to target cells in a 3D live cell environment fully automatically, by using a patch clamp technique combined with deep learning (Koos et al. 2021); (3) Deep Visual Proteomics, which derives from the CAMI system and is applicable for ultrasensitive proteomics (Mund et al. 2022); and (4) Mito-Raman, which uses intelligent microscopy and Raman spectroscopy to perform automatic cell mitosis profiling (Voros *et al,* in prep.). These methods build on the image analysis and deep machine learning techniques we have developed so far. We aim to further improve them by incorporating novel methodologies, hoping to make these single-cell isolation methods the most suitable approaches for each dedicated modality. A more detailed prescription of methods (1), (2) and (3) is presented in Chapter 4.

CHAPTER 2

Single cell image analysis techniques in microscopy

> When you are so lucky that your PhD work – which focused on satellite image analysis – can directly be applied in your first paid job, it is great. This happened to me, and I was even luckier, as my first PhD student took over that early work and further developed it to detect overlapping single cells in microscopy images. We scored best in class. BUT, science never sleeps, or more precisely, scientists never sleep (not much at least) and we have eye-witnessed the revolution of deep learning. In 2-3 years we have completely rewritten all the achievements of the last two decades in microscopy imaging. Let me, my dear Reader, introduce you to this exciting journey, and show you the little pieces of bits we have contributed to this field.

In case you are not a comfortable user of mathematical statistics, it is sufficient to read the first paragraph of section 2.1 for the understanding of this chapter. The reason for the latter of section 2.1 is to show the Author's view on how image analysis tasks can be formulated and solved within a probability theory framework.

2.1 Introduction to a probabilistic view – Is it all about priors?

The challenge of single cell segmentation can be viewed as a special case of a general image understanding problem: the task is the identification of region R in the image domain corresponding to some entity or entities in the scene.

In order to solve this problem in any particular case, we have to construct, even if only implicitly, a probability distribution on the space of regions $P(R|I, K)$. This distribution depends on the current image data I, and on any prior knowledge K we may have about the region or about its relation to the image data, as encoded in the likelihood $P(I|R, K)$ and the prior $P(R|K)$ appearing in the Bayes' decomposition of $P(R|I, K)$ (or equivalently in their energies $-\ln P(I|R, K)$ and $-\ln$

P(R|K)). Then this probability distribution can be used to make estimates of the region we are looking for.

In the algorithmic solution of realistic problems, the prior knowledge K, and in particular prior knowledge about the 'shape' of the region, as described by P(R|K), is critical. The single cell nucleus extraction problem provides a good example: R takes the form of a collection of approximately circular connected components of similar size. There is thus a great deal of prior knowledge about the region sought. Then the question is how to incorporate such prior knowledge into a model for R. If the prior knowledge included in the model is not sufficient, it should necessarily be provided by the user.

I find it especially useful to present data analysis in biomedical imaging in a probabilistic way, as every computational method in this thesis is about the engineering of K, the prior knowledge, into image analysis models. To maximize P(R|K), or in other words to find the best solution given the image and our prior knowledge (e.g. best segmentation of single cells, or best illumination corrected image), we use a great range of optimizers (e.g. energy minimization, deep learning training). This chapter of the thesis presents a range of possibilities for the incorporation of K. First, we introduce CIDRE (Smith et al. 2015), an energy minimization framework to correct illumination problems in microscopy images. The major essence of CIDRE is that K is a composition of three types of priors, and while maximizing P(R|K) we used a robust solver that optimizes correction values for each pixel (i.e. iteratively approximates the optimal solution of an equation system for millions of variables simultaneously). In this way CIDRE provides state-of-the-art image correction, even in case of a low number of images. Second, we introduce higher-order active contours (HOACs), a differential geometry framework to incorporate long range interactions into image segmentation, and describe shapes with predefined properties, such as circles or ellipsoids (Horváth et al. 2009). Interestingly, here, the construction of the energy functional itself and its appropriate parameterization will give us K. We extended HOAC models for single cell segmentation, even in case of very dense and overlapping cells. Finally, we will present the nucleAIzer model, a deep learning framework we originally developed for the Data Science Bowl competition and reached the highest score at the time of its publication (Reka Hollandi et al. 2020). The idea we presented here is a learning model that adapts to the data itself, and creates learning examples to fine-tune itself. Such self-adaptive systems are still rare, and have a very interesting methodology to represent K, the prior knowledge. We will also present our recently developed augmentation techniques to artificially extend K, i.e. to generate more versatile training sets.

2.2 Illumination correction methods

Relevant publication:

Smith K., Li Y., Piccinini, F., Csucs G., Balazs C., Bevilacqua A., **Horvath, P.** (2015)
CIDRE: an illumination-correction method for optical microscopy.
Nature Methods

No optical system is perfect. Manufacturing defects, poor settings, vignetting and uneven light sources all contribute to uneven illumination in every image we take. The amount of distortion can vary due to a number of factors, but even seemingly minor shifts in illumination can lead to undesirable effects in photostitching or bias measurement (Fig. 1).

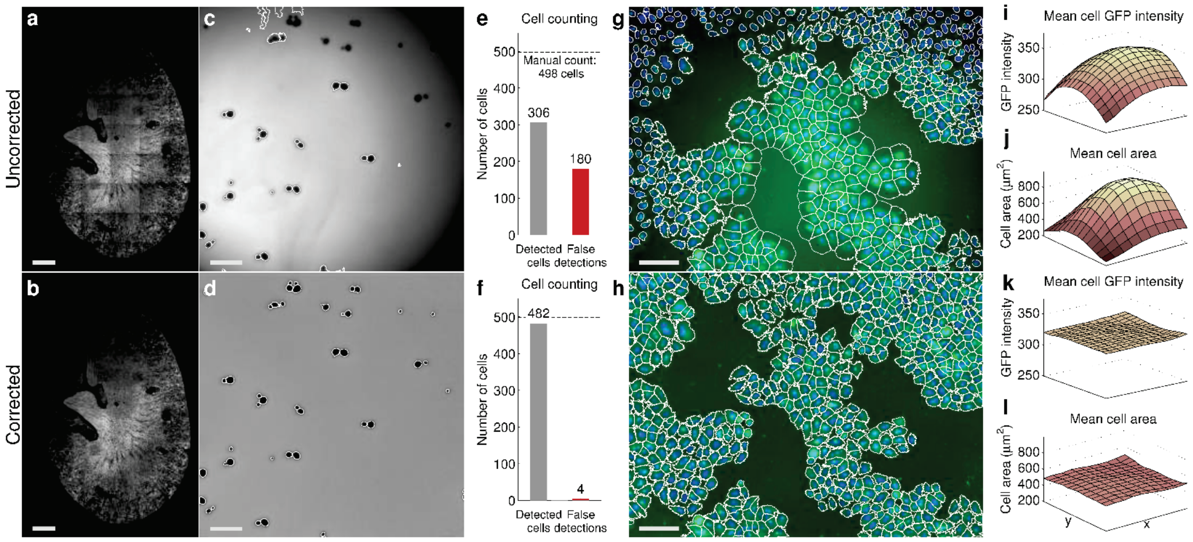


Figure 1. Uneven illumination can adversely affect measurements. Original images appearing in the top row suffer from uneven illumination. The same images appear below after being corrected using the CIDRE approach. **(a-b)** A 9x6 photostitched image of a mouse kidney section. **(c-d)** A stained medium containing yeast cells. Detections from CellProfiler are marked in yellow. **(e-f)** Results from an automatic cell counting algorithm on 40 images similar to (c) and (d). **(g-h)** HeLa cells from a high content screen. Detections are marked in yellow. **(i-l)** Measured intensity and cell area for each region of the image averaged over 3,456 images similar to (g) and (h). Image source: (Smith et al. 2015)

Although measurements from images are routinely used in research, scientists tend to underestimate the consequences of illumination distortion. For example, a task as simple as automatic cell counting can become unreliable as a result of uneven illumination (**Fig. 1c-f**). In a test using 40 images of yeast cells in a fluorescent medium, a 36% increase in missed detection rate was attributed to uneven illumination (**Fig. 1e-f**). False detections increased by 35%. When asked to estimate these increases, nearly three-quarters of respondents believed they would be less than 10%.

Because of the difficulties perceiving gradual illumination change, researchers often misjudge the magnitude of intensity loss. Our experiments on ten datasets representing ordinary microscope setups revealed that between 10% and 40% less light is typically recorded at the corners of the image than in the center. For sensors with a large field-of-view, the attenuation is even more severe. Novel cameras with larger sensor areas (sCMOS and EMCCD), which have become popular in next generation microscopes, routinely lose half the intensity.

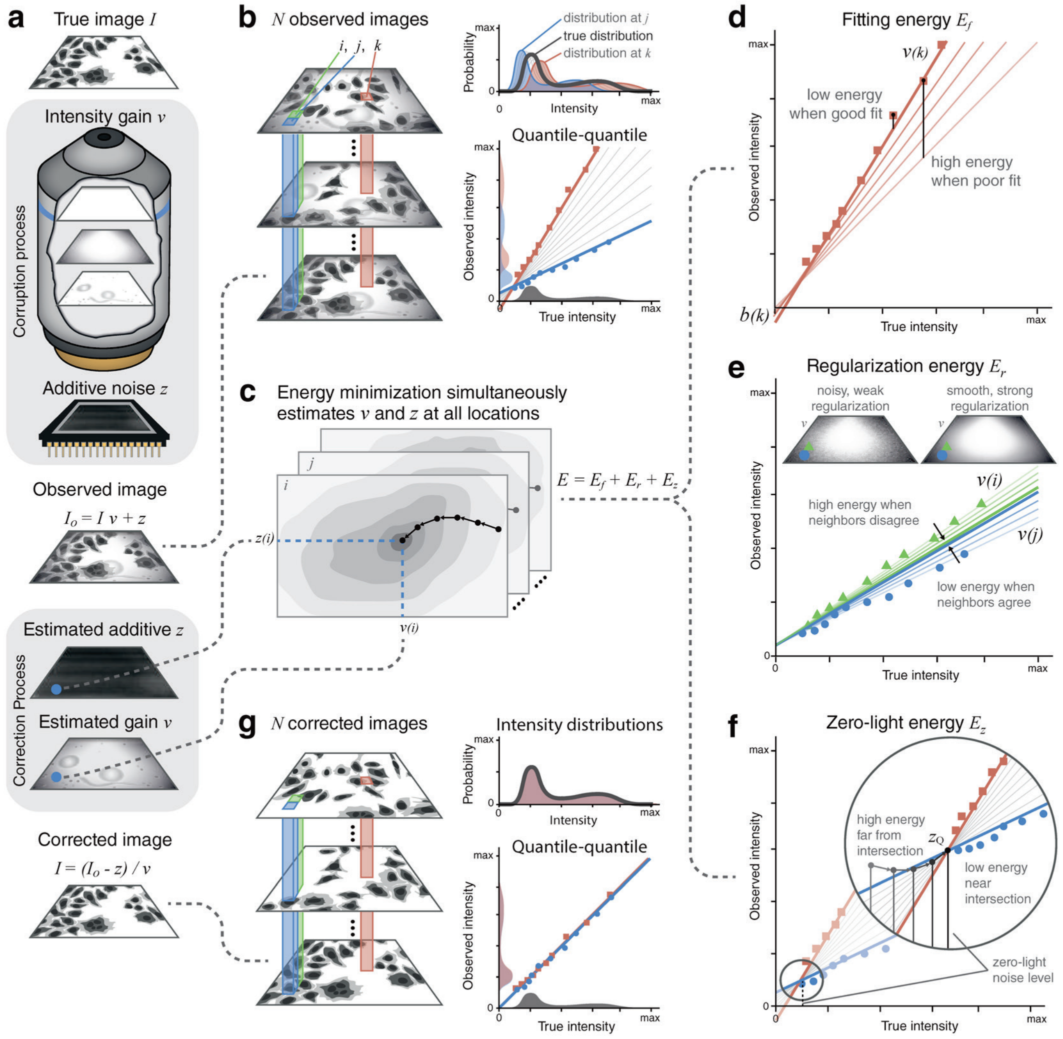


Figure 2. CIDRE is a retrospective method for illumination correction that does not require calibration images. (a) The true image is corrupted by misaligned optics, vignetting, dust, etc., which is modeled as a linear intensity gain function v. The observed image also contains thermal and readout noise modelled as an additive term z. **(b)** To recover the true image, we consider the local distributions of observed intensities collected from many images (red and blue). Each is related to the true underlying distribution of the specimen (gray) by a linear transform parametrized by v and b which correspond to slope and y-intercept in a Quantile-Quantile plot between the true distribution and the local distribution. **(c)** We estimate these parameters simultaneously for all locations by minimizing a robust regularized energy function composed of several terms. **(d)** The first is a robust regression term that ensures parameter values fit the data. **(e)** A regularization term reduces noise and guarantees a smooth correction surface by forcing neighboring distributions (green) to agree on similar parameter values. **(f)** The third term estimates z, the intensity recorded by the sensor when no light is present. Because the z surface is nearly flat, it can be estimated by finding the common point where all regression lines intersect. **(g)** Applying the reverse

transform using estimates of v and z, local intensity distributions take the shape of the true distribution, and the uncorrupted image is recovered. Image source: (Smith et al. 2015)

We reviewed the causes of uneven illumination. Images recorded by the sensor are corrupted by vignetting, misaligned optics, a non-uniform light source, or obstructions such as dust (Fig. 2a). These distortions can be modelled by a linear intensity gain function (v) that attenuates the signal from the true uncorrupted image. Thermal noise and readout noise make contributions independent of the signal. Thermal noise is generated by heat produced by the system, while readout noise occurs in the analog-to-digital conversion process. We call these additive noise sources the zero-light noise (z) because they can be estimated by capturing dark frame images with the shutter closed. Thus, the process of image formation can be described by Io = I*v+z where Io is the intensity value observed by the sensor. The reverse process can be used to recover the uncorrupted image

$$I = \frac{I_o - z}{v}$$

The correction process seems trivial at the first glance, but this appearance is deceptive because it is impossible to know v and z exactly. Correction methods currently practiced by the community deal with this issue in several ways. Prospective methods form empirical estimates of v and z from special calibration images that must be collected during image acquisition. Retrospective single-image methods focus on smoothing the appearance of the image, while other retrospective methods build a correction model from groups of images, although they either neglect part of the image formation process or ignore it altogether. We considered twelve common correction methods described in (Smith et al. 2015). Of these twelve correction methods, only the gold standard methods have the potential to obtain the true correction model. However, they require special calibration images to do so. The rest either ignore the zero-light term or simply smooth the image without regard for the relative intensity of objects in the image. For details see (Smith et al. 2015).

We used ten datasets representing a variety of typical microscopy setups using different microscopes, methods of sample preparation, staining, light sources, magnifications, and types of sensors (Fig. 3). Measuring the correction quality is difficult because we do not have access to the uncorrupted image. As a solution, we proposed to collect hundreds of pairs of overlapping images for each dataset, precisely align them, and measure the disagreement in the overlapping regions. The scores we report are these values normalized by the disagreement between uncorrected image pairs (Fig. 3).

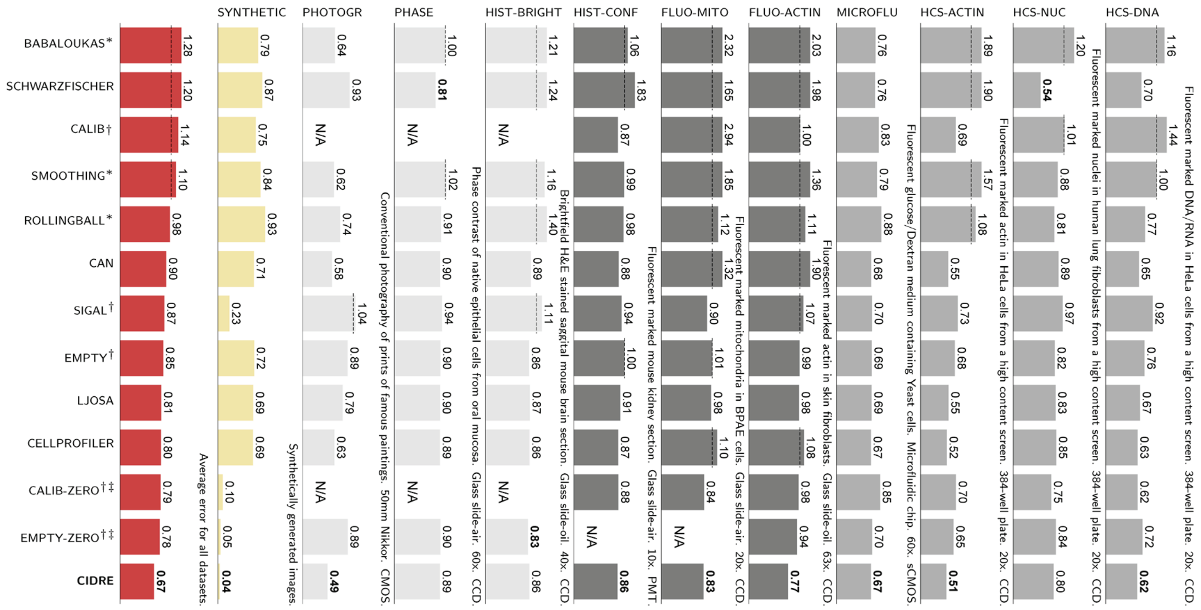


Figure 3. Evaluation of various illumination correction methods (rows) on eleven data sets (columns). Scores are the mean absolute differences between pairs of overlapping test images after correction, computed pixel-wise and normalized by the disagreement between uncorrected image pairs. A score of 1 implies equivalent disagreement to the uncorrected image pairs (indicated by dashed lines), 0 implies no disagreement. Image source: (Smith et al. 2015)

To address these limitations, we proposed a new correction model, CIDRE. CIDRE learns a correction model directly from a collection of images using an energy minimization approach (Fig. 2). We assume that objects appear everywhere in the image with equal probability. This implies that the distribution of intensities collected from an infinite number of uncorrupted images is the same for any location, but also means that time-lapse images or image sets with stationary objects can be problematic. Intensity distributions from a finite set of observed images (blue and red in Fig. 2b) are simply linear transforms of a sampling of that true distribution (gray). CIDRE estimates v, b, and z simultaneously for all locations by minimizing a robust regularized energy function comprising several terms (Fig. 2c). The first term is a robust regression that ensures fitting the data (Fig. 2d). The second term reduces noise and guarantees a smooth correction surface by encouraging neighboring distributions to agree on similar values for v (Fig. 2e). The third term estimates the intensity recorded by the sensor when no light is present. The optimization procedure estimates parameters by minimizing these energies. We refer back to section 2.1, K – the prior knowledge – here is the weighted sum of the three terms mentioned above.

Like other retrospective correction methods, CIDRE is dependent on the amount and quality of data provided. Too few images or images with low intensity information can yield poor results. Experiments with a varying number of training images show that for most applications, 10 images are sufficient for a good performance, while for datasets with low information content, approximately 100 images are necessary.

CIDRE offers numerous advantages over contemporary correction methods. It does not require calibration images, so it can be applied to previously collected data. It is the only computational method to estimate the zero-light term, and simultaneously learns v and z for every location, ensuring that the parameters interact to provide a good correction. The robust regression protects against outliers and utilizes every bit of provided data. CIDRE consistently performed well on every dataset, and outperformed the other twelve correction methods by a substantial margin, including the gold-standard methods (Fig. 3).

2.3 Single cell segmentation

After illumination problems are corrected, identification of the cell’s nucleus is the starting point for many approaches of microscopy-based cellular analyses. The precise location of the nucleus is the basis for a variety of quantitative assessments of essential cellular functions, and is the first step in determining the boundaries of individual cells, allowing for a variety of further analyses. Dominant approaches for this task have been based on classic image processing algorithms (e.g., thresholding and seeded watershed; (Carpenter et al. 2006)), guided by shape and spatial priors (Molnar et al. 2016). These methods require expert knowledge to properly adjust the parameters, which typically must be retuned when experimental conditions change.

Recently, deep learning has revolutionized an assortment of tasks in image analysis, from image classification to face recognition. It is also responsible for breakthroughs in diagnosing retinal images, classifying skin lesions with superhuman performance (Esteva et al. 2017), and correcting artifacts in fluorescence images. A recent work reviewed in (Moen et al. 2019) indicates that deep learning is effective for nucleus segmentation (Falk et al. 2019), however, these methods often fail to properly separate touching nuclei well, and most importantly, they lack robustness to unseen domains (Reka Hollandi et al. 2020). We have recently published a comprehensive review about this topic (Reka Hollandi et al. 2022).

We remark that the complete understanding of the models below requires solid background in differential geometry, variational calculus and statistics. Here we only give a high-level overview, and refer to the relevant publications everywhere.

2.3.1 Segmentation using higher-order active contours

Relevant publication:
Horvath, P., Jermyn, I., Kato, Z., Zerubia, J. (2009)
A higher-order active contour model of a "gas of circles"
Pattern Recognition

As discussed above, the identification of fluorescently stained cell nuclei is the basis for cell recognition, segmentation and feature extraction in high-content microscopy experiments. The nuclear morphology of individual cells is also one of the most important indicators of phenotypic variation. However, the cells used in the experiments can lose their contact inhibition, and therefore stack on top of each other, making the detection of individual cells extremely difficult with current segmentation methods. The model presented can recognize cell nuclei and their morphology, even in highly confluent cell cultures with many overlapping nuclei. We have combined the active contour model "gas of circles", which favors circular shapes but allows small variations around them, with a new data model. We have demonstrated the power of our method on microscopic images of cells, comparing the results with those obtained from a widely used approach, and with manual image segmentations executed by experts.

The term “gas of circles” refers to regions in the image domain composed of an unknown number of circles of approximately the same radius (Horváth et al. 2009). The model is constructed using higher-order active contours (HOACs) in order to include non-trivial prior knowledge about region shape without constraining topology. Our main theoretical contribution is an analysis of the local minima of the HOAC energy that allows us to guarantee stable circles, fix one of the model parameters, and constrain the rest. We originally applied the model to tree crown extraction from aerial images of plantations (Fig 4).

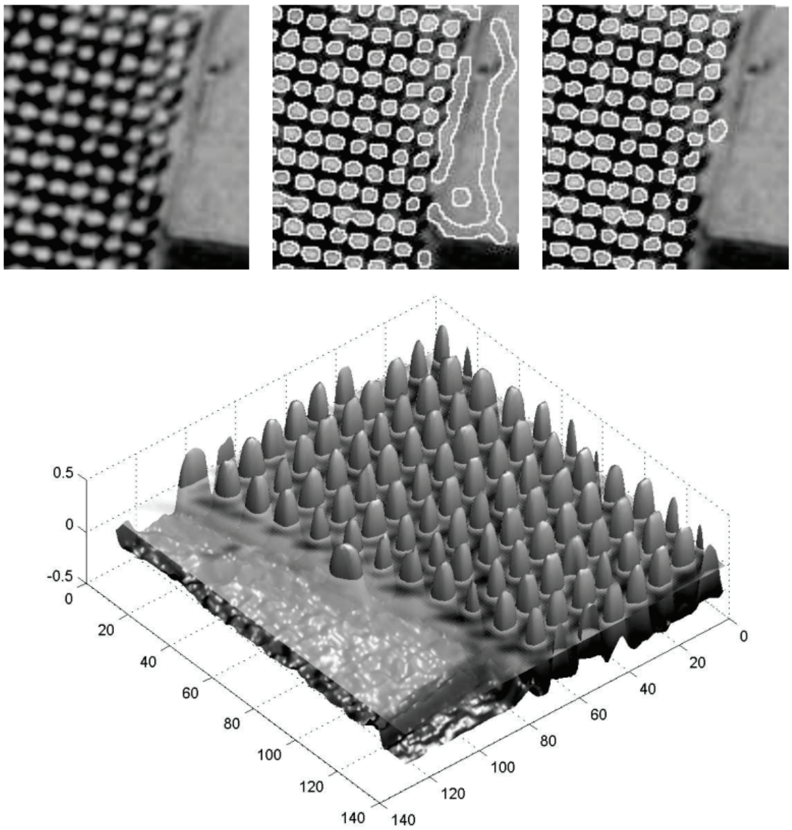


Figure 4. Tree crown extraction from aerial images of plantations. *Upper row; left*: an image of regularly planted poplars with different fields on the right; *middle:* result with the 'gas of circles' model; *right:* result with the inflection point 'gas of circles' model. *Bottom:* The corresponding field surface, thresholded with the original image. Image source: Horvath PhD thesis.

Numerical experiments both confirmed the theoretical analysis and show the empirical importance of the prior shape information (for the stability calculations, see (Horváth et al. 2009)).

We further developed the 'gas of circles' model to enable single cell segmentation in fluorescence images (Molnar et al. 2016). The fundamental idea of the "multi-layered gas of circles" (ML GOC) model is the observation that fluorescently labelled single cells located above each other have multiples of the intensity of a single cell alone (Fig. 5). Therefore we have created multiple layers of the GOC model and an additional energy that allows interaction between the layers, and punishes overlapping circular objects unless fluorescence intensity in the image is high enough to justify multiples of cells (Fig. 5 f).

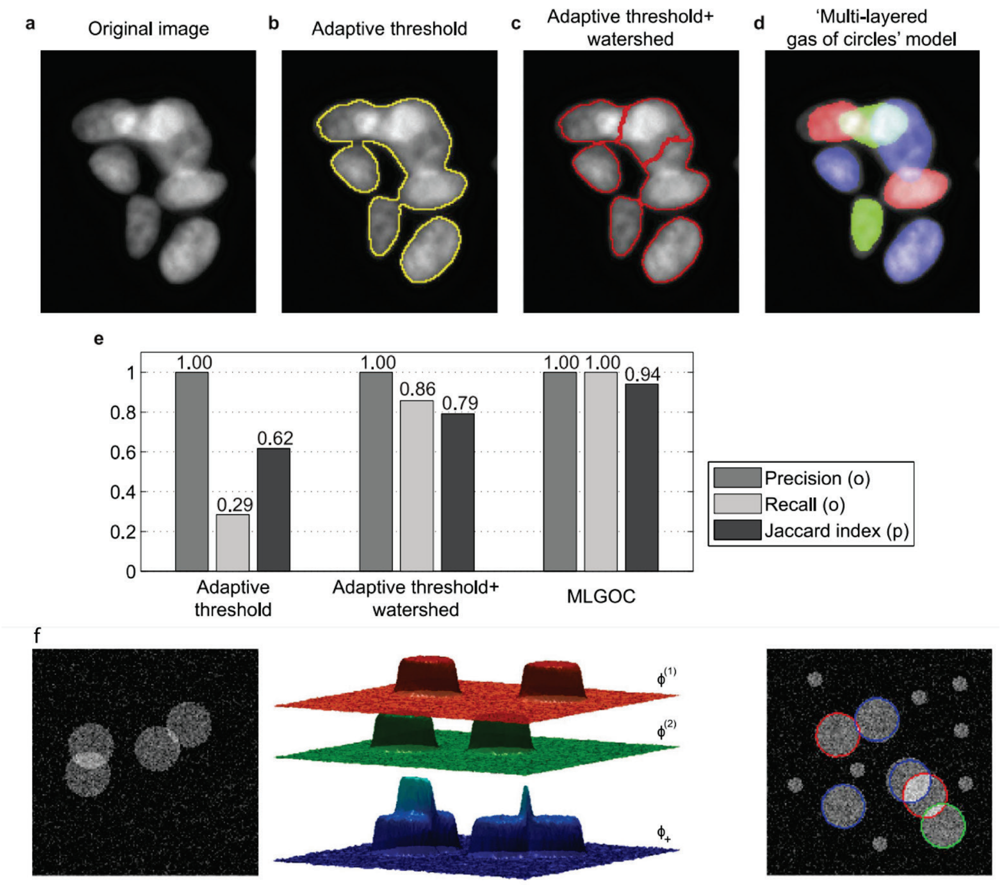


Figure 5. Comparison of different methods on microscopic images containing overlapping cells. *Top row from left to right:* **(a)** Original image; **(b)** Result (Region of Interest) obtained by adaptive threshold using CellProfiler; **(c)** Results of CellProfiler standard segmentation method; **(d)** Results with the proposed "multilayer gas of circles" method; **(e)** Precision, recall and Jaccard index of segmented objects ('o' and 'p' indicate that the metrics are computed at the object and pixel level, respectively). **(f)** Illustration of the proposed data model and behavior of the geometric model. Image source: (Molnar et al. 2016)

2.3.2 NucleAIzer: deep learning-based single cell segmentation

Relevant publication:

Hollandi, R., Szkalisity, A.,..., Carpenter, A. E., Smith K., **Horvath, P.** (2020)
nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer.
Cell Systems

The 2018 Data Science Bowl (DSB) organized by Kaggle, Booz Allen Hamilton and the Broad Institute challenged participants to push the state of the art in nucleus segmentation. The goal of the challenge was to develop fully automated and robust methods effective in a variety of conditions, including differing cell lines, treatments, and types of light microscopy. The challenge attracted thousands of data scientists from all around the world. Approaches using deep learning dominated the competition, achieving scores that shattered what was previously possible: the best performing traditional methods, such as those discussed in the previous section, ranked no higher than 1,000 out of 3,891 submissions. The top deep-learning-based methods relied on only a handful of different architectures, namely Mask R-CNN, U-Net, and feature-pyramid networks; the factors that the participants commonly believed to have most influence over their method's ranking were the amount of data, pre-processing and the methods used to augment the data.

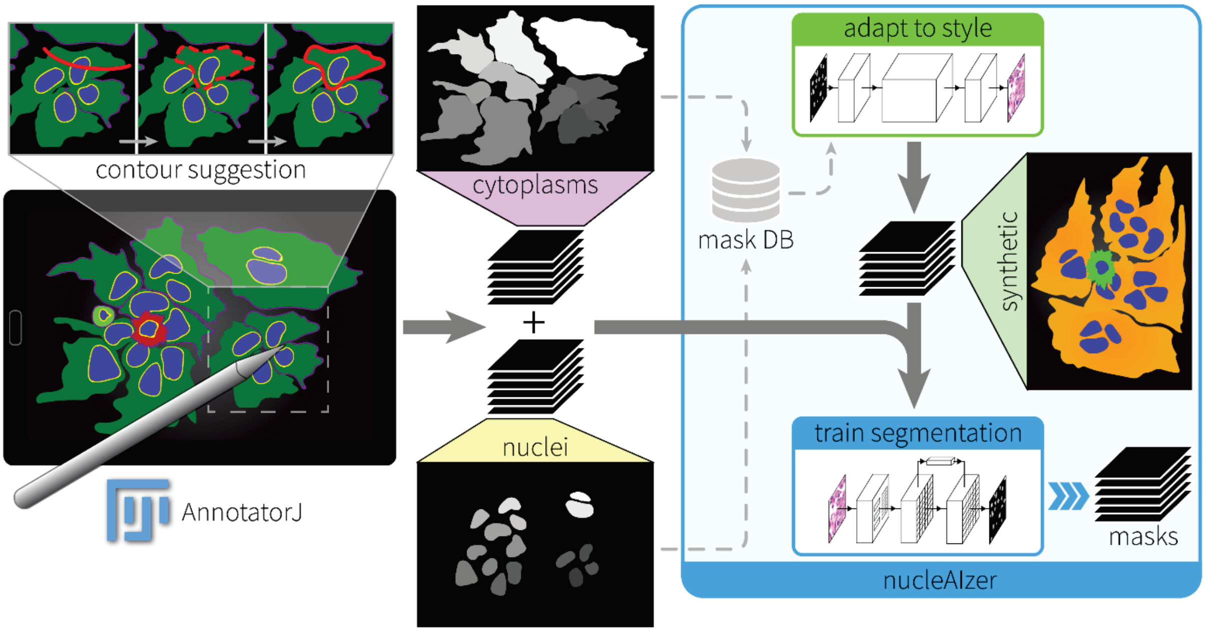


Figure 6. A single cell segmentation pipeline using nucleAIzer. Objects annotated (e.g. in this case in AnnotatorJ (Réka Hollandi et al. 2020)) can be exported to proper training data format suitable to train different types of deep learning models. The annotated object masks are stored in a mask database (denoted as 'mask DB' in the figure) which is used in the mask generation and image style transfer step of the nucleAIzer pipeline (green box in the figure). The resulting synthetic microscopy images adapted to the new experiment's style (appearance) are generated and forwarded to train a segmentation model, together with the annotated masks, and the trained model yields segmentation masks as output. Image source: Reka Hollandi's PhD thesis, with the permission of the Author.

We presented an approach superior to any other submissions, which we named nucleAIzer (www.nucleaizer.org). Unlike the previous best submissions, nucleAIzer applies image style transfer (Isola et al. 2017), i.e. an image-to-image translation using a pixel-wise mapping from one image to the other, which ensures that the generated synthetic output image resembles the original image as closely as possible. It aims to overcome one of the greatest challenges of deep learning: the extent of the annotated training set. In particular, we have addressed the unsupervised domain adaptation problem in which the target (test) samples are drawn from a different distribution than the labelled training samples, but we have access to some unlabelled samples from the target distribution. We have augmented the training samples by creating realistic-looking artificial sample images with the texture, coloration and pattern elements from source images not included in the training set, using image style transfer (Fig 6). Combining this with a segmentation network based on Mask R-CNN (He et al. 2017), an instance segmentation and classification network, along with boundary correction using U-Net (Ronneberger, Fischer, and Brox 2015), a semantic segmentation network for biomedical images and mathematical morphology, our method has outperformed all other methods reported on the final DSB leaderboard. We have also demonstrated that our method outperforms similar baselines on public fluorescent and histology datasets. Our trained model does not require parameter tuning or specialized knowledge for use, and can be applied on a wide variety of conditions and imaging modalities. Some example segmentations are shown in Figure 7 and the Reader can try the online portal under www.nucleaizer.org.

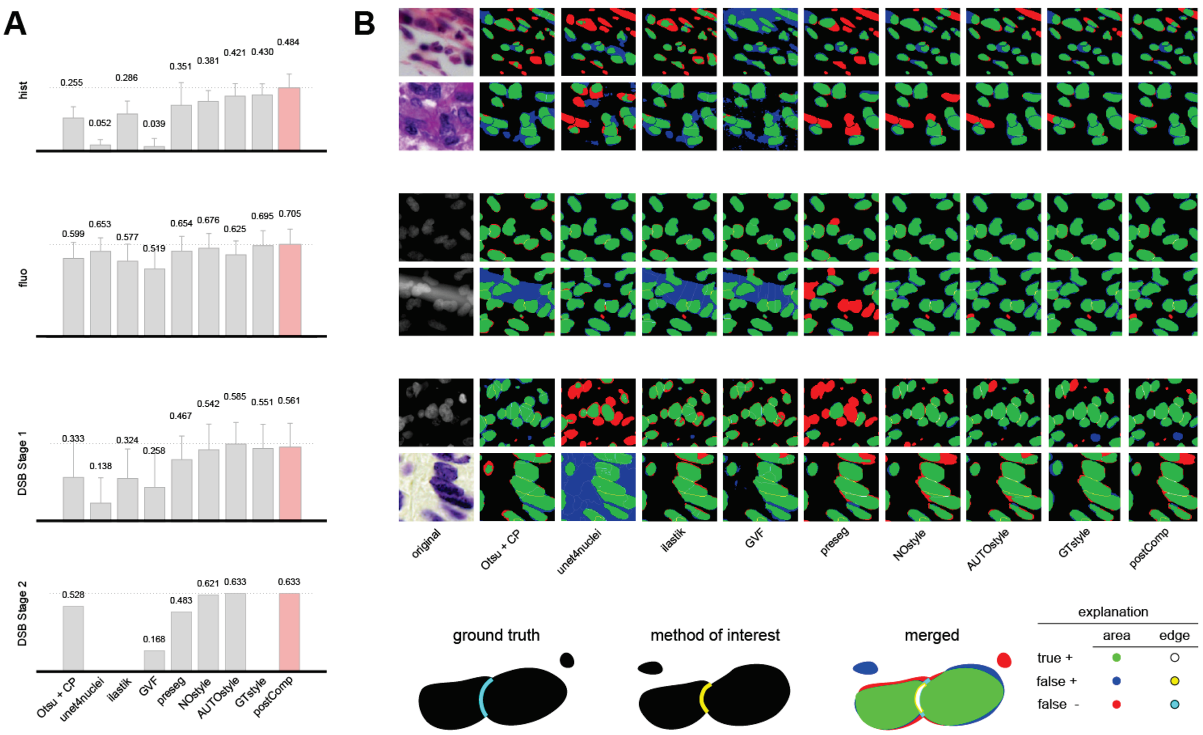


Figure 7. Quantitative and qualitative comparison of segmentation results. (A) Mean IoU scores with error bars (standard deviation) on the four test image sets. The highest score is marked with a dashed line and in pink colour. **(B)** Example colour-coded segmentation results compared to ground truth annotations on difficult image regions cropped. Two examples are shown per test set, and rows correspond to A. See colour coding explained in the legend in the bottom row. We remark that ground truth annotations were not available for DSB stage 2. Image source: (Reka Hollandi et al. 2020)

As shown above, augmentation techniques can largely improve segmentation accuracy, as deep learning approaches for object segmentation require a large, and often pixel-wise annotated dataset for training. This task relies on high-quality samples, and only domain experts can accurately annotate images. Besides, analyzing biological images is challenging because of their heterogeneity and, sometimes, poorer quality compared to natural images. In addition, ground truth masks might be imperfect due to the annotator-related bias, which introduces further uncertainty. Consequently, a plethora of annotated samples is required, making object segmentation a laborious process. One of the techniques utilized to improve the model is data augmentation of the training set. Conventionally, a transformation (i.e. rotation, flipping, noise addition, etc.) or a series of transformations are applied on the original images. Data augmentation has become the *de facto* technique in deep learning, especially in the case of heterogeneous or small datasets, to improve the accuracy of cell-based analysis.

Another option of improving performance relies on augmenting both the training and the test datasets, then performing the prediction both on the original and on the augmented versions of the image, followed by merging the predictions. This approach is called *test-time augmentation* (TTA) (Fig. 8). Experiments show that TTA helps to eliminate overconfident incorrect predictions.

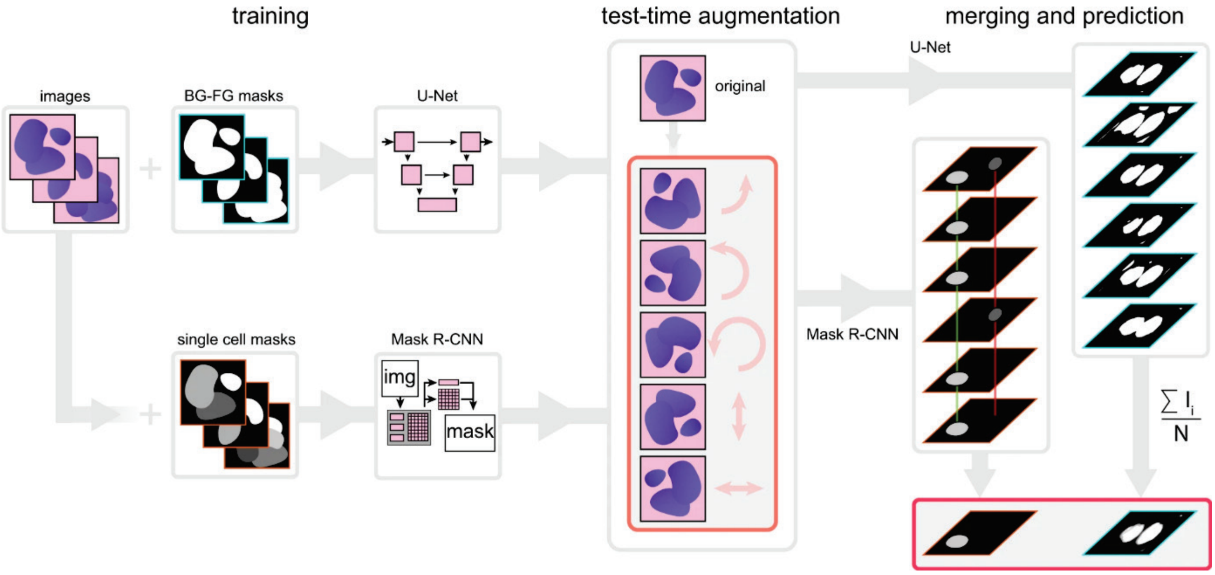


Figure 8. Principle of the proposed test-time augmentation techniques. Several augmented instances of the same test images are predicted, and the results are transformed back and merged. In the case of U-Net, pixel-wise majority voting was applied, while for Mask R-CNN a combination of object matching and majority voting was applied. Image source: (Moshkov et al. 2021)

In (Moshkov et al. 2021) we assessed the impact and described cases of utilizing test-time augmentation for deep-learning models trained on microscopy datasets. We have trained deep learning models for semantic segmentation (when the network only distinguishes the foreground from the background, using the U-Net architecture) and instance segmentation (when the network assigns labels to separate objects, using the Mask R-CNN architecture) (Fig. 8). Test-time augmentation has outperformed single instance predictions in each test case, and could further improve the best result of the DSB, as demonstrated by the improvement of the score, changing from 0.633 to 0.644.

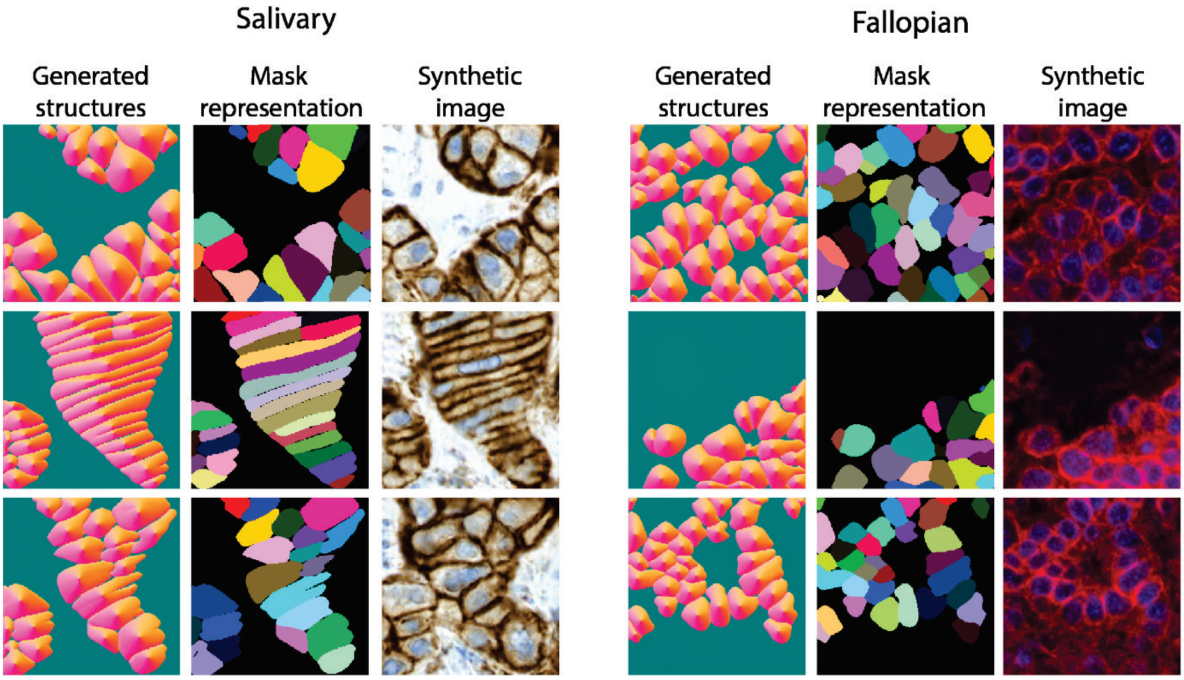


Figure 9. Synthesized flows, the reconstructed masks and their corresponding microscopy images generated by our method using two complex tissue culture types.

Finally, we have recently been working on smart augmentation methods that not only mimic the texture of the cells as presented in (Reka Hollandi et al. 2020), but learns the structure and the appearance of cell cultures and tissues simultaneously, by using a complex parallel deep learning system (Tasnadi *et al,* submitted). Our pipeline uses regular and conditional generative adversarial networks (GANs) for image-to-image translation to construct synthetic microscopy images along with their corresponding masks, in order to simulate the distribution and shape of the objects and their appearance. The major benefit of our method is that it not only proposes synthetic images, but also the corresponding labelled mask images. The synthetic samples are then used to pre-train instance segmentation networks (such as the nucleAIzer). The proposed method further increases accuracy when combined with other augmentation techniques, such as basic transformations or nonlinear geometric augmentations like elastic deformation.

CHAPTER 3

Phenotypic analysis

Statistical learning-based, single cell level decisions comprise one of the great advancements of bioimage analysis. This is especially true with the rise of automation both in sample handling and in image acquisition. The level of complexity concerning biology-related questions is very high. Personally, what I love the most about developing and applying the introduced tools is the feeling of discovery: being a pioneer in finding new genes or drugs to modify a disease or solve fundamental biological questions derived from large imaging scenarios.

Phenotypic image analysis is the task of recognizing variations in cell properties using microscopic image data. These variations, produced by a complex network of interactions between genes and the environment, may hold the key to uncover important biological phenomena or to understand the response to a medical treatment. Today, phenotypic analysis is rarely performed completely by hand. The abundance of high-dimensional image data produced by modern high-throughput microscopes necessitates computational solutions. Over the past decade, a number of software tools have been developed to address this need. They use statistical learning methods to infer relationships between a cell's phenotype and data from the image.

3.1 Single cell phenotyping

Relevant publication:
Piccinini, F., Balassa, T., Szkalisity, A., …, Smith K., **Horvath, P.** (2017)
ACC: Discovery software for phenotypic image analysis.
Cell Systems

We have introduced Advanced Cell Classifier (ACC; www.cellclassifer.org) (Piccinini et al. 2017), a machine-learning software designed to provide a faster and more complete understanding of large datasets, and to train predictive models as accurately as possible. In

2011, we released ACC version 1.0 (ACC v1.0), a graphical image analysis software tool that offers access to a variety of machine-learning methods and provides accurate analysis (Horvath et al. 2011). Several large-scale cell-based phenotypic HCS studies have made use of ACC, including at least 25 human genome-wide RNAi screens and numerous extensive drug screens. These studies cover a wide variety of biological topics, ranging from the studies of influenza A virus (Banerjee et al. 2014) to studies of acute lymphoblastic leukemia (Fischer et al. 2015).

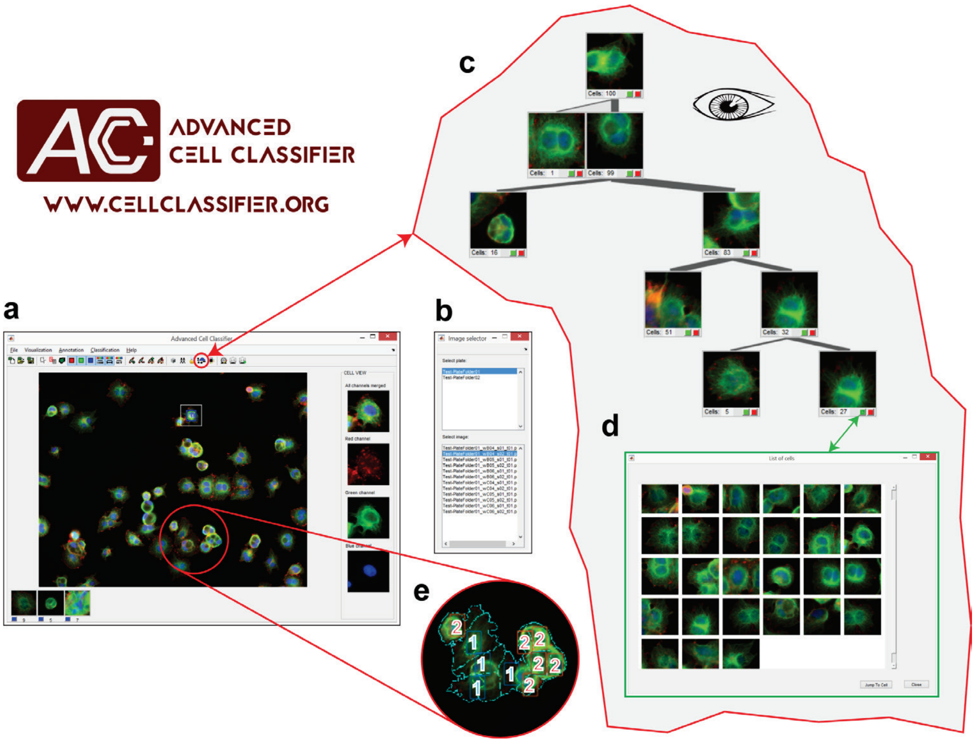


Figure 10. The Advanced Cell Classifier interface. (a) The main ACC interface is intuitive and easy to use for non-experts. **(b)** The image selector window allows the user to select different images, easily navigating between plates and images. **(c)** Phenotype finder, a tool that uses machine learning to efficiently discover new cellular phenotypes. It organizes non-annotated cells into a browsable hierarchical tree, based on their appearance. **(d)** Cells automatically discovered by phenotype finder as new phenotypes. **(e)** Once the classes of interest are defined and a classifier is trained, phenotypes of non-annotated cells can be predicted with a single click. Image source: (Piccinini et al. 2017).

ACC is a user-friendly software tool with the goal of improving the collection and understanding of image data and the accuracy of the analysis (Fig. 10). It allows researchers, even those without computer vision or machine-learning knowledge, to efficiently characterize and exploit their cell-based and image-based HCS experiments, leading to new discoveries. It is capable of answering the questions raised in the introduction (ie. 'Have I entirely discovered my data? Or at least partially?' and 'Is my analysis as accurate as possible?') by using intelligent methods to explore and annotate large single-cell image data, including (1) an active learning approach to improve the accuracy of the classifier, and (2) similar cell search, an algorithm to find similar cells and increase the number of annotations for rare phenotypes (Fig. 11); moreover, (3) phenotype finder, a novel method to automatically discover new and biologically relevant cell phenotypes is included (Fig. 12). In addition, it provides an easy-to-use report generator to automatically obtain statistics on cell distribution and class incidence; a user-friendly interface; detailed documentation, video tutorials, and online resources; as well as improved data visualization methods. The source code of ACC is freely distributed as an open-source tool at www.cellclassifier.org.

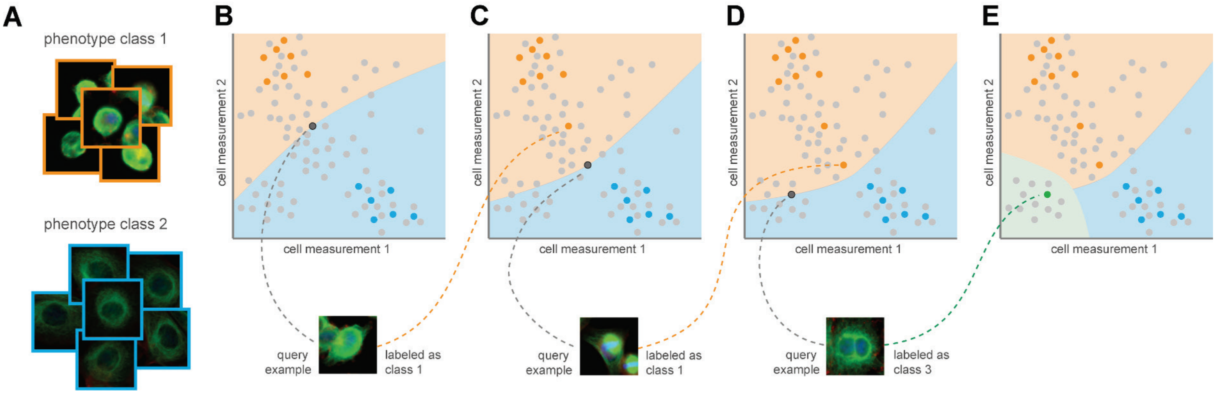


Figure 11. Active learning tool of ACC. The active learning annotation tool helps to refine the decision boundary of the classifier. It prioritizes the most informative examples and presents them to the user for annotation. **(A)** Known phenotypic classes (blue and orange) with cells initially annotated to these classes. **(B)** Two-dimensional illustration of a synthetic feature. Bold coloured points represent annotated examples, grey points represent unannotated examples. The orange and blue regions define the classifier's predictions of the two phenotypes. The active learning tool uses a query strategy to determine which cell would be informative to the classifier (black point). **(C)** When a cell is annotated, the decision boundaries change and a new cell is chosen. **(D)** In some cases, the classifier may have low certainty for cells that do not belong to any existing phenotypes. The user may decide to create a new class, and **(E)** the decision boundaries of the multi-dimensional feature space change accordingly. Image source: (Piccinini et al. 2017) .

ACC provides several new innovative tools designed to explore and collect the data necessary to train classifiers more efficiently and effectively. The ability of the classifier to correctly recognize cell types ultimately depends on the quality of data provided. While there is no universally accepted recipe for generating quality training data, many principles and techniques can be applied in practice to improve the efficiency and quality of the annotation process. The most fundamental principle is to ensure that the training data are complete in the sense that they include examples of all the important phenotypes present in the screen. Although it may seem obvious, practically speaking, this can be tedious when the amount of data is very large. Another common issue is imbalance between classes. Often, interesting phenotypes are in the minority or occur very infrequently. If the data are imbalanced due to the presence of a rare class, the lack of representative data will make learning difficult. Given a single example of a rare cell, ACC can quickly identify additional, previously unidentified examples by using the similar cell search feature, thereby helping to balance the dataset and improve classification performance. Another way to improve data collection is to avoid redundant annotations and to prioritize annotations that are most useful for boosting classification performance. ACC uses active learning to carefully select the most informative examples for labelling, and thereby it avoids irrelevant examples and refines regions where the classifier is uncertain.

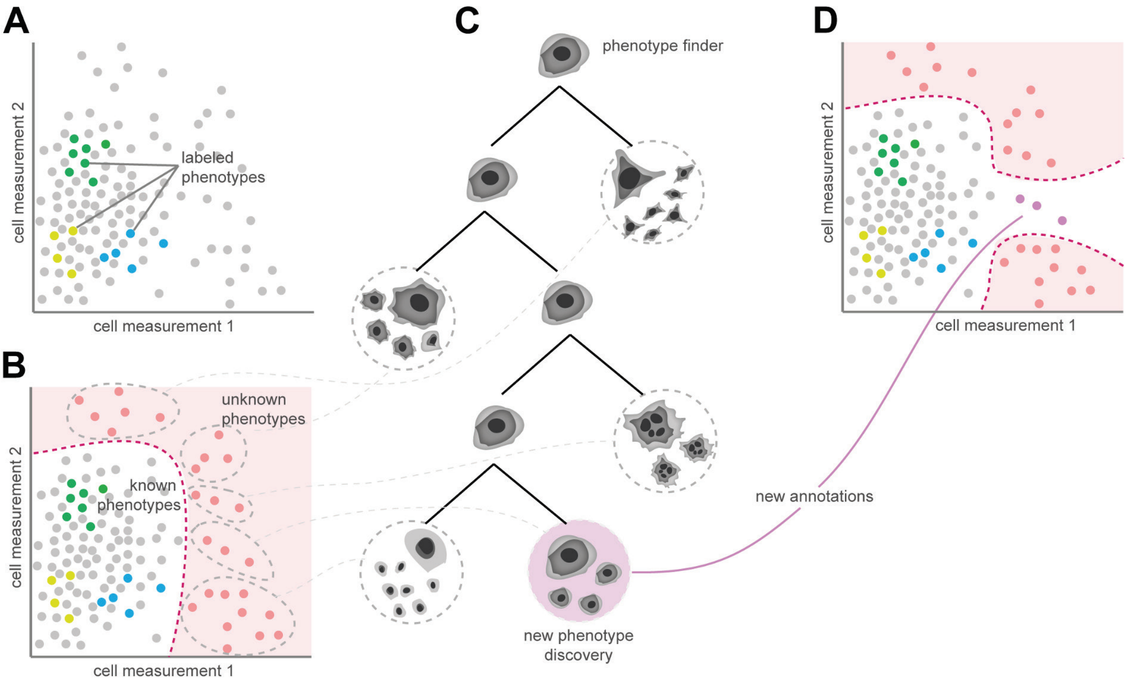


Figure 12. The Phenotype Finder Tool off ACC. The phenotype finder tool organizes cells into a browsable hierarchy. **(A)** Cells are represented by dots embedded in a two-dimensional synthetic feature space. Sets of cells annotated by the expert are shown in green, yellow and blue. **(B)** A one-class classifier is used to automatically determine which cells are the least similar to the known cell types (pink region). **(C)** Cells with the least similarity to known examples are sampled and clustered to construct a dendrogram. The expert can browse and analyze the representative cells shown in the tree, create a new phenotype class from these cells, or add cells to an existing phenotype. **(D)** If a new phenotype is discovered, the region of non-annotated cells changes in the multidimensional feature space. Image source: (Piccinini et al. 2017).

Data quality can also be improved by the iterative refinement of the classifier. During data collection, ACC can train a classifier on existing annotated data, and can display predicted annotations on unlabelled data. By correcting erroneous predictions, the user adds valuable data points to the training set, which can help to correct predictive errors.

In summary, ACC includes powerful new methods to mine microscopic image data, discover new phenotypes, and improve recognition performance. While no single method can be regarded as a silver bullet that solves all annotation problems, in our experience, the most effective strategy is to alternate between discovery tools, adjusting it to the demands of the actual biological task. ACC gives the user access to a large variety of state-of-the-art machine-learning algorithms, has an intuitive user interface with advanced visualization, and allows for efficient navigation of image data. It is easy to use, well documented, and comes with helpful video tutorials. By using synthetic data and existing screens, we have demonstrated that the discovery tools in ACC improve the quality of training datasets and ultimately create classifiers with better phenotype recognition. Using our software, it is possible to discover interesting cell phenotypes hidden in large datasets.

3.2 The regression plane concept

Relevant publication:
Szkalisity, A., Piccinini, F., Beleon, A., …,Honti, V., **Horvath, P.** (2021)
Regression plane concept for analysing continuous cellular processes with machine learning
Nature communications

Biological processes are inherently continuous, and the chance of phenotypic discovery is significantly restricted by discretizing them. Using multi-parametric active regression we have introduced the Regression Plane (RP), a user-friendly discovery tool enabling class-free phenotypic supervised machine learning, to describe and explore biological data in a continuous manner (Szkalisity et al. 2021). Regression Plane (RP) is an interface for fully supervised, continuous machine learning, appropriate for image-based single-cell analysis. The idea originates from a study of cell entry of influenza A virus, revealing that histone deacetylase-mediated reorganization of the microtubules led to various endosomal morphological and trafficking phenotypes that affected influenza infection (Yamauchi et al. 2011). The scatteredness of late endosomes and lysosomes (single output variable) was determined using regression instead of classification. Restricting the output to a single dimension prohibited the modelling of branching, circulating, parallel and crossing processes, therefore we extended the approach to utilize a 2D plane (Fig. 13). Considering cellular steady states as graph nodes and gradual changes between the states as edges, the biological systems that correspond to planar graphs can be modelled with RP. Further extension of the modelling to 3D would increase the complexity of labelling and raise the chance of annotation errors. Additionally, to improve the quality of the annotated sets and decrease the time required from experts, we have incorporated active learning methods appropriate for regression-based phenotyping.

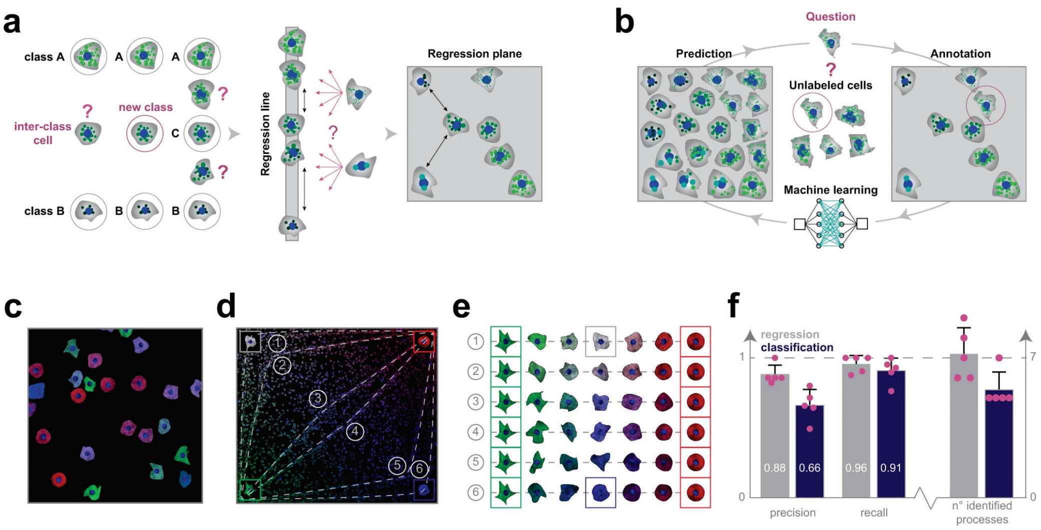


Figure 13. The regression plane concept. (a) The classical way to model a biological process includes the phenotypical analysis of cells (i.e. subdividing cells into classes). However, in a high-content screening scenario, the multitude of different phenotypes makes it extremely challenging to create a set of representative classes. A possible solution builds on using a regression line, allowing to represent a single effect without the need of discretization. Nonetheless, biological processes are typically characterized by numerous ongoing effects. Thus, the regression plane represents a good trade-off between visualization capabilities and annotation complexity. **(b)** Active regression. **(c)** Synthetic dataset. Image from the synthetic dataset, generated using SIMCEP. **(d)** Experimental design. The designed processes overlayed on the space of perturbations. 6 processes are tracks in the space, and an extra process is formed of uniformly distributed cells (latent process 7). **(e)** Designed processes. The 6 continuous processes are modelled between two fixed endpoints: green cells of highly irregular shape and red, rounded cells. To assign a colour to the middle point of each process we interpolated between white (process 1) and blue (process 6). **(f)** Classification *vs.* regression applied on synthetic data. Image source: (Szkalisity et al. 2021).

We tested the capabilities of RP on 2 different time-resolved datasets. First, RP was demonstrated to be capable of reproducing an unsupervised mitotic time model developed in the MitoCheck project. (Cai et al. 2018) analyzed cell mitosis by performing time-lapse experiments to establish a canonical model for the morphological changes appearing during the mitotic progression of human cells. We intended to analyze this dataset without using prior feature information about the underlying process, by exploiting regression techniques to characterize mitosis.

In our analysis, a field expert created a regression plane representing the process of mitosis (Fig. 14a). After prediction, the cells followed the designed circular path recalling canonical mitotic phases (Fig. 14b-d), while they also represented subtle phenotypic changes and single-cell differences in the regression plane. Additionally, we investigated whether the fluorescent tags have an effect on the distribution of cells on the regression plane, and in most cases we did not observe undesired cellular behaviour due to the perturbations. Finally, we compared the results of the original methodology presented by Cai *et al,* (multi-dimensional dynamic time warping [DTW] for creating the standard mitotic time, Fig. 14e) with the results obtained by RP

(Fig. 14f), and we concluded that RP is capable of reproducing a mitotic time model equivalent to the original one. This indicates that RP can compete with complex analysis techniques, such as DTW. Moreover, RP provides the flexibility to customize the output space, enabling higher resolution analysis of user-defined sections of the biological process.

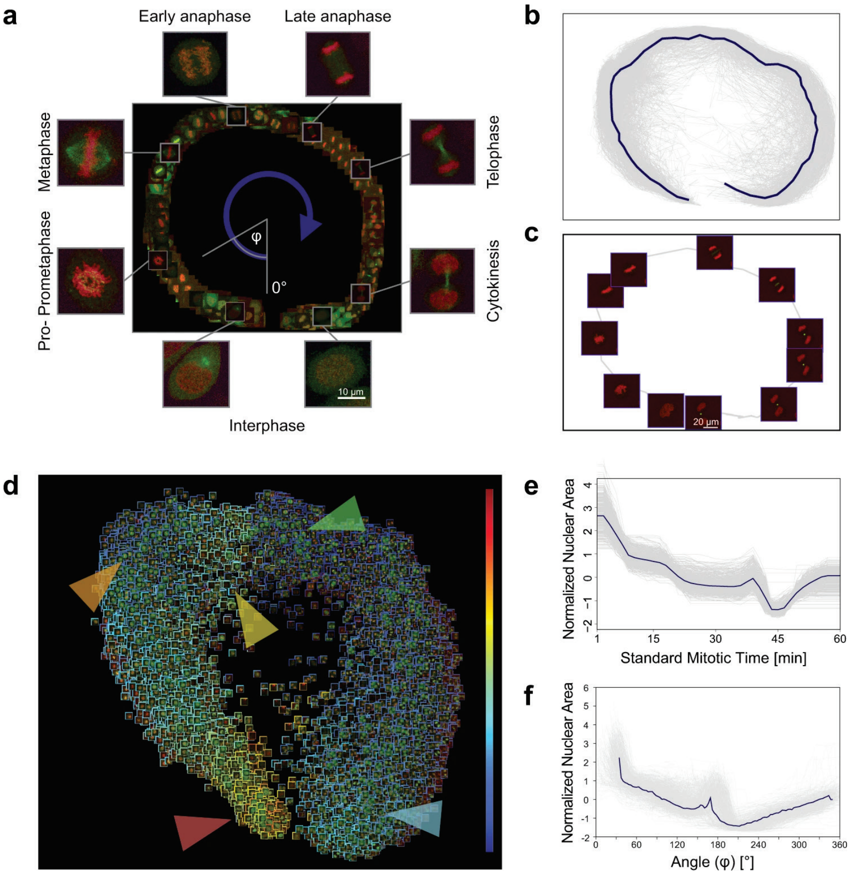


Figure 14. Testing the regression plane method on mitosis. (a) Regression plane of 585 cells annotated by a microscopy expert. **(b)** Trajectories for all the predicted cells. The median curve is shown in solid blue. **(c)** Example of a single-cell trajectory with representative cell icons visualized. **(d)** Regression plane with all (n = 19,920) predicted cells. **(e)** Trend for the normalized nuclear area according to standard mitotic time. Grey lines represent single-cell trajectories. **(f)** Trend for the normalized nuclear area according to the regression plane. Grey lines represent single-cell trajectories. The coordinates predicted by RP were converted to 1D by taking the angle argument of the polar coordinate representation as illustrated in (a). Image source: (Szkalisity et al. 2021)

3.3 A case study: the discovery of Neuropilin-1, a host factor for SARS-CoV-2

Relevant publication:

Daly, J.L., Simonetti, B., ...Horvath, P., ..., Yamauchi, Y. (2020)

Neuropilin-1 is a host factor for SARS-CoV-2 infection

Science

The methods presented above were used in several studies, leading to a wide range of fundamental biological and clinical discoveries, such as the cell entry factors of influenza A virus (Banerjee et al. 2014), the description of NUP98, a large multiprotein component of the nuclear pore complex (NPC) which regulates the transport of macromolecules between the nucleus and cytoplasm (Laurell et al. 2011), or personalized treatment for childhood leukemia (Frismantas et al. 2017). Here I will present an application which is very timely, namely the discovery of neuropilin-1, the second known host factor of SARS-CoV-2 coronavirus (Daly et al. 2020). Based on our previous studies performing whole-genome screens on flu, we already knew that neuropilin-1 served an entry factor in viruses. Thus, in early 2020, soon after the COVID-19 pandemic was declared by the WHO, we started SARS-CoV-2 screening to identify the main determinants of its infectivity.

It is well known that virus-host interactions determine cellular entry and spreading in tissues. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the earlier SARS-CoV use angiotensin-converting enzyme 2 (ACE2) as a receptor; however, their tissue tropism differs, raising the possibility that additional host factors are involved. The spike protein of SARS-CoV-2 contains a cleavage site for the protease furin which is absent from SARS-CoV. Neuropilin-1 (NRP1), which is known to bind furin-cleaved substrates, potentiates the infectivity of SARS-CoV-2. NRP1 is abundantly expressed in the respiratory and olfactory epithelium, with highest the expression levels in endothelial and epithelial cells. We found that the furin-cleaved S1 fragment of the spike protein binds directly to cell surface NRP1, and blocking this interaction with a small-molecule inhibitor or a monoclonal antibody reduced viral infection in cell culture. The significant role of NRP1 in SARS-CoV-2 infection may suggest potential targets for future antiviral therapeutics.

Here I shortly describe the image analysis and machine learning-based single cell phenotyping steps of the discovery process. HeLa cells were imaged with a confocal laser scanning microscope (SP5II AOBS, Leica Microsystems) attached to an inverted epifluorescence microscope (DMI600, Thermo Fischer Scientific) with a 40X/1.25na objective. Projected images taken with a 20x objective were used for image analysis for single-cell and multinucleated cell infection image analysis with supervised machine learning. First, images of each fluorescence channel were corrected using the CIDRE illumination correction method (Smith et al. 2015). Individual cell nuclei were detected by the deep machine learning-based segmentation algorithm NucleAIzer (Reka Hollandi et al. 2020). Cellular cytoplasm was detected both on the green and red channels using UNET to enhance fluorescence images. The method was trained to precisely delineate the signals often being faint in the cytoplasm. Cellular phenotypes were assigned to each individual nucleus, differentiating between infected cells containing a single nucleus (single cell infection) and those that contain multiple nuclei (multi-nuclei infection) as observed in the distinct cell-cell fusion syncytia phenotype.

Supervised machine learning was used for phenotypic assignment. The decisions were based on single-cells and the cellular microenvironment's morphology and intensity features (Piccinini et al. 2017). Final statistics included the number of multi-nucleated cells, the average number of 190 nuclei in these cells and the count of other phenotypic classes (Fig. 15).

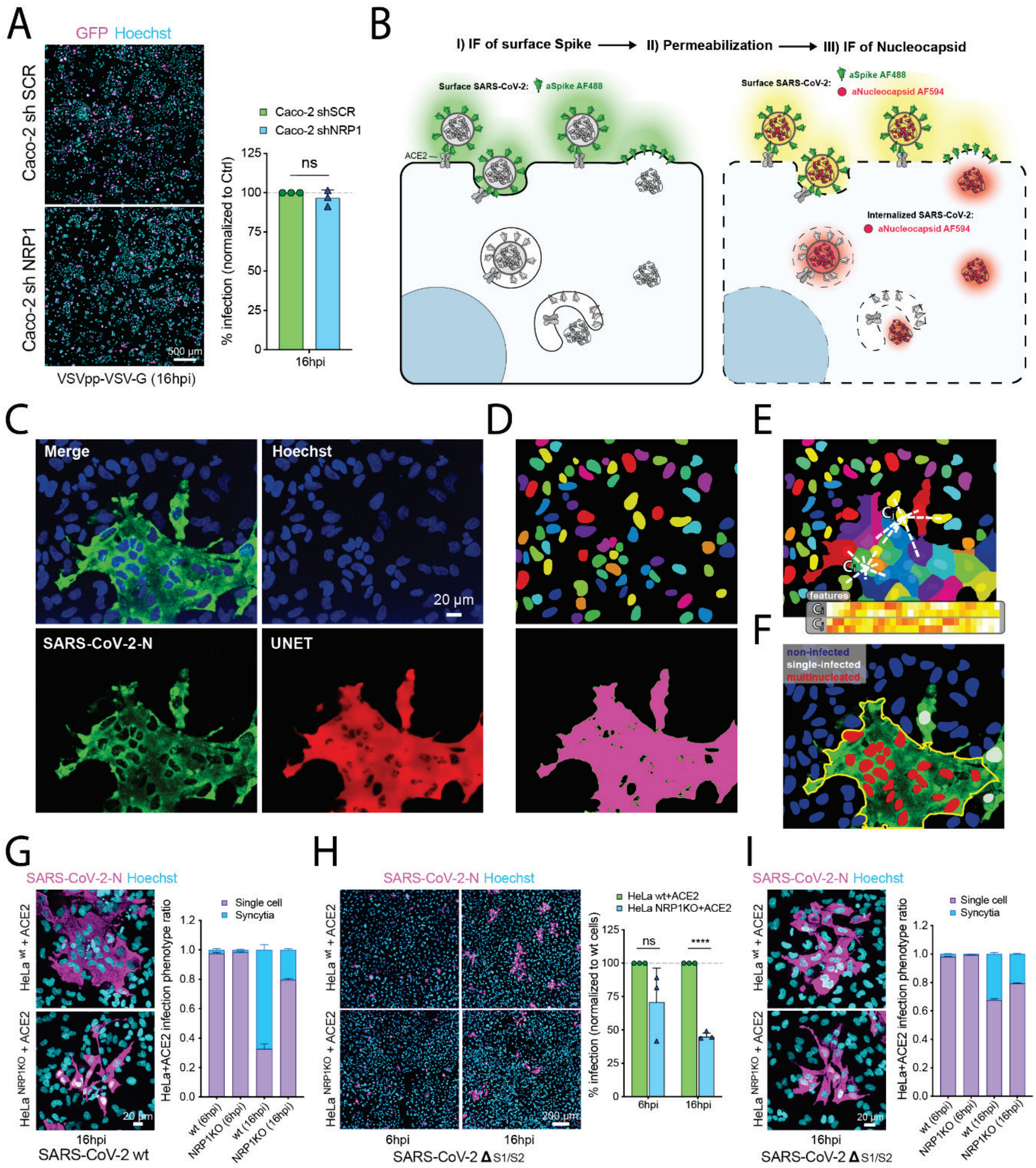


Figure 15. Image processing and phenotyping of SARS-CoV-2. (A) Caco-2 shSCR and shNRP1 cells were infected with VSV pseudotyped with VSV-G for 16 hours. **(B)** Schematic of the two-step staining procedure used to distinguish external and internal virus particles. **(C)** Original image of SARS-CoV-2 N signal (green) and enhanced image (red) using UNET deep learning algorithm. **(D)** Single-cell segmentation of the nuclei using the nucleAIzer deep learning algorithm, and the cytoplasmic region based on global thresholding of the UNET enhanced image. **(E)** Morphology, shape and intensity features of single-cells and their microenvironment are extracted. Features

include morphology, intensity and texture descriptor numbers. Ci: features of the i-th cell, Cj: features of the j-th cell. **(F)** Machine learning-based phenotyping of single cells, differentiating between non-infected, single-nuclei infected and multinucleated cells. **(G)** Ratio of syncytia and single cell infection phenotypes. **(H)** HeLawt+ACE2 and HeLaNRP1 KO+ACE2 cells were infected with SARS-CoV-2 ΔS1/S2. Cells were fixed at 6 or 16 hpi and stained as in (G), and viral infectivity was quantified (N=3). **(I)** Ratio of syncytia and single cell infection phenotypes. Image source: (Daly et al. 2020)

CHAPTER 4

Computer-aided correlative microscopy

> Finally, let me incorporate my other hobby, electromechanics into my thesis. So far, I have introduced you, my dear Reader, to image correction, and how to find cells, and how to determine their phenotypes. Right now I would like to show that adding a little bit to this would further improve fundamental biology research and translational personalized medicine, opening the way for a better understanding of major issues at a single cell level. This bit was the physical extraction of single cells from their native environment. To achieve that, we built intelligent machines controlled by computers. Some use little needles and 3D navigation in the human brain, others use laser beams to isolate a single cancer cell, but all these approaches share a common feature: cell extraction is executed at micrometer precision.

Quantifying the heterogeneity of cell populations is important in many fields, such as cancer research and neurobiology, but techniques for isolating individual cells are limited. We have developed novel automated, high-throughput, non-destructive and cost-effective isolation methods that can capture individual target cells using commonly available techniques. High-resolution microscopy, image analysis, machine learning, patch clamping, and laser microdissection microscopy enable scalable molecular genetic, proteomic and electrophysiological analysis of single cells focused on morphology and 2D and 3D location within the sample.

Specifically, we have developed a technology to increase the accuracy and throughput of microscopy-based single-cell isolation by automating the processes of target selection and isolation (Brasko et al. 2018). Computer-assisted microscopy isolation (CAMI) combines image analysis algorithms, machine-learning and high-throughput microscopy to recognize individual cells in cell suspensions or tissues, and automatically guides extraction by laser capture microdissection (LCM) or micromanipulation. To demonstrate the capabilities of our approach, we conducted experiments that require targeted single-cell isolation to collect individual cells without disturbing their microenvironment. We showed that CAMI-selected cells can be successfully used for digital PCR (dPCR) and next-generation sequencing.

Based on our promising results, we further developed CAMI and introduced the Deep Visual Proteomics concept, which combines intelligent laser micro capturing with ultrasensitive proteomics (Mund et al. 2022). 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. Deep Visual Proteomics (DVP) 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. We confirmed DVP's practical usefulness on a variety of single cell experiments both on cell cultures and tissue sections. The ability of DVP to retain precise spatial proteomics information in the cellular context has implications for the molecular profiling of clinical samples.

I can proudly remark that the CAMI system and its driving software, BIAS, was installed in numerous world-leading labs and pharma companies worldwide.

Finally, as the "queen of challenges" in spatial single cell isolation, we have successfully introduced a deep learning driven needle into a slice of human brain tissue to measure a single cell's electrophysiology response, and have successfully extracted single cells for genetic profiling (Koos et al. 2021). Patch clamp recording of neurons is a labor-intensive and time-consuming procedure. We have presented AutoPatcher, a tool that performs electrophysiological recordings in label-free tissue slices, fully automatically. The automation process covers the detection of cells in label-free images, calibration of the micropipette's movement, approaching the cell with the pipette, formation of whole-cell configuration, and recording. Cell detection is based on deep learning. The model was trained on a new image database of neurons within unlabelled brain tissue slices. The pipette tip detection and approaching phases use image analysis techniques for precise movements. High-quality measurements were performed on hundreds of human and rodent neurons. We have also demonstrated that further molecular and anatomical analyses can be performed on the recorded cells. Our tool can multiply the number of daily measurements to support brain research.

4.1 Computer-assisted, microscopy-based single-cell isolation

Relevant publication:

Brasko C., Smith K., Molnar, Cs., …, Tamas, G., **Horvath, P.** (2018)

Intelligent image-based *in situ* single-cell isolation.

Nature Communications

On a personal note, CAMI is the system that I am the most proud of related to my scientific career. We have developed CAMI to increase the accuracy and throughput of microscopy-based single-cell isolation by automating the processes of target selection and isolation. It combines image analysis algorithms, machine-learning and high-throughput microscopy to recognize individual cells in cell suspensions or tissues, and automatically guides extraction by laser capture microdissection (LCM) or micromanipulation. To demonstrate the capabilities of our approach, we conducted a set of experiments that require targeted single-cell isolation to collect individual cells without disturbing their microenvironment. We showed that CAMI-selected cells can be successfully used for digital PCR (dPCR) and next-generation sequencing (Fig. 16).

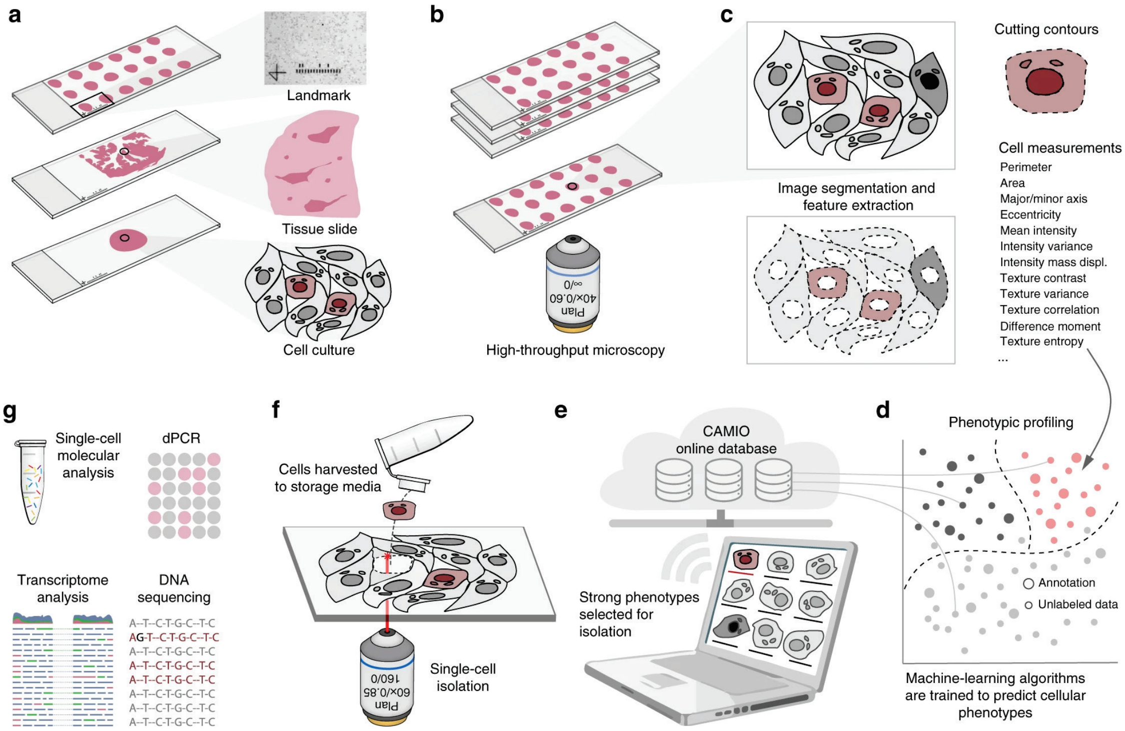


Figure 16. Summary of the computer-assisted microscopy isolation (CAMI) technology. (a) Tissue or cultured samples are prepared. **(b)** Samples are imaged with an automated high-throughput microscope. **(c)** Image analysis software identifies and segments cells. **(d)** Advanced Cell Classifier or BIAS software trains and optimizes machine-learning algorithms to automatically recognize cellular phenotypes. **(e)** Interface to review and select imaged cells. **(f)** Selected cells are extracted by micromanipulation or laser microdissection combined with a catapulting system. **(g)** Outside the CAMI workflow, the collected cells can be characterized at the molecular level (e.g., by digital PCR or next-generation sequencing). Image source: (Brasko et al. 2018)

As shown in the diagram summarizing CAMI technology in Fig. 16, samples are collected in variable formats etched with registration landmarks, and are potentially treated with compounds according to the assay (sample preparation phase). The samples may come from tissues or cell cultures, and are imaged with an automated high-throughput microscope. Images from the microscope are sent to our image analysis software that uses state-of-the-art algorithms to correct illumination, identify and segment cells and extract multiparametric cellular measurements. ACC or BIAS software trains machine-learning algorithms to automatically recognize the cellular phenotype of every cell in the sample, based on their extracted properties. If the user wishes, he/she may add or remove cells, or can correct mistakes in the contour and classified phenotype. Selected cells are then extracted by micromanipulation or laser microdissection combined with a catapulting system, and are collected in a microtube or high-throughput format for molecular characterization such as sequencing or dPCR. The software components we developed to support this technology are freely available. As a proof of principle in (Brasko et al. 2018), we conducted three sets of experiments to demonstrate the capabilities of the technology to target, isolate, and analyze individual cells without disturbing their microenvironment. The rationale behind the selected experimental designs is that none of these analyses could have been executed by conventional automated isolation techniques (e.g., FACS), and alternative solutions would have required laborious manual operation.

4.2 Deep Visual Proteomics

Relevant publication:
Mund, A., Coscia, F., Kriston, A., Hollandi, R., …, **Horvath, P.***, Mann, M.* (2022)
Deep Visual Proteomics defines single-cell identity and heterogeneity
Nature Biotechnology

Modern microscopy's versatility, resolution and multi-modal nature yields increasingly detailed images of single-cell heterogeneity and tissue organization. Early methods combined proteomics with imaging modalities such as imaging mass spectrometry (IMS) or imaging based multiplexed proteomic approaches. Using such methods, a predefined subset of proteins is usually targeted for analysis, but this subset is far short of the actual complexity of the proteome. Taking advantage of the substantially increased sensitivity of technologies based on mass spectrometry (MS), we aimed to enable the analysis of proteomes within their native, subcellular context to explore their contribution to health and disease states (Mund et al. 2022). Similar to CAMI, we combined submicron-resolution imaging, image analysis for single-cell phenotyping based on artificial intelligence (AI) and isolation with an ultra-sensitive proteomics workflow (Fig. 17). We introduced the software ‘BIAS’ (Biology Image Analysis Software), which coordinates image acquisition and laser microdissection (LMD) microscopes.

The technique named deep visual proteomics (DVP) fluently combines microscopy of cell cultures or tissue samples (formalin-fixed, paraffin-embedded, i.e. FFPE) with processing of the samples using the cell segmentation methods described earlier, and phenotyping of the identified samples. The cells found to be of interest are then laser microdissected, either completely automatically or on the basis of expert advice, and ultrasensitive proteomics is performed on the extracted cells. 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.

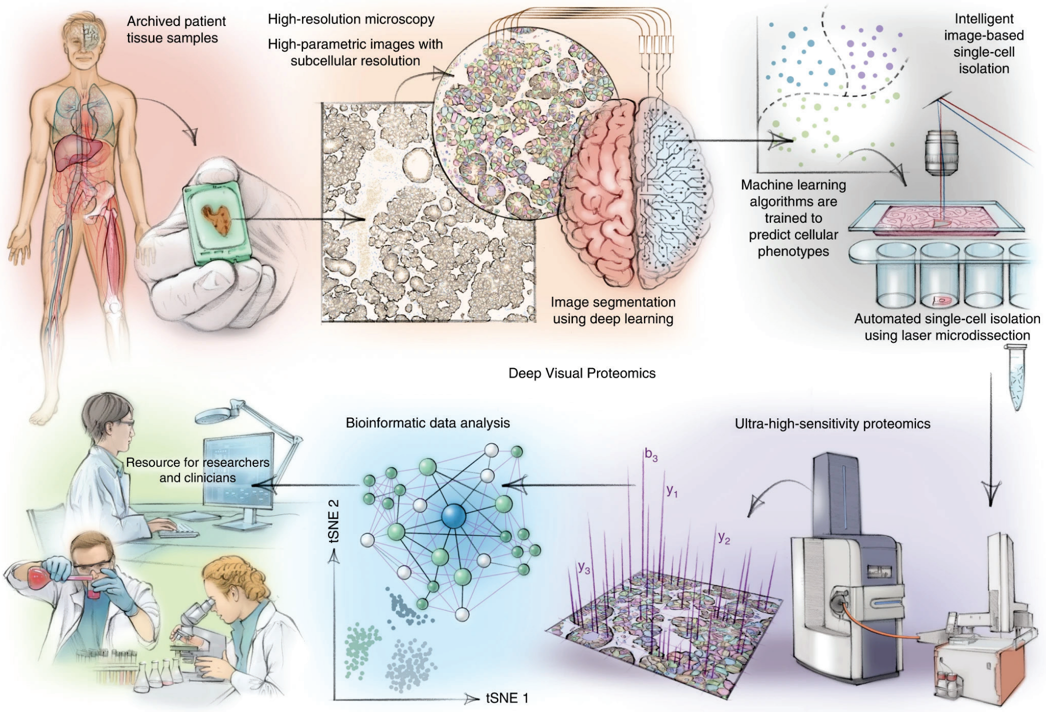

Archived patient tissue samples
High-resolution microscopy
High-parametric images with subcellular resolution
Intelligent image-based single-cell isolation
Machine learning algorithms are trained to predict cellular phenotypes
Image segmentation using deep learning
Automated single-cell isolation using laser microdissection
Deep Visual Proteomics
Bioinformatic data analysis
Ultra-high-sensitivity proteomics
Resource for researchers and clinicians
tSNE 2
tSNE 1
b3
y1
y2
y3

Figure 17. The schematic concept of deep visual proteomics (DVP). DVP combines high-resolution imaging, AI-guided image analysis for single-cell classification and isolation with an ultra-sensitive proteomics workflow. DVP links data-rich imaging of cell culture or archived patient biobank tissues with deep-learning-based cell segmentation and machine-learning-based identification of cell types and states. (Un)supervised AI-classified cellular or subcellular objects of interest undergo automated LMD and MS-based proteomic profiling. Subsequent bioinformatics data analysis enables data mining to discover protein signatures, providing molecular insights into proteome variation in health and disease states, at the level of single cells. tSNE, t-distributed stochastic neighbor embedding. Image source: (Mund et al. 2022)

The software tools presented in chapter 2 and 3 were incorporated into a novel professional software named BIAS (Biological Image Analysis Software, Single-Cell Technologies Ltd.). To physically extract the cellular features discovered with BIAS, we have developed an interface between scanning and laser microdissection microscopes. BIAS transfers cell contours between the microscopes, preserving full accuracy. After optimization, the isolation microscopes are capable of autonomously excise 1,250 high-resolution contours per hour. To prevent potential laser-induced damage to cell membranes, we excise contours with an offset.

An extensive experimental verification of DVP is available in (Mund et al. 2022). Here we present two examples of the potential use of this pipeline. To explore the sensitivity, specificity and robustness of our DVP workflow, we obtained normal human fallopian tube tissue and separated ciliated cells 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. 18). We solely detected FOXJ1 (ciliated cells) in FOXJ1-stained cells (Fig. 18 a, c), along with more than 5,000 other quantified proteins with excellent correlations of biological replicates. Bioinformatic analysis of differences in protein abundance mirrored the biologic features of the distinct cell types (Fig. 18 b, c, d). This was driven by known protein markers of ciliated cells, and was expanded to proteins not yet revealed to be 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 allow the discovery of cell type and cell state markers, and provide unbiased information to understand disease states at the global proteome level.

A more complex example is shown in Figure 19. Understanding the molecular alterations in melanoma development and progression is key to identifying therapeutic vulnerabilities. Regarding that the pathogenic mutations in melanoma are largely cataloged (Pollock et al. 2003), we aimed to directly study spatially resolved proteomes of distinct cellular phenotypes of melanoma progression (Fig. 19 a,b). We co-stained FFPE-embedded primary tumor material preserved for 17 years with two markers, SOX10 and CD146, to map melanoma cells. As overexpression of CD146 is implicated in melanoma progression, and immunotherapy against CD146 targets metastases, we used CD146 as a disease progression marker in our analysis. Deep learning predicted five classes with clearly defined spatial distribution: class 1, melanoma *in situ*; class 2, predominantly tumor; class 3, cells of the tumor microenvironment; class 4, enriched in CD146-high regions; and class 5, enriched in CD146-low regions. We used high-content imaging to determine the required number of cells to identify statistically and analytically robust cellular phenotypes for precise cell type and state isolation within a spatial region. For this reason, we typically collected around 100 cells per sample. Including replicates, we isolated and profiled 27 different samples obtained from seven unique regions of the same tissue section, including normal melanocytes, melanoma *in situ* and primary melanoma from the radial and vertical growth phases (Fig. 19 a–d). We found high quantitative reproducibility among biological replicates, resulting in disease state and region-specific proteomes (Fig. 19 e–g). Pre-cancerous (melanoma *in situ*) and primary melanoma showed differences in proteins involved in immune cell signalling and cell metabolism, and coincided with reduced melanogenesis. The advanced stages (radial and vertical melanoma growth phases) showed well-defined metabolic activation along with disease progression, a known hallmark of human cancers. Expression of proteins involved in oxidative phosphorylation and mitochondria function gradually increased from melanocytes, through melanoma *in situ* to invasive melanoma, indicating a dependency on mitochondrial respiration in the advanced tumor stages (Fig. 19 h–j). Conversely, proteins involved in antigen presentation and interferon response were downregulated in the advanced stages compared to melanoma *in situ*, which is in line with immune evasion strategies described for melanoma.

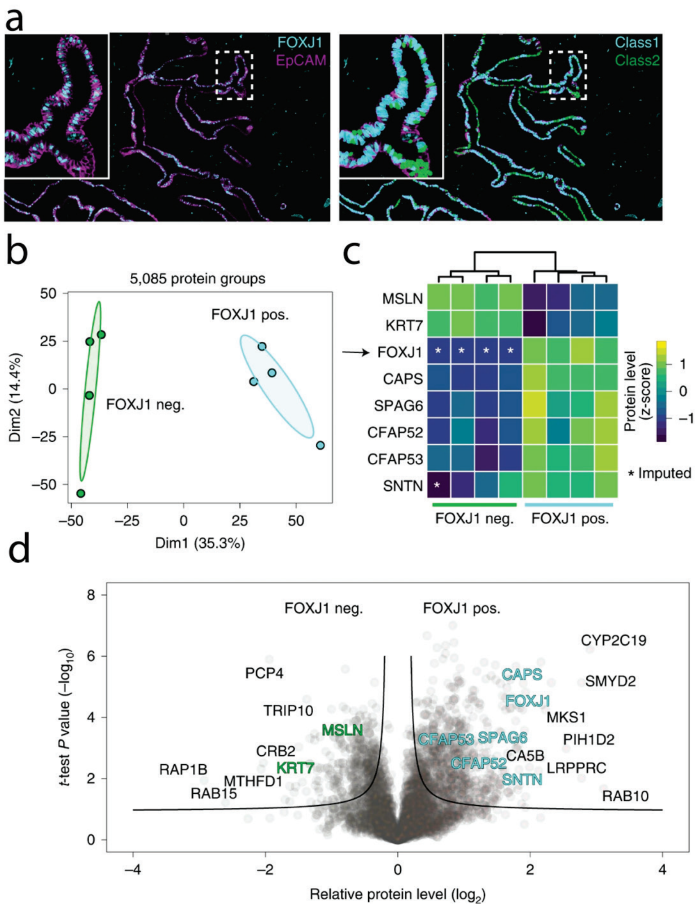


Figure 18. Immunofluorescence staining of the human fallopian tube epithelium with FOXJ1 and EpCAM antibodies, detecting ciliated and epithelial cells, respectively. (a) *Left panel*: Ciliated (FOXJ1-positive) and secretory (FOXJ1-negative) cells. *Right panel*: Cell classification based on FOXJ1 intensity. Class 1 (FOXJ1-positive) and class 2 (FOXJ1-negative); magnification factor = ×387. **(b)** PCA of FOXJ1-positive and FOXJ1-

negative cell proteomes. **(c)** Heat map of known protein markers for secretory and ciliated cells. Protein levels are z-scored. Asterisks represent imputed data. **(d)** 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. Significantly enriched cell-type-specific proteins are displayed above the black lines. Image source: (Mund et al. 2022)

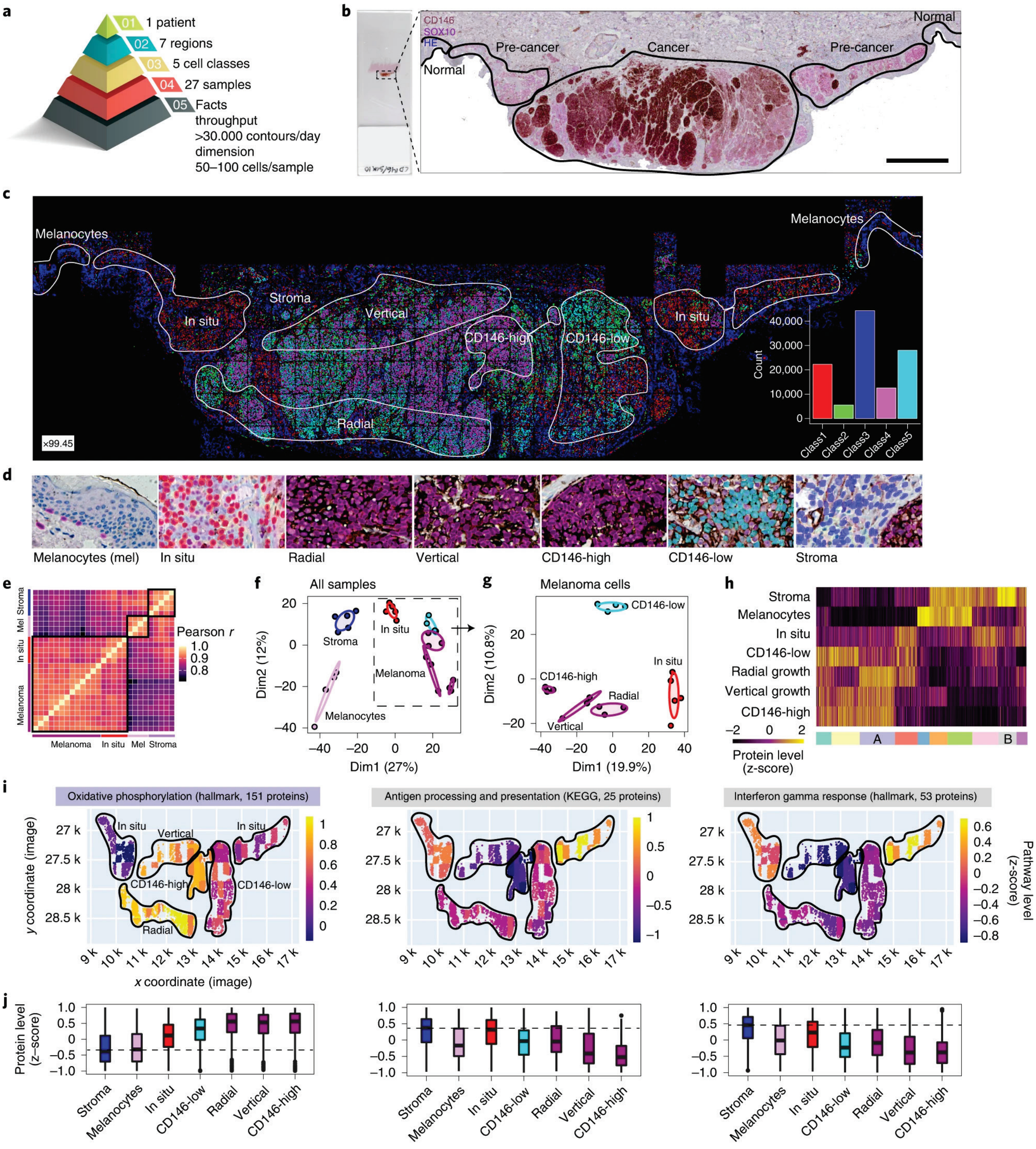


Figure 19. DVP sample isolation workflow to profile primary melanoma. (a) Sample isolation and characteristics **(b)** DVP applied to primary melanoma. **(c)** Pathologist-guided and AI-based cell classification. **(d)** Example pictures of the seven identified classes. **(e)** Correlation matrix (Pearson r) of all the 27 proteome samples assessed. **(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 protein groups revealed as significant by ANOVA (FDR < 0.05). **(i)** Tissue heat map, mapping the proteomics results onto the imaging data. **(j)** Box plots of z-scored protein levels. Image source: (Mund et al. 2022)

4.3 AutoPatcher

Relevant publication:

Koos, K., Olah, G., Balassa, T.,... , Tamas, G., **Horvath, P.** (2021)

Automatic deep learning-driven label-free image-guided patch clamp system

Nature Communications

We have presented a system to overcome time-consuming and expertise-intense neuron characterization and collection. This fully automated, differential interference contrast microscopy (DIC or label-free in general) image-guided patch clamping system (DIGAP or AutoPatcher) combines 3D infrared video microscopy, cell detection using deep convolutional neural networks and a glass microelectrode guiding system in order to approach, attach, break-in, and record biophysical properties of the target cell.

The steps of the visual patch clamp recording process are illustrated in Fig. 20. Before the first use of the system, the pipette has to be calibrated, so that it can be moved relative to the field of view of the camera (1). A position update is executed after every pipette replacement (2) using the built-in pipette detection algorithms (3) to overcome the problem caused by pipette length differences. At this point, the system is ready to perform patch clamp recordings. We have acquired and annotated a single cell image database on label-free neocortical brain tissues, which is, to our knowledge, the largest 3D set of this kind. A deep convolutional neural network has been trained for cell detection. The system can automatically select the detected cell for recording (4). When a cell is selected, multiple subsystems are started simultaneously that perform the patch clamping: (i) A subsystem controls the movement of the micropipette next to the cell. If any obstacle is found in the way, an avoidance algorithm tries to bypass it (5). (ii) A cell tracking system follows the possible shift of the cell in 3D (6). (iii) During the whole process, a pressure regulator system assures that the demanded pressure on the pipette tip is available (7). Once the pipette touches the cell (cell-attached configuration), the system performs gigaseal formation (8), then breaks the cell membrane (9) and automatically starts the electrophysiological measurements (10). When the recording is completed, the operator can decide whether to start the process again on a new target cell or to continue with one or both of the following manual steps. The nucleus or the cytoplasm of the patched cell can be harvested (11), or the recorded cells can be anatomically reconstructed in the tissue (12).

At the end of the measurements, the implemented pipette cleaning method can be performed, or the next patch clamp recording can be started after pipette replacement, from the pipette tip position update step (3). An event logging system collects information during the patch clamp process, including the target locations and outcome success, and report files can be generated at the end. The report files are compatible with the Allen Cell Types Database.

Our system was tested on rodent and human samples *in vitro*. The quality of the electrophysiological measurements strongly correlates to that made by a trained experimenter. We have used the system for harvesting cytoplasm and nucleus from the recorded cells, and performed anatomical reconstruction on the samples. Our system can operate on unstained tissues using deep learning and reaches cell detection accuracy of human experts. Besides, it enables the multiplication of the number of recordings while preserving high-quality measurements.

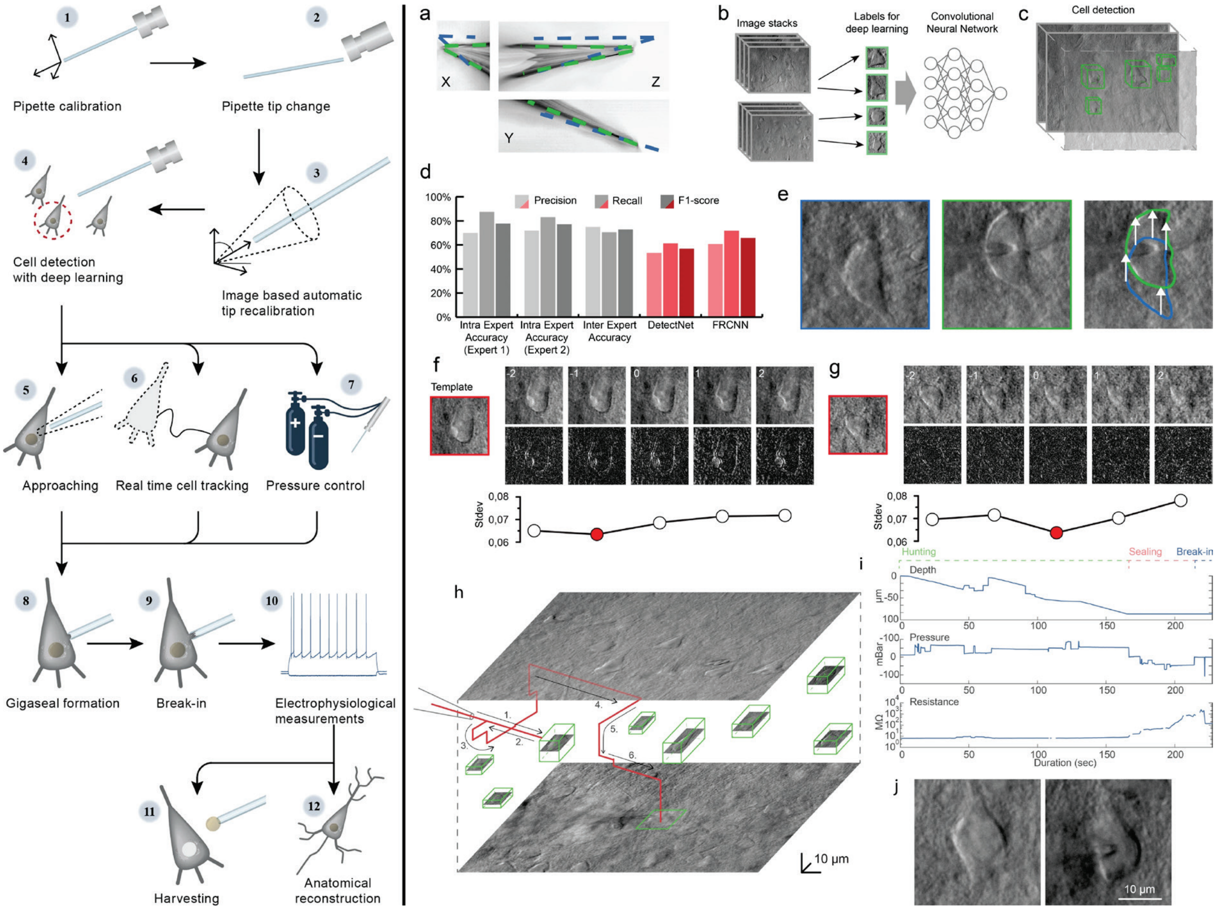


Figure 20. (left) Steps of AutoPatcher procedures. (1) Pipette calibration by the user, **(2)** pipette replacement after recording, **(3)** image-based automatic pipette tip detection, **(4)** automatic cell detection, **(5)** pipette navigation to the target cell, **(6)** 3D cell tracking, **(7)** pressure regulation, **(8)** gigaseal formation, **(9)** break-in, **(10)** electrophysiological recording, **(11)** nucleus and cytoplasm harvesting, **(12)** anatomical reconstruction of the recorded cell. **(right) Computational steps of AutoPatcher (a)** Result of the Pipette Hunter detection model. **(b)** Training dataset generation. **(c)** After the training session, the DIGAP system detects cells in unstained living neocortical tissues. **(d)** Accuracy of the automated cell detection pipeline. **(e)** Lateral tracking of the cell's movement. **(f, g)** Z-tracking of the cell's movement. The template image was captured at the optimal focal depth (in red boxes) before starting the tracking. **(h)** Trajectory of the pipette tip (red line) with obstacle avoidance (numbered) in the tissue, and the spatial location of the detected cells (green boxes). **(i)** Plots of the depth of the pipette tip in the tissue, the applied air pressure, and the measured pipette tip resistance during the approach. **(j)** Image of a cell before and after performing patch clamp recording on it. Image source: (Koos et al. 2021)

We have tested four different object detection deep learning architectures (Szegedy et al. 2015). Of these, the user can choose according to the requirements and the available resources. For this work we used DetectNet. By using these tools, the training processes generated models that recognize neurons in their original environment in DIC images (Fig. 20 b). We also implemented a procedure that extends 2D detection by uniting overlapping bounding boxes along the Z-axis in the image stacks in order to complete object detection in the 3D space (Fig. 20 c). To evaluate the performance of the proposed frameworks we measured precision, recall and F1 score on a validation dataset (Fig. 20 d). Due to the elasticity of the tissue, the movement of the pipette can significantly deform it, and thus can shift the location of the cell of interest. In order to precisely re-define the pipette's trajectory, the location of the target cell needs to be tracked. We have developed an online system that performs tracking in the lateral and Z directions (Fig. 20 e-g). A representative procedure is demonstrated in Figure 20 h-j, trajectory, pressure, and resistance data are visualized.

Conclusions and outlook

In this thesis I have presented how single cells can be profiled both morphologically and molecularly, even in case of a complex spatial environment. Image correction and segmentation methods have been developed to accurately find cells in the samples, and phenotyping algorithms have been developed to automatically identify cellular types and decide on their classifications. With single cell manipulation techniques we are able to extract the desired cell from its native environment for single-cell characterization.

By providing these technologies, I believe that we have opened a door to more precise single cell discoveries. The presented methods can be implemented in an almost standard lab setup, and enable scientists to do their research of interest. I foresee several interesting forthcoming works based on the introduced workflows.

Despite the advancements shown in this thesis, I think this is only the beginning, and there is still a lot to do. The plan of my research groups for the future are to further advance on the fields described here, as well as to make use of the developed technologies both in fundamental research (see the Human Mitosis Atlas below) and in single cell-based cancer therapies (hopeAI). Here I share my point of view on how I envision the advancement of the fields.

Image analysis has probably benefited the most from the rise of deep learning, and this is likely the beginning only. With the exponential improvement of hardware and quantum computing being at the corner, there is a good reason to believe that we will be able to digitally process much larger quantities of image data and will be able to ask more systematic questions, such as structures and patterns in cancers across populations at a single cell level. To this end, my research group will continue developing deep learning tools that perform better segmentation and more global image understanding. Another area that will likely rise in the near future is 3D image generation and processing, due to the fact that spheroid, organoid and other tissue engineered 3D models approximate the physiology of cellular behavior better than flat biology. In fact, we have already made the first baby-steps towards this goal, and will put more resources into the efforts to standardize, accelerate and improve 3D image analysis models.

With the presented tools in hand, we have already taken the first steps in a large-scale initiative, the Human Cancer Mitosis Atlas (MITO-Omics). This will comprehensively describe molecular and morphological changes of cells during mitosis at a nearly infinite resolution. Our plan is to extend our current data to all types of human cancers. MITO-Omics will be a great resource for drug development and fundamental clinical research to better understand and fight against cancers.

Finally, my personal aim is to bring our single-cell methods to personalized cancer therapy. Doing so, I would like to profile cancers at a single cell level, in order to open the way to proposing advanced therapies based on the revealed genomic and proteomic alterations. Definitely, this will be a joint venture between the academia and the industry. We have already taken the first steps towards this goal, and the first clinical success story is delivered (unpublished). Currently we are in discussion with decision makers to make this approach accessible to clinical practice.

Key publications

Single cell image analysis techniques in microscopy (Thesis 1)

Smith K., Li Y., Piccinini, F., Csucs G., Balazs C., Bevilacqua A., **Horvath, P.** (2015)
CIDRE: an illumination-correction method for optical microscopy.
Nature Methods

Horvath, P., Jermyn, I., Kato, Z., Zerubia, J. (2009)
A higher-order active contour model of a "gas of circles" and its application to tree crown extraction
Pattern Recognition., 42(5):699-709

Hollandi, R., …, **Horvath, P.** (2020)
nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer.
Cell Systems

Single cell phenotyping (Thesis 2)

Piccinini, F., Balassa, T., Szkalisity, A., Molnar, Cs., Paavolainen, L., Kujala, K., Buzas, K., Sarazova, M., Pietiainen, V., Kutay, U., Smith K., **Horvath, P.** (2017)
ACC: Discovery software for phenotypic image analysis.
Cell Systems

Szkalisity, A., …, **Horvath, P.** (2021)
Regression plane concept for analysing continuous cellular processes with machine learning
Nature communications

Daly, J.L., …., **Horvath, P.**, …, Yamauchi, Y. (2020)
Neuropilin-1 is a host factor for SARS-CoV-2 infection
Science

Single cell isolation (Thesis 3)

Brasko C., Smith K., Molnar, Cs., …,**Horvath, P.** (2018)
Intelligent image-based in situ single-cell isolation.
Nature Communications

Koos, K., Olah, G., Balassa, T., Mihut, N., Rozsa, M., Ozsvar, A., Tasnadi, E., Barzo, P., Farago, N., Puskas, L., Molnar, G., Molnar, J., Tamas, G., **Horvath, P.** (2021)
Automatic deep learning-driven label-free image-guided patch clamp system
Nature Communications

Mund, A., Coscia, F., Hollandi, R., Kovacs, F., Kriston, A.,...,**Horvath, P.***, Mann, M.* (2022)
AI-driven Deep Visual Proteomics defines cell identity and heterogeneity
Nature Biotechnology

Relevant review articles

Horvath, P., Aulner, N., Bickle, M., Davies, A., Del Nery, E., Ebner, D., Montoya, M., Ostling, P., Pietiainen, V., Price, L., Shorte, S., Turcatti, G., von Schantz, C., Carragher, N. (2016)
Screening out irrelevant cell-based models of disease
Nature Reviews Drug Discovery

Caicedo, J. C., ..., Molnar, Cs., ..., **Horvath, P.**, Linington, R. G., Carpenter, A. E. (2017)
Data-analysis strategies for image-based cell profiling.
Nature Methods

Carragher, N., Piccinini, F., Tesei, A., Trask, J., Bickle, M., **Horvath, P.** (2018)
Concerns, challenges and promises of high-content analysis of 3D cellular models.
Nature Reviews Drug Discovery

Smith K., Piccinini, F., Balassa, T., Koos, K., Danka, T., Azizpour H., **Horvath, P.** (2018)
Phenotypic image analysis software tools for exploring and understanding big image data from cell-based assays.
Cell Systems

Peirsman, A., Blondeel, E., Ahmed, T., ..., Buzas, K., Carragher, N., Diosdi, A., **Horvath, P**., Piccinini, F., ..., De Wever, O. (2021)
MISpheroID: a knowledgebase and transparency tool for minimum information in spheroid identity
Nature methods

Hollandi, R., Moshkov, N., Paavolainen L., Tasnadi, E., Piccinini, F., **Horvath, P**. (2022)
Nucleus segmentation: towards automated solutions
Trends in Cell Biology (CellPress)

References

Altschuler, Steven J., and Lani F. Wu. 2010. “Cellular Heterogeneity: Do Differences Make a Difference?” *Cell* 141 (4): 559–63.

Babaloukas, Georgios, Nicholas Tentolouris, Stavros Liatis, Alexandra Sklavounou, and Despoina Perrea. 2011. “Evaluation of Three Methods for Retrospective Correction of Vignetting on Medical Microscopy Images Utilizing Two Open Source Software Tools.” *Journal of Microscopy* 244 (3): 320–24.

Banerjee, Indranil, Yasuyuki Miyake, Samuel Philip Nobs, Christoph Schneider, Peter Horvath, Manfred Kopf, Patrick Matthias, Ari Helenius, and Yohei Yamauchi. 2014. “Influenza A Virus Uses the Aggresome Processing Machinery for Host Cell Entry.” *Science* 346 (6208): 473–77.

Bodzon-Kulakowska, Anna, and Piotr Suder. 2016. “Imaging Mass Spectrometry: Instrumentation, Applications, and Combination with Other Visualization Techniques.” *Mass Spectrometry Reviews*. https://doi.org/10.1002/mas.21468.

Brasko, Csilla, Kevin Smith, Csaba Molnar, Nora Farago, Lili Hegedus, Arpad Balind, Tamas Balassa, et al. 2018. “Intelligent Image-Based in Situ Single-Cell Isolation.” *Nature Communications* 9 (1): 226.

Caicedo, Juan C., Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, et al. 2017. “Data-Analysis Strategies for Image-Based Cell Profiling.” *Nature Methods* 14 (9): 849–63.

Cai, Yin, M. Julius Hossain, Jean-Karim Hériché, Antonio Z. Politi, Nike Walther, Birgit Koch, Malte Wachsmuth, et al. 2018. “Experimental and Computational Framework for a Dynamic Protein Atlas of Human Cell Division.” *Nature* 561 (7723): 411–15.

Carpenter, Anne E., Thouis R. Jones, Michael R. Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A. Guertin, et al. 2006. “CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes.” *Genome Biology* 7 (10): R100.

Daly, James L., Boris Simonetti, Katja Klein, Kai-En Chen, Maia Kavanagh Williamson, Carlos Antón-Plágaro, Deborah K. Shoemark, et al. 2020. “Neuropilin-1 Is a Host Factor for SARS-CoV-2 Infection.” *Science* 370 (6518): 861–65.

Espina, Virginia, John Milia, Glendon Wu, Stacy Cowherd, and Lance A. Liotta. 2006. “Laser Capture Microdissection.” *Cell Imaging Techniques*. https://doi.org/10.1007/978-1-59259-993-6_10.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.” *Nature* 542 (7639): 115–18.

Falk, Thorsten, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, et al. 2019. "U-Net: Deep Learning for Cell Counting, Detection, and Morphometry." *Nature Methods* 16 (1): 67–70.

Fischer, Ute, Michael Forster, Anna Rinaldi, Thomas Risch, Stéphanie Sungalee, Hans-Jörg Warnatz, Beat Bornhauser, et al. 2015. "Genomics and Drug Profiling of Fatal TCF3-HLF-Positive Acute Lymphoblastic Leukemia Identifies Recurrent Mutation Patterns and Therapeutic Options." *Nature Genetics* 47 (9): 1020–29.

Frismantas, Viktoras, Maria Pamela Dobay, Anna Rinaldi, Joelle Tchinda, Samuel H. Dunn, Joachim Kunz, Paulina Richter-Pechanska, et al. 2017. "Ex Vivo Drug Response Profiling Detects Recurrent Sensitivity Patterns in Drug-Resistant Acute Lymphoblastic Leukemia." *Blood* 129 (11): e26–37.

Giladi, Amir, and Ido Amit. 2017. "Immunology, One Cell at a Time." *Nature*. https://doi.org/10.1038/547027a.

He, Kaiming, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. 2017. "Mask R-CNN." *2017 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2017.322.

Heppner, G. H. 1984. "Tumor Heterogeneity." *Cancer Research* 44 (6): 2259–65.

Hollandi, Réka, Ákos Diósdi, Gábor Hollandi, Nikita Moshkov, and Péter Horváth. 2020. "AnnotatorJ: An ImageJ Plugin to Ease Hand Annotation of Cellular Compartments." *Molecular Biology of the Cell* 31 (20): 2179–86.

Hollandi, Reka, Nikita Moshkov, Lassi Paavolainen, Ervin Tasnadi, Filippo Piccinini, and Peter Horvath. 2022. "Nucleus Segmentation: Towards Automated Solutions." *Trends in Cell Biology* 32 (4): 295–310.

Hollandi, Reka, Abel Szkalisity, Timea Toth, Ervin Tasnadi, Csaba Molnar, Botond Mathe, Istvan Grexa, et al. 2020. "nucleAIzer: A Parameter-Free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer." *Cell Systems* 10 (5): 453–58.e6.

Horvath, Peter, Thomas Wild, Ulrike Kutay, and Gabor Csucs. 2011. "Machine Learning Improves the Precision and Robustness of High-Content Screens: Using Nonlinear Multiparametric Methods to Analyze Screening Results." *Journal of Biomolecular Screening* 16 (9): 1059–67.

Horváth, P., I. H. Jermyn, Z. Kato, and J. Zerubia. 2009. "A Higher-Order Active Contour Model of a 'gas of Circles' and Its Application to Tree Crown Extraction." *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2008.09.008.

Houle, David, Diddahally R. Govindaraju, and Stig Omholt. 2010. "Phenomics: The next Challenge." *Nature Reviews. Genetics* 11 (12): 855–66.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. "Image-to-Image Translation with Conditional Adversarial Networks." *2017 IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.632.

Koos, Krisztian, Gáspár Oláh, Tamas Balassa, Norbert Mihut, Márton Rózsa, Attila Ozsvár, Ervin Tasnadi, et al. 2021. "Automatic Deep Learning-Driven Label-Free Image-Guided Patch Clamp System." *Nature Communications* 12 (1): 936.

Laurell, Eva, Katja Beck, Ksenia Krupina, Gandhi Theerthagiri, Bernd Bodenmiller, Peter Horvath, Ruedi Aebersold, Wolfram Antonin, and Ulrike Kutay. 2011. "Phosphorylation of Nup98 by Multiple Kinases Is Crucial for NPC Disassembly during Mitotic Entry." *Cell* 144 (4): 539–50.

Liesegang, Thomas J. 2001. "The Sequence of the Human Genome. Venter JC,∗∗E-Mail: Humangenome@celera.com Adams MD, Myers EW, et Al. Science 2001;291:1304–1351." *American Journal of Ophthalmology*. https://doi.org/10.1016/s0002-9394(01)01077-7.

Moen, Erick, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. 2019. "Deep Learning for Cellular Image Analysis." *Nature Methods* 16 (12): 1233–46.

Molnar, Csaba, Ian H. Jermyn, Zoltan Kato, Vesa Rahkama, Päivi Östling, Piia Mikkonen, Vilja Pietiäinen, and Peter Horvath. 2016. "Accurate Morphology Preserving Segmentation of Overlapping Cells Based on Active Contours." *Scientific Reports* 6 (August): 32412.

Moshkov, Nikita, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. 2021. "Author Correction: Test-Time Augmentation for Deep Learning-Based Cell Segmentation on Microscopy Images." *Scientific Reports* 11 (1): 3327.

Mund, Andreas, Fabian Coscia, András Kriston, Réka Hollandi, Ferenc Kovács, Andreas-David Brunner, Ede Migh, et al. 2022. "Deep Visual Proteomics Defines Single-Cell Identity and Heterogeneity." *Nature Biotechnology*, May. https://doi.org/10.1038/s41587-022-01302-5.

Pelkmans, Lucas. 2012. "Cell Biology. Using Cell-to-Cell Variability--a New Era in Molecular Biology." *Science* 336 (6080): 425–26.

Peng, Tingying, Kurt Thorn, Timm Schroeder, Lichao Wang, Fabian J. Theis, Carsten Marr, and Nassir Navab. 2017. "A BaSiC Tool for Background and Shading Correction of Optical Microscopy Images." *Nature Communications* 8 (June): 14836.

Piccinini, Filippo, Tamas Balassa, Abel Szkalisity, Csaba Molnar, Lassi Paavolainen, Kaisa Kujala, Krisztina Buzas, et al. 2017. "Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data." *Cell Systems* 4 (6): 651–55.e5.

Pollock, Pamela M., Ursula L. Harper, Katherine S. Hansen, Laura M. Yudt, Mitchell Stark, Christiane M. Robbins, Tracy Y. Moses, et al. 2003. "High Frequency of BRAF Mutations in Nevi." *Nature Genetics* 33 (1): 19–20.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-319-24574-4_28.

Schie, Iwan W., and Thomas Huser. 2013. "Methods and Applications of Raman Microspectroscopy to Single-Cell Analysis." *Applied Spectroscopy*. https://doi.org/10.1366/12-06971.

Singh, S., M-A Bray, T. R. Jones, and A. E. Carpenter. 2014. "Pipeline for Illumination Correction of Images for High-Throughput Microscopy." *Journal of Microscopy* 256 (3): 231–36.

Smith, Kevin, Yunpeng Li, Filippo Piccinini, Gabor Csucs, Csaba Balazs, Alessandro Bevilacqua, and Peter Horvath. 2015. "CIDRE: An Illumination-Correction Method for Optical Microscopy." *Nature Methods* 12 (5): 404–6.

Smith, Kevin, Filippo Piccinini, Tamas Balassa, Krisztian Koos, Tivadar Danka, Hossein Azizpour, and Peter Horvath. 2018. "Phenotypic Image Analysis Software Tools for Exploring and Understanding Big Image Data from Cell-Based Assays." *Cell Systems*. https://doi.org/10.1016/j.cels.2018.06.001.

Sommer, Christoph, Christoph Straehle, Ullrich Kothe, and Fred A. Hamprecht. 2011. "Ilastik: Interactive Learning and Segmentation Toolkit." *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. https://doi.org/10.1109/isbi.2011.5872394.

Strack, Rita. 2022. "Spatial Proteomics with Subcellular Resolution." *Nature Methods* 19 (7): 780.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. "Going Deeper with Convolutions." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2015.7298594.

Szkalisity, Abel, Filippo Piccinini, Attila Beleon, Tamas Balassa, Istvan Gergely Varga, Ede Migh, Csaba Molnar, et al. 2021. "Regression Plane Concept for Analysing Continuous Cellular Processes with Machine Learning." *Nature Communications* 12 (1): 2532.

Tay, Savaş, Jacob J. Hughey, Timothy K. Lee, Tomasz Lipniacki, Stephen R. Quake, and Markus W. Covert. 2010. "Single-Cell NF-kappaB Dynamics Reveal Digital Activation and Analogue Information Processing." *Nature* 466 (7303): 267–71.

Toth, Timea, Tamas Balassa, Norbert Bara, Ferenc Kovacs, Andras Kriston, Csaba Molnar, Lajos Haracska, Farkas Sukosd, and Peter Horvath. 2018. "Environmental Properties of Cells Improve Machine Learning-Based Phenotype Recognition Accuracy." *Scientific Reports* 8 (1): 10085.

Yamauchi, Yohei, Heithem Boukari, Indranil Banerjee, Ivo F. Sbalzarini, Peter Horvath, and

Ari Helenius. 2011. “Histone Deacetylase 8 Is Required for Centrosome Cohesion and Influenza A Virus Entry.” *PLoS Pathogens* 7 (10): e1002316.

CIDRE: an illumination-correction method for optical microscopy

Kevin Smith[1], Yunpeng Li[2], Filippo Piccinini[3], Gabor Csucs[1], Csaba Balazs[1], Alessandro Bevilacqua[3,4] & Peter Horvath[5–7]

Uneven illumination affects every image acquired by a microscope. It is often overlooked, but it can introduce considerable bias to image measurements. The most reliable correction methods require special reference images, and retrospective alternatives do not fully model the correction process. Our approach overcomes these issues for most optical microscopy applications without the need for reference images.

No optical system is ideal. Inhomogeneous illumination is present in every image acquired by a microscope. Many factors, including misaligned optics, dust, nonuniform light sources and vignetting, contribute to uneven illumination[1]. It is increasingly common for light microscopes to be used as quantitative instruments even though seemingly minor shifts in illumination can corrupt measurements and invalidate subsequent analyses. For example, we found that uneven illumination increased the false detections and missed detections by CellProfiler[2] on images of yeast cells by 35% when illumination correction was neglected (**Supplementary Fig. 1c–f**). Other routine measurements can be affected as well. Uneven illumination substantially reduced the measurements of the mean intensity and mean area of GFP-stained HeLa cells in the corner of the image relative to the center (**Supplementary Fig. 1g–l**).

The consequences of ignoring uneven illumination are often underestimated, as reflected in our survey of microscope users (**Supplementary Note 1**). The magnitude of intensity loss attributed to vignetting, that is, falloff of intensity from the center of the image, is often substantially stronger than assumed. Data from 11 ordinary microscope setups revealed that between 10% and 40% less light is typically recorded at the dimmest region of the image (**Supplementary Data 1**). Intensity loss is even more severe for cameras with large sensor areas or wide apertures, such as scientific complementary metal-oxide semiconductor (sCMOS) devices, which can experience a falloff greater than 50% (**Supplementary Data 1**).

The most common approach for correcting uneven illumination reverses the image formation process, attempting to recover the true image, I, from the image observed by the sensor, I_0. Distortions to the observed image are modeled by a linear intensity gain function v and an additive term z; $I_0(x) = I(x)v(x) + z(x)$, where $I_0(x)$ is the intensity observed at location x. The intensity gain models attenuations to the signal (**Fig. 1a**). An additive or zero-light term models contributions present even if no light is incident on the sensor, mainly camera offset and fixed-pattern thermal noise. It is usually nearly uniform, varying by only a few intensity values. The uncorrupted image is recovered by reversing the image formation process.

$$I(x) = \frac{I_0(x) - z(x)}{v(x)} \qquad (1)$$

Although simple at first glance, in practice v and z cannot be known exactly, which has prompted the development of a variety of correction methods (**Supplementary Note 2**). Prospective methods estimate the correction surfaces from special reference images collected during acquisition[3,4], whereas retrospective methods rely on the actual image data[5,6]. Prospective methods are regarded as more reliable because they empirically estimate the terms in equation (1), whereas retrospective methods are more practical because they do not require special acquisitions.

We introduce a new retrospective method, corrected intensity distributions using regularized energy minimization (CIDRE), which achieves correction quality similar to that of prospective methods. Unlike existing retrospective methods, CIDRE estimates both v and z. The key insight to our approach is that the distribution of intensities from a single location across many images is related to an underlying distribution common to all locations by a linear transform. This assumes that objects may appear anywhere in the image with equal probability. Local-intensity distributions from a finite set of observed images are simply linear transforms of a sampling of the underlying distribution (**Fig. 1b**). The parameters of the transform can be visualized in a quantile-quantile plot, where v corresponds to the slope and b corresponds to the y intercept.

CIDRE estimates v, b and z simultaneously for all locations by minimizing a regularized energy function composed of several terms (**Fig. 1c**). The first term is a robust regression that ensures v and b fit the data (**Fig. 1d**). The second term reduces noise and guarantees a smooth correction surface, encouraging neighboring distributions to agree on similar values for v (**Fig. 1e**). The third term estimates z by finding the common point where all regression lines intersect (**Fig. 1f**). Corrected

[1]Scientific Center for Optical and Electron Microscopy, ETH Zürich, Zurich, Switzerland. [2]Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [3]Advanced Research Center on Electronic Systems, University of Bologna, Bologna, Italy. [4]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy. [5]Institute of Biochemistry, ETH Zürich, Zurich, Switzerland. [6]Synthetic and System Biology Unit, Biological Research Center, Szeged, Hungary. [7]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. Correspondence should be addressed to K.S. (kevin.smith@scopem.ethz.ch).

RECEIVED 22 APRIL 2014; ACCEPTED 16 DECEMBER 2014; PUBLISHED ONLINE 16 MARCH 2015; DOI:10.1038/NMETH.3323

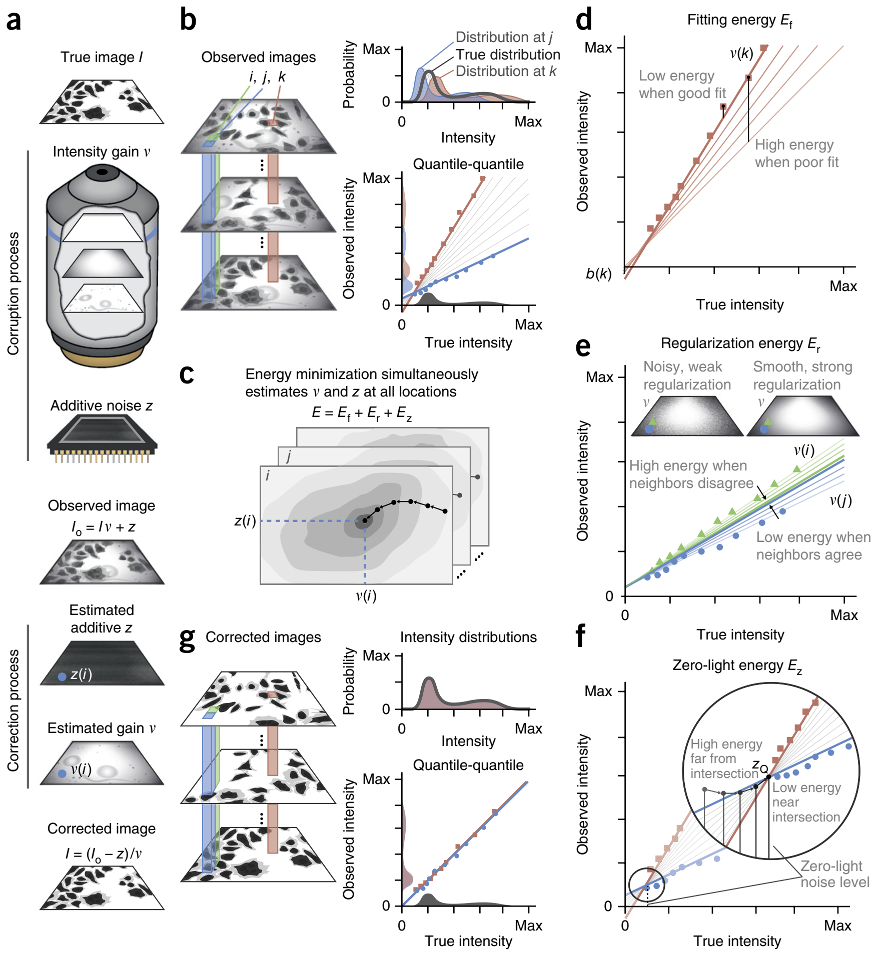


Figure 1 | Summary of illumination correction using CIDRE. (**a**) The observed image is corrupted by misaligned optics, vignetting, dust, etc., and this corruption is modeled as a linear intensity gain function v. The observed image also contains contributions from camera offset and thermal sources, modeled as an additive term z. (**b**) To recover the true image, we consider the local distributions of observed intensities (red, blue and green) collected from many images. Each is related to the true underlying distribution of the specimen (gray, upper plot) by a linear transform parameterized by v and b, which correspond to slope and y intercept, respectively, in a quantile-quantile plot between the true distribution and the local distribution. (**c**) The plot shows an estimation of the parameters in **b** simultaneously for all locations by minimizing a regularized energy function composed of several terms. (**d–f**) Terms composing the energy function in **c**. The plot in **d** shows a robust regression term that ensures parameter values fit the data. The plot in **e** shows a regularization term that reduces noise and guarantees smoothness of the correction surface, forcing neighboring distributions (green and blue) to agree on similar parameter values. The plot in **f** shows how z, a nearly uniform-intensity surface representing the intensity recorded when no light is present, is estimated by finding the common point where all regression lines intersect. (**g**) When we apply the reverse transform using estimates of v and z, local-intensity distributions (red, blue and green) take the shape of the true distribution (gray, upper and lower plots), and the uncorrupted images can be recovered.

images are obtained using equation (1) (**Fig. 1g**). A complete explanation is provided in **Supplementary Note 3**.

We compared our approach to 12 commonly used methods in a series of tests on 12 data sets. Eleven of these data sets represent typical microscopy setups with various microscopes, sample preparations, staining, light sources, magnifications and sensors (**Supplementary Table 1** and **Supplementary Data 1**). As a gold standard, we used prospective correction in which v is estimated as the average of several reference images or empty images and z is estimated as the average of several dark-frame images. To measure correction quality, we collected hundreds of pairs of overlapping images for each data set, precisely aligned them and reported the mean of absolute differences for each pair of corresponding pixels in the overlapping regions (**Supplementary Data 2**). The scores are normalized by the mean score of the uncorrected pairs (**Fig. 2**). Although this provides a reasonable estimate of the correction quality, unavoidable differences between image pairs may cast some doubt on our measure. To address this, we generated a twelfth data set of synthetic images and distorted them using a known model (**Supplementary Note 4**), allowing us to directly measure the disagreement between the true and corrected images.

Our tests highlight the inadequacies of current illumination-correction practices. Single-image retrospective methods generally made the illumination more uneven instead of correcting it. Multi-image retrospective methods showed some improvement, but no method consistently achieved satisfactory performance (that is, better than untreated). The gold-standard prospective methods performed best among existing methods, but their reliance on reference images limits their usefulness.

CIDRE is a retrospective method and thus does not require calibration images; therefore, it can always be applied to previously collected data. Unlike other retrospective methods, it is capable of estimating the z term, which helped it to perform

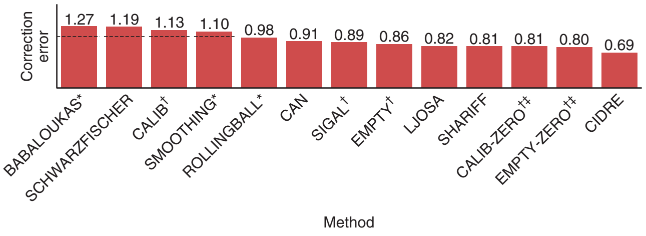


Figure 2 | Comparison of 13 illumination correction methods, averaged over 12 image collections. Scores are the mean absolute differences between pairs of overlapping test images after correction, computed pixelwise and normalized by the disagreement between uncorrected image pairs. A score of 1 implies equivalent disagreement to the uncorrected image pairs, indicated by dashed lines. A score of 0 implies a perfect correction, though it is not achievable in practice. *Single-image methods. †Prospective methods. ‡Gold-standard methods. Methods and data sets are described in the Online Methods.

well consistently on every data set. In terms of correction quality, it surpassed all tested methods including the gold standard, and it also substantially reduced errors in cell counting, cell intensity and cell area measurements (**Supplementary Fig. 1**). CIDRE is available as open-source software in Matlab (https://www.github.com/smithk/cidre/) and as an ImageJ plug-in (**Supplementary Software**).

Although CIDRE is useful for many applications, it is not suitable in certain conditions. The key assumption is violated if the images are highly correlated. In time-lapse images, for example, this may cause artifacts in the correction, although combining images from different sites can help reduce this danger. Like other retrospective methods, CIDRE performs best with many images containing ample intensity information. With 1,000 images or more, CIDRE was substantially better than the gold standard on average (**Fig. 2** and **Supplementary Fig. 2**). Fewer images or sparse intensity information reduced correction quality. We found that for most applications, ten images were sufficient to ensure an improvement in illumination quality, whereas approximately 100 images were necessary to match the gold standard (**Supplementary Fig. 3**).

Uneven illumination is a common but misunderstood phenomenon, and results reported by many researchers have undoubtedly been affected by it. Proper illumination-correction procedures should be followed whenever acquiring images for quantitative microscopy, and CIDRE is a simple, freely available tool that can help ensure the quality of such measurements.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank R. Dechant, T. Schwarz, A. Kaufmann and J. Kusch for data collection; L. Lenherr Smith for help with the manuscript; and our colleagues who completed the survey. This work was funded by SystemsX.ch, the Swiss national initiative for Systems Biology, through the SyBIT project. P.H. acknowledges support from the Finnish TEKES FiDiPro and the Hungarian National Brain Research Programme (MTA-SE-NAP B-BIOMAG).

AUTHOR CONTRIBUTIONS

P.H., F.P., A.B. and G.C. initiated the project. K.S. and P.H. designed the correction method. K.S. and Y.L. designed the optimization. K.S. and C.B. implemented the software. F.P. and K.S. collected the image data. K.S. performed the analysis and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Goldman, D.B. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 2276–2288 (2010).
2. Carpenter, A.E. *et al. Genome Biol.* **7**, R100 (2006).
3. Young, I.T. *Curr. Protoc. Cytom.* **14**, 2.11 (2001).
4. Model, M. *Curr. Protoc. Cytom.* **68**, 10.14 (2014).
5. Ljosa, V. & Carpenter, A.E. *PLoS Comput. Biol.* **5**, e1000603 (2009).
6. Shariff, A., Kangas, J., Coelho, L.P., Quinn, S. & Murphy, R.F. *J. Biomol. Screen.* **15**, 726–734 (2010).

ONLINE METHODS

Software. Our software, CIDRE, is available as an ImageJ plug-in and as a Matlab script. The open-source code for CIDRE is available as **Supplementary Software** and for download at https://www.github.com/smithk/cidre/. Compiled versions of the code and updates are available at http://www.scopem.ethz.ch/research/software/cidre.html.

Microscopy data sets. We collected 11 data sets designed to represent a variety of microscopy setups commonly used for research to evaluate illumination-correction methods (**Supplementary Table 1**). Misaligned optics, dust, scratches, vignetting, and nonuniform light sources contribute to uneven illumination in these data sets. The most common source of uneven illumination is vignetting, a radial intensity falloff from the center of the image[7] caused by blockages of extreme light paths, varying angular sensitivity of the sensor, and natural geometric falloff[1]. Images were acquired with eight different microscopes and one conventional SLR camera. We varied the microscope itself, sample preparation, staining, mounting, objectives, light sources, type of sensor, image resolution, and bit depth. In addition, we generated a set of synthetic images to form a twelfth data set allowing us to directly measure correction quality. Each data set contains several sets of images: approximately 100 images of a reference material (for example, fluorescent slide), 10–100 'empty' images (for example, the culture medium or the glass slide), 10–100 dark-frame images where no light is incident on the sensor (acquired with the light source switched off or otherwise blocked), 10–100 images in a series with increasing exposure times, 100–200 pairs of images used to test the correction models, 1,000–3,000 images used to train the correction models, and 1,000–3,000 segmentations of the training images obtained with CellProfiler[2]. Below we give a brief description of each data set (**Supplementary Data 1**).

HCS-DNA: this data set contains propidium iodide (P3566 Invitrogen)-stained ATCC HeLa cells from a high-content screen. The compound stains DNA and RNA, visualizing the nuclei and cytoplasm as well as clumps of Semliki Forest virus. Images were collected on the RFP channel of a fluorescence wide-field microscope using 384-well plates with a plastic-air interface. The microscope was a Molecular Devices ImageXpress Micro with a 20×/0.75-NA objective with a LED-liquid light guide (Lumencor Spectra X). A CCD sensor (Photometrics CoolSNAP HQ) acquired 1,392 × 1,040 12-bit images. The mean signal intensity in the images was low (219). Reference images were acquired at various locations on a plastic fluorescent reference slide and empty wells containing fluorescent medium. The intensity attenuation measured as the difference between the 99th percentile and 1st percentile intensities computed from the mean of the reference images was 42.6%. The confluency, measured as the percentage of the image area occupied by cells, was 21.5%.

HCS-NUC: this data set contains DAPI-stained MRC5 human fetal lung fibroblasts (SV40 immortalized) from a high-content screen. Images were collected using a second microscope with a similar setup to that of HDC-DNA but with a DAPI-adapted filter set. The mean signal intensity was moderate (549), intensity attenuation was 18.7%, and cell confluency was low (2.7%).

HCS-ACTIN: this data set contains actin-stained HeLa cells from a high-content screen. Images were collected using the same microscope and setup as for HCS-NUC, except that a different filter set and light source were used (Lambda LS arc lamp, Sutter Instruments, with a 300-W xenon bulb, PerkinElmer). The mean signal intensity was bright (1,776), intensity attenuation was 23.3%, and confluency was high (88.2%).

MICROFLUID: this data set contains yeast cells in a synthetic medium containing glucose and dextran coupled with Alexa Fluor 680, used to study regulation of the PKA pathway through V-ATPase. A microfluidic setup was used (CellAsic Microfluidic plate Y04C), and images were acquired using a Nikon Ti-Eclipse microscope with a 60×/1.4-NA objective and LED-liquid light guide (CoolLED pE-2). A high-resolution sCMOS sensor (Hamamatsu Orca Flash 4.0) was used to acquire 2,048 × 2,048 16-bit images. Reference images were acquired from various locations on a plastic fluorescent reference slide and empty wells containing fluorescent medium. The mean signal intensity was bright (1,086), intensity attenuation was very strong 50.4%, and cell confluency was very low (0.6%).

FLUO-ACTIN: this data set contains muntjac skin fibroblast cells on a glass slide (FluoCells prepared slide F36925). The prominent filamentous actin in these cells was labeled with Alexa Fluor 488. Images were collected using a fluorescence wide-field microscope with a glass-oil interface. The microscope model was a Zeiss 200M with a 63×/1.4-NA objective and a HBO-direct light source with a 100-W Mercury arc lamp. A CCD sensor (Hamamatsu C8484) acquired 1,344 × 1,024 12-bit images. Reference images were acquired at various locations on a plastic fluorescent reference slide. The mean signal intensity was low (213), intensity attenuation was weak (7.5%) and cell confluency was moderate (32.5%).

FLUO-MITO: this data set contains bovine pulmonary artery endothelial (BPAE) cells on a glass slide (FluoCells prepared slide F36924). Mitochondria were labeled with red fluorescent MitoTracker Red CMXRos. Images were collected using a fluorescence wide-field microscope with a glass-air interface. The microscope model was an Olympus IX81 with a 20×/0.75-NA objective and a Lambda LS arc lamp (Sutter Instruments) with a 175-W xenon bulb (PerkinElmer). The CCD sensor (Hamamatsu Orca ER) acquired 1,344 × 1,024 12-bit images. Reference images were acquired at various locations on a plastic fluorescent reference slide. The mean signal intensity was low (338), intensity attenuation was strong (40.3%), and cell confluency was low (13%).

FLUO-EMCCD: this data set contains bovine pulmonary artery endothelial (BPAE) cells on a glass slide (FluoCells prepared slide F14781). Anti-bovine alpha-tubulin mouse monoclonal in conjunction with goat anti-mouse antibodies label the microtubules. Images were collected using a fluorescence wide-field microscope with a glass-air interface. The microscope model was a Leica DMI6000B with a 40×/0.7-NA objective and a 100-W metal-halide lamp. An EMCCD sensor (Andor iXon 885) acquired 1,004 × 1,002 16-bit images with the EM gain set at a medium setting (3). Reference images were acquired at various locations on a plastic fluorescent reference slide. The mean signal intensity was bright (2,155), intensity attenuation was 38%, and cell confluency was moderate (34.3%).

HIST-CONFOCAL: this data set contains a 16-µm cryostat section of mouse kidney on a glass slide stained with Alexa Fluor 488 to label elements of the glomeruli and convoluted tubules (FluoCells prepared slide F24630). Images were collected using a fluorescence confocal microscope with a glass-air interface.

doi:10.1038/nmeth.3323

The microscope model was a Zeiss LSM 510 with a 10×/0.3-NA objective and a 30-mW argon 488-nm laser with an optical fiber light guide. A Hamamatsu photomultiplier tube (PMT) sensor acquired 512 × 512 12-bit images. Reference images were acquired at various locations on a plastic fluorescent reference slide. The mean signal intensity was medium (614), intensity attenuation was strong (43.9%), and confluency was high (54%).

HIST-BRIGHT: this data set contains a sagittal section of mouse brain on a glass slide labeled with hematoxylin and eosin (HE) stain. Bright-field images were collected using a wide-field microscope with a glass-oil interface. The microscope model was an Olympus IX81 with a 40×/1.3-NA objective and Lambda LS arc lamp (Sutter Instruments) with a 175-W xenon bulb (PerkinElmer). A CCD sensor (Hamamatsu Orca ER) acquired 1,344 × 1,024 12-bit images. Reference images were acquired using empty locations on the glass slide. The mean signal intensity was bright (614), and intensity attenuation was 32%.

PHASE: this data set contains native epithelial cells from oral mucosa on a glass slide. Phase-contrast images were collected using a wide-field microscope with a glass-air interface. The microscope model was an Olympus IX81 with a 10×/0.3-NA objective and 100-W halogen lamp. A CCD sensor (Hamamatsu Orca ER) acquired 1,344 × 1,024 12-bit images. Reference images were acquired using empty locations on the glass slide. The mean signal intensity was moderate (923), intensity attenuation was 12.4%, and cell confluency was low (7.1%).

PHOTOGRAPHY: this data set contains conventional photographs of projections of paintings and drawings from the National Gallery of Art. A consumer digital SLR camera was (Nikon D90) was used with an 18- to 105–mm zoom lens (set at 50 mm). Reference images were acquired using empty sheets of white paper. A CMOS sensor (Nikon DX format) acquired 1,072 × 712 8-bit images. The mean signal intensity was bright (150), and intensity attenuation was 17%.

SYNTHETIC: this data set contains synthetically generated images of HeLa cells. Uncorrupted 1,392 × 1,040 12-bit images were synthetically generated and artificially subjected to illumination distortions using a physically plausible model (**Supplementary Note 4**). Reference images were synthetically generated in a similar manner. Because the corruption process is known and the image pairs are perfectly aligned, this data set allows us to directly measure differences between corrections and uncorrupted images.

Baseline illumination-correction methods. We compared our method to 12 commonly used prospective and retrospective approaches for illumination correction (**Supplementary Note 2**). Each method is summarized below, starting with prospective correction methods, which require special calibration images to be acquired along with the data.

CALIB-ZERO models the illumination gain v as the average of a set of images of a plastic fluorescent reference slide[8] at various locations. The zero-light term z is modeled by averaging a set of images taken at various locations with the shutter closed or the light source turned off or otherwise blocked. Image correction follows equation (1), reversing the image formation process. This approach is considered a gold standard because it empirically models both correction terms. However, it has several drawbacks (**Supplementary Note 2**).

CALIB models the illumination gain v as above but ignores the zero-light term[8]. Image correction is done by normalizing the image by the gain $I(x) = I_0(x)/v(x)$. This is an incomplete model, but it can give reasonable results when the zero-light term is small relative to the calibration intensity.

EMPTY-ZERO models the illumination gain v as the average of empty' images taken at various locations[3]. In our experiments, this includes either the culture medium without any cells or the glass slide. The zero-light term z is modeled by averaging a set of images taken with the shutter closed. Image correction reverses the image formation process (equation (1)). This approach also serves as a gold standard because it empirically models the correction terms, but it suffers from the same drawbacks as CALIB-ZERO (**Supplementary Note 2**). This method is appropriate for bright-field images or fluorescence imaging when the medium fluoresces.

EMPTY models the illumination gain v as above but ignores the zero-light term[8]. Image correction is done by normalizing the image by the gain $I(x) = I_0(x)/v(x)$. Although flawed, this approach can give reasonable results when z is small.

SIGAL uses a specially acquired set of images of a homogeneous medium captured with increasing exposure times to build the correction model[9]. Estimates of v and z are obtained by performing a least-squares fitting on data from the exposure series, giving additional weight to data from zero exposure time to anchor the zero-light level. Image correction reverses the image formation process (equation (1)). Because the exposure series often contains only a few data points (7 in the original paper[9], 13–100 in our data) and are not regularized, estimates of v and z tend to be noisy and non-smooth.

Next, we describe four purely computational retrospective methods that build the correction model from more than one image.

CAN builds a model of the illumination gain by sorting image data by location[10]. The top 10th-percentile data are used to fit a polynomial surface in the log-intensity space to estimate v. Images are corrected according to $I(x) = I_0(x)/v(x)$. This approach ignores the zero-light term, but it can give reasonable results when that term is small relative to the signal.

LJOSA forms the illumination gain estimate v by fitting a polynomial or spline surface to an average of the images to be corrected[5]. The fitted surface is used for correction according to $I(x) = I_0(x)/v(x)$. This approach ignores the zero-light term, but it can give reasonable results when that term is small relative to the signal.

SCHWARZFISCHER is a method designed to correct time-lapse fluorescence images, though it can be used for static images if the background contains illumination information[11]. Each image is broken into overlapping tiles, and each tile is clustered as either background or signal on the basis of local statistics. Next, background models are constructed for each image using natural-neighbor interpolation to smooth and fill in missing values. A foreground model is estimated using least-squares fitting at every location on the background data. The images are corrected with a formula similar to equation (1). This method can compensate for photobleaching as well as uneven illumination, but it becomes unreliable for high-confluency images because it depends on a large background area (which is sparse when many cells are present). Comparing intensities between images

doi:10.1038/nmeth.3323

corrected with this method is problematic because the correction model is adapted to each image.

SHARIFF is the default method[6] used in the software package CellProfiler[2], which supports several options for illumination correction. The intensity gain is estimated using the mean image intensity computed at every location. It is smoothed with a median filter and rescaled to an appropriate range. Images are corrected according to $I(x) = I_0(x)/v(x)$.

Finally, we describe three single-image retrospective correction methods. These methods are applied independently to every image and do not require any calibration images. Although they are useful for improving the appearance of the image, these methods are inappropriate if intensity measurements will be made from the image because they nonuniformly alter the signal.

ROLLINGBALL, or the top-hat transform, applies a shape filter that subtracts a geometric opening of the image from the image[12]. The 'ball' in this case refers to the shape of the filter kernel, which is sized to be larger than the diameter of the largest expected object. For each data set, we tested several kernel sizes and chose the one with the best performance.

SMOOTHING also relies on subtracting a filtered version of the image from the original[13]. In this case, a Gaussian smoothing kernel is used, which is sized to be larger than the diameter of the largest expected object.

BABALOUKAS selects locations from the background of the image, fits a polynomial to those points, and subtracts the fitted surface from the original image[14]. In the original paper, background points were selected manually. We automatically selected background points using the clustering technique proposed by Schwarzfischer *et al.*[11].

Evaluation protocol. To assess the quality of correction (**Fig. 2** and **Supplementary Figs. 2** and **3**), we defined an evaluation procedure, error measure, and benchmark score. Measuring correction quality is difficult because we cannot directly compare against the uncorrupted images. Some authors use the coefficient of variation (CV) of corrected reference images[15], but differences between the reference images and actual data cast doubt on the validity of this measure. Our solution was to collect hundreds of pairs of overlapping images for each data set, precisely align them, and report the mean of absolute differences for each pair of corresponding pixels in the overlapping regions (**Supplementary Data 2**). We collected 100–200 pairs of test images for every data set. The various correction methods were applied to each data set. Prospective methods used sets of reference images to train the model, whereas multi-image retrospective methods were trained using a collection of 1,000–3,000 images acquired under similar conditions as the test set. Single-image retrospective methods operate directly on the test images. Each pair of test images was aligned using a subpixel registration technique[16]. The overlapping regions from each test pair were used in the evaluation (**Supplementary Data 2**). Whenever possible, we used a four-quadrant solution wherein quadrant IV of the first image overlapped quadrant II of the second image. Our reasoning is that the most dramatic intensity difference often occurs between a corner of the image and the center, and this format allowed us to directly compare the corner of one image to the center of the other. However, our ability to automate the data collection process was limited by the microscope software and size of the specimen, so this layout was not always possible. In two cases (FLUO-MITO and PHASE), we were forced to overlap the left half of the first image with the right half of the second image. In the case of the SYNTH data set, each pair of test images is perfectly aligned, and so the entire image overlaps. For a given test pair, the overlap regions are two separate acquisitions of the same field of view, denoted IM1 and IM2. The principal sources of disagreement between IM1 and IM2 are uneven illumination, noise, alignment errors, and photobleaching. The error is the mean of absolute differences between all aligned pixel pairs in the overlapping regions. More precisely

$$e = \frac{1}{N}\sum_{n}\sum_{l}\left|I_1^n(l) - I_2^n(l)\right|$$

where e is the error, n indexes the image pair, l indexes locations in the overlapping region, and N is the total pixel pairs across all images pairs. To correct for photobleaching, and because the outputs of the correction methods have different dynamic ranges (some methods are purely multiplicative, some purely subtractive, etc.), we standardized the intensity distributions of the images, thus ensuring comparability across correction methods. The median and s.d. of the set of test image sets were normalized to those of the uncorrected IM1 image set. Finally, the mean error for all image pairs is normalized by the benchmark error, that of the uncorrected image pairs, with

$$s_{\text{method}} = \frac{\bar{e}_{\text{method}}}{\bar{e}_{\text{UNCORRECTED}}}$$

This yields a score s in the interval $[0, \infty)$, where 0 indicates perfect correction (IM1 and IM2 are the same), 1 indicates disagreement equivalent to the uncorrected images, and >1 indicates greater disagreement than the uncorrected images (i.e., the correction process increased disagreement between image pairs).

Assumptions. Our correction method makes the following assumption: the probability of the specimen appearing at location x in an image is uniform for all x, and there is no correlation between the images. In other words, the content of the image can appear anywhere in the image with equal probability. For example, a cultured cell or a piece of tissue are just as likely to appear in the corner of the image as in the center. A second assumption is based on empirical observations: the zero-light noise $z(x)$ should be approximately uniform for all x (flat across the image). This is based on evidence from measurements of the zero-light noise from the data we collected and reliable sensor manufacturing. To verify this assumption we captured and averaged dark frames from each data set. We found that the s.d. of such images was less than 0.1% of the dynamic range, and often less than 0.02%.

Illumination-correction formulation. Our method for illumination correction, CIDRE (corrected intensity distributions using regularized energy minimization), operates on a set of observed images corrupted by illumination distortion that were acquired under similar conditions. CIDRE recovers the uncorrupted images by first estimating the unknown parameters v and z through an energy-minimization technique and then applying equation (1) to recover the uncorrupted images (**Supplementary Note 3**).

doi:10.1038/nmeth.3323

It is assumed that objects may appear anywhere in the image with equal probability. If this assumption is valid, then the distribution of intensities at a single location taken from infinitely many images is related to an underlying distribution common to every location by a linear transform modeling the corruption process. Local-intensity distributions from a finite set of observed images are simply linear transforms of a sampling of the underlying distribution parameterized by v, b and z. CIDRE estimates these parameters simultaneously for all locations by using a quasi-Newton method[17] to minimize a regularized energy function comprising a robust regression term, a smoothness term, and an offset term. The robust regression term ensures v and b fit the data (**Fig. 1d**). The smoothness term enforces a smooth correction surface and reduces noise by encouraging neighboring distributions to agree on similar values for v (**Fig. 1e**). The offset term estimates z by finding the common point where all regression lines intersect (**Fig. 1f**). The results of the optimization are robust estimates for v and z, which allow us to apply equation (1) and recover the uncorrupted images.

7. Ray, S.F. *Applied Photographic Optics* 3rd edn. (Focal, 2002).
8. Varga, V.S. *et al. Cytometry A* **60**, 53–62 (2004).
9. Sigal, A. *et al. Nat. Methods* **3**, 525–531 (2006).
10. Can, A. *et al.* in *Proc. IEEE Int. Symp. Biomed. Imaging* 288–291 (IEEE, 2008).
11. Schwarzfischer, M. *et al.* in *Micro. Image Anal. Appl. Biol.* (MIAAB, Heidelberg, Germany, 2011).
12. Sternberg, S.R. *Computer* **16**, 22–34 (1983).
13. Leong, F.J., Brady, M. & McGee, J.O. *J. Clin. Pathol.* **56**, 619–621 (2003).
14. Babaloukas, G., Tentolouris, N., Liatis, S., Sklavounou, A. & Perrea, D. *J. Microsc.* **244**, 320–324 (2011).
15. Olsen, D. *et al. Remote Sens.* **2**, 464–477 (2010).
16. Guizar-Sicairos, M., Thurman, S.T. & Fienup, J.R. *Opt. Lett.* **33**, 156–158 (2008).
17. Liu, D.C. & Nocedal, J. *Math. Program.* **45**, 503–528 (1989).

doi:10.1038/nmeth.3323

Pattern Recognition 42 (2009) 699 – 709

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A higher-order active contour model of a 'gas of circles' and its application to tree crown extraction

P. Horváth[a,b], I.H. Jermyn[a,*], Z. Kato[b], J. Zerubia[a]

[a]Ariana (joint INRIA/I3S research group), INRIA, B.P. 93, 06902 Sophia Antipolis, France
[b]Image Processing and Computer Graphics Department, P.O. Box 652, University of Szeged, 6701 Szeged, Hungary

ARTICLE INFO

Article history:
Received 22 November 2006
Received in revised form 3 September 2008
Accepted 5 September 2008

Keywords:
Tree
Crown
Extraction
Aerial image
Higher-order
Active contour
Gas of circles
Prior
Shape

ABSTRACT

We present a model of a 'gas of circles': regions in the image domain composed of a unknown number of circles of approximately the same radius. The model has applications to medical, biological, nanotechnological, and remote sensing imaging. The model is constructed using higher-order active contours (HOACs) in order to include non-trivial prior knowledge about region shape without constraining topology. The main theoretical contribution is an analysis of the local minima of the HOAC energy that allows us to guarantee stable circles, fix one of the model parameters, and constrain the rest. We apply the model to tree crown extraction from aerial images of plantations. Numerical experiments both confirm the theoretical analysis and show the empirical importance of the prior shape information.

1. Introduction

Forestry is a domain in which image processing and computer vision techniques can have a significant impact. Resource management and conservation require information about the current state of a forest or plantation. Much of this information can be summarized in statistics related to the size and placement of individual tree crowns (e.g. mean crown area and diameter, density of the trees). Currently, this information is gathered using expensive field surveys and time-consuming semi-automatic procedures, with the result that partial information from a number of chosen sites frequently has to be extrapolated. An image processing method capable of automatically extracting tree crowns from high resolution aerial or satellite images and computing statistics based on the results would greatly aid this domain.

The tree crown extraction problem can be viewed as a special case of a general image understanding problem: the identification of the region R in the image domain Ω corresponding to some entity or entities in the scene. In order to solve this problem in any particular case, we have to construct, even if only implicitly, a probability distribution on the space of regions $\mathrm{P}(R|I,K)$. This distribution depends on the current image data I and on any prior knowledge K we may have about the region or about its relation to the image data, as encoded in the likelihood $\mathrm{P}(I|R,K)$ and the prior $\mathrm{P}(R|K)$ appearing in the Bayes' decomposition of $\mathrm{P}(R|I,K)$ (or equivalently in their energies $-\ln \mathrm{P}(I|R,K)$ and $-\ln \mathrm{P}(R|K)$). This probability distribution can then be used to make estimates of the region we are looking for.

In the automatic solution of realistic problems, the prior knowledge K, and in particular prior knowledge about the 'shape' of the region, as described by $\mathrm{P}(R|K)$, is critical. The tree crown extraction problem provides a good example: particularly in plantations, R takes the form of a collection of approximately circular connected components of similar size. There is thus a great deal of prior knowledge about the region sought. The question is then how to incorporate such prior knowledge into a model for R. If the model does not include enough prior knowledge, it will be necessary for the user to provide it.

The simplest prior information concerns the smoothness of the region boundary. For example, the Ising model and many active contour models [1–3] use a combination of region boundary length and region area as their prior energies, but curvature can be used too [1]. Such models are integrals over the region boundary of a function of various derivatives of the boundary. In consequence, they

[*] Corresponding author. Tel.: +33 4 9238 7683; fax: +33 4 9238 7643.
E-mail addresses: Peter.Horvath@sophia.inria.fr (P. Horváth), Ian.Jermyn@sophia.inria.fr (I.H. Jermyn), kato@inf.u-szeged.hu (Z. Kato), Josiane.Zerubia@sophia.inria.fr (J. Zerubia)
URLs: http://www-sop.inria.fr/ariana (P. Horváth), http://www-sop.inria.fr/ariana (I.H. Jermyn), http://www.inf.u-szeged.hu/~kato/ (Z. Kato), http://www-sop.inria.fr/ariana (J. Zerubia).

0031-3203/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.patcog.2008.09.008

capture local differential geometric information, corresponding to local interactions between boundary points, but can say nothing more global about the shape of the region. To go further, one must introduce longer range interactions. There are two principal ways to do this: one is to introduce hidden variables, given which the original variables of interest are (more or less) independent. Marginalizing over the hidden variables then introduces interactions between the original variables. Another is to include explicit long-range interactions between the original variables.

The first approach has been much investigated, in the form of template shapes and their deformations. Here a probability distribution or an energy is defined based on a distance measure of some kind between regions. One region, the template, is fixed, while the other is the variable R. Template regions may be learned from examples [4–9] or fixed by hand [10]; similarly the distance function maybe based, for example, on the learned covariance of a Gaussian distribution [5–9], or chosen *a priori* [4,10,11]. The most sophisticated methods use the kernel trick to define the distance as a pull-back from a high-dimensional space, thereby allowing more complex behaviours [12]. Multiple templates may also be used, corresponding to a mixture model [12,13].

These methods assign high probability to regions 'close' to certain points in the space of regions. The set of regions with high probability is thus in some sense bounded. As such, it is difficult to construct models of this type that favour regions for which the topology, and in particular the number of connected components, is unknown *a priori*, because the set of regions in this case is unbounded, and cannot be described as variations around one or more templates. There are many problems, however, for which the topology is unknown *a priori*, for example, the extraction of networks, or the extraction of an unknown number of objects of a particular type from astronomical, biological, medical, or remote sensing images. For this type of prior knowledge, a different type of model is needed. Higher-order active contours (HOACs) are one such category of models.

HOACs [14] take the second approach mentioned above. They introduce explicit long-range interactions between region boundary points via energies that contain multiple integrals over the boundary, thus avoiding the use of template shapes. HOAC energies can be made intrinsically Euclidean invariant, and, as required by the above analysis, incorporate sophisticated prior information about region shape without necessarily constraining region topology. As with other methods incorporating significant prior knowledge, it is not necessary to introduce extra knowledge via an initialization close to the target region: a generic initialization suffices, thus rendering the method quasi-automatic. Rochery et al. [14] applied the method to road extraction from satellite and aerial images using a prior which favours network-like objects.

In this paper, we describe a HOAC model of a 'gas of circles': the model favours regions composed of an *a priori* unknown number of circles of a certain radius. For such a model to work, the circles must be stable to small perturbations of their boundaries, i.e. they must be local minima of the HOAC energy, for otherwise a circle would tend to 'decay' into other shapes. The main theoretical contribution of this paper is an analysis of the stability of local minima of the HOAC energy that allows us to ensure that circles of a given radius are stable. In addition, it allows us to fix one of the model parameters in terms of the others, and to constrain the rest. This type of calculation has wide applicability to other active contour models and to other shapes. For example, it shows that no stable circle is possible using a classical active contour model containing only boundary length and interior area terms. The calculation proceeds by performing a functional Taylor expansion of the HOAC energy around a circle (or more generally, any shape), and then demanding that the first order term be zero for all perturbations, and that the second order term be positive semi-definite. Gradient descent experiments using the HOAC energy, with parameters fixed using the stability calculations, produce stable circles of the expected radii, thereby demonstrating empirically the coherence between the stability calculations and the numerical computations used in practice to minimize the energy.

The model has many potential applications, to medical, biological, physical, and remote sensing imagery in which the entities to be identified are circular. We choose to apply it to the problem of extracting tree crowns from aerial imagery, using the 'gas of circles' model as a prior energy, and an appropriate likelihood. We will see that the extra prior knowledge included in the 'gas of circles' model permits the separation of trees that cannot be separated by simpler methods, such as maximum likelihood or classical active contours. We focus on images of plantations and orchards, for which the model is well adapted. The case of general forests is much harder, and will be left for future work.

In the next section, we present a brief introduction to HOACs. In Section 3, we describe the 'gas of circles' HOAC model, the stability analysis, and the results of geometric experiments. In Section 4, we apply the new model to tree crown extraction. We describe a likelihood energy for trees, and then present experimental results on synthetic data and on aerial images. We conclude in Section 5, and discuss some open issues with the model.

2. Higher-order active contours

HOAC models, like all active contour models, represent a region R by its boundary, ∂R, a closed 1-chain γ in the image domain Ω ([15] is a useful reference for the following discussion). Although region boundaries correspond to a special subset of closed 1-chains known as domains of integration, active contour energies themselves are defined for general 1-chains. It is convenient to use this more general context to distinguish HOAC energies from classical active contours, because it allows for notions of linearity to be used to characterize the complexity of energy functionals.

Using this representation, HOAC energies can be defined as follows [14]. Let γ be a 1-chain in Ω, and $\operatorname{dom}\gamma$ be its domain. Then $\gamma^n : (\operatorname{dom}\gamma)^n \to \Omega^n$ is an n-chain in Ω^n. We define a class of $(n-p)$-forms on Ω^n that are 1-forms with respect to $(n-p)$ factors and 0-forms with respect to the remaining p factors (by symmetry, it does not matter which p factors). These forms can be pulled back to $(\operatorname{dom}\gamma)^n$ by γ^n. The Hodge duals of the p 0-form factors with respect to the induced metric on $\operatorname{dom}\gamma$ can then be taken independently on each such factor, thus converting them to 1-forms, and rendering the whole form an n-form on $(\operatorname{dom}\gamma)^n$. This n-form can then be integrated on $(\operatorname{dom}\gamma)^n$.

In the $(n,p)=(n,0)$ cases, we are simply integrating a general n-form on the image of γ^n in Ω^n, thus defining a linear functional on the space of n-chains in Ω^n, and hence an nth-order monomial on the space of 1-chains in Ω. Taking arbitrary linear combinations of such monomials then gives the space of polynomial functionals on the space of 1-chains. By analogy we refer to the general (n,p) cases as 'generalized nth-order monomials' on the space of 1-chains in Ω, and to arbitrary linear combinations of the latter as 'generalized polynomial functionals' on the space of 1-chains in Ω. HOAC energies are generalized polynomial functionals. Standard active contour energies are generalized *linear* functionals on 1-chains in this sense, hence the term 'higher-order'.

The (1, 1) case is simply the boundary length in some metric. The (1, 0) case gives the region area in some metric. An interesting application of the (2, 2) case to topology preservation is described by Sundaramoorthi [16]. We specialize to the (2, 0) case. Let F be a 2-form on Ω^n. Using the antisymmetry of F together with the symmetry of γ^2, we can write the energy functional in this case as

$$E(\gamma) = \int_{(\partial R)^2} F = \int_{(\operatorname{dom}\gamma)^2} (\gamma \times \gamma)^* F$$
$$= \iint_{(\operatorname{dom}\gamma)^2} \mathrm{d}t\, \mathrm{d}t'\ \tau(t) \cdot F(\gamma(t), \gamma(t')) \cdot \tau(t'), \tag{2.1}$$

where $F(x, x')$, for each $(x, x') \in \Omega^2$, is a 2×2 matrix, t is a coordinate on $\operatorname{dom}\gamma$, and $\tau = \dot{\gamma}$ is the tangent vector to γ.

By imposing Euclidean invariance on this term, and adding linear terms, Rochery et al. [14] defined the following higher-order active contour prior:

$$E\mathrm{g}(\gamma) = \lambda_C L(\gamma) + \alpha_C A(\gamma) - \frac{\beta_C}{2} \iint \mathrm{d}t\, \mathrm{d}t'\ \tau(t') \cdot \tau(t)\ \Phi(R(t, t')), \tag{2.2}$$

where L is the boundary length functional, A is the interior area functional and $R(t, t') = |\gamma(t) - \gamma(t')|$ is the Euclidean distance between $\gamma(t)$ and $\gamma(t')$. Rochery et al. [14] used the following interaction function Φ:

$$\Phi(z) = \begin{cases} \frac{1}{2}\left(1 - \frac{z-d}{\varepsilon} - \frac{1}{\pi} \sin \frac{\pi(z-d)}{\varepsilon}\right) & |z - d| < \varepsilon, \\ H(d - z) & \text{else.} \end{cases} \tag{2.3}$$

In this paper, we use this same interaction function with $d = \varepsilon$, but other monotonically decreasing functions lead to qualitatively similar results.

3. The 'gas of circles' model

For certain ranges of the parameters involved, the energy in equation (2.2) favours regions in the form of networks, consisting of long narrow arms with approximately parallel sides, joined together at junctions, as described by Rochery et al. [14]. It thus provides a good prior for network extraction from images. This behaviour does not persist for all parameter values, however, and we will exploit this parameter dependence to create a model for a 'gas of circles', an energy that favours regions composed of an *a priori* unknown number of circles of a certain radius.

For this to work, a circle of the given radius must be stable, that is, it must be a local minimum of the energy. In Section 3.1, we show that stable circles are indeed possible provided certain constraints are placed on the parameters. More specifically, we expand the energy E_g in a functional Taylor series to second order around a circle of radius r_0. The constraint that the circle be an energy extremum then requires that the first order term be zero, while the constraint that it be a minimum requires that the operator in the second order term be positive semi-definite. These requirements constrain the parameter values. In Section 3.2, we present numerical experiments using E_g that confirm the results of this analysis.

3.1. Stability analysis

We denote a member of the equivalence class of maps representing the 1-chain defining the circle by γ_0, and a small perturbation by $\delta\gamma$. To second order,

$$E\mathrm{g}(\gamma) = E\mathrm{g}(\gamma_0 + \delta\gamma) \simeq E\mathrm{g}(\gamma_0) + \left\langle \delta\gamma \middle| \frac{\delta E\mathrm{g}}{\delta\gamma} \right\rangle_{\gamma_0} + \frac{1}{2} \left\langle \delta\gamma \middle| \frac{\delta^2 E\mathrm{g}}{\delta\gamma^2} \middle| \delta\gamma \right\rangle_{\gamma_0}, \tag{3.1}$$

where $\langle\cdot|\cdot\rangle$ is a metric on the space of 1-chains.

Since γ_0 represents a circle, it is easiest to express it in terms of polar coordinates r, θ on Ω. For a suitable choice of coordinate on S^1, a circle of radius r_0 centred on the origin is then given by $\gamma_0(t) = (r_0(t), \theta_0(t))$, where $r_0(t) = r_0$, $\theta(t) = t$, and $t \in [-\pi, \pi)$. We are interested in the behaviour of small perturbations $\delta\gamma = (\delta r, \delta\theta)$. Because the energy E_g is defined on 1-chains, tangential changes in γ do not affect its value. We can therefore set $\delta\theta = 0$, and concentrate on δr.

On the circle, using the arc length parameterization t, the integrands of the different terms in E_g are functions of $t - t'$ only; they are invariant to translations around the circle. In consequence, the second derivative $\delta^2 E\mathrm{g}/\delta\gamma(t)\delta\gamma(t')$ is also translation invariant, and this implies that it can be diagonalized in the Fourier basis of the tangent space at γ_0. It is thus easiest to perform the calculation by expressing δr in terms of this basis: $\delta r(t) = \sum_k a_k \mathrm{e}^{\mathrm{i} r_0 k t}$, where $k \in \{m/r_0 : m \in \mathbb{Z}\}$. Below, we simply state the resulting expansions to second order in the a_k for the three terms appearing in Eq. (2.2). Details can be found in Appendix A.

The boundary length and interior area of the region are given to second order by

$$L(\gamma) = \int_{-\pi}^{\pi} \mathrm{d}t\ |\tau(t)| \simeq 2\pi r_0 \left\{1 + \frac{a_0}{r_0} + \frac{1}{2} \sum_k k^2 |a_k|^2 \right\} \tag{3.2}$$

$$A(\gamma) = \int_{-\pi}^{\pi} d\theta \int_0^{r(\theta)} \mathrm{d}r'\ r' \simeq \pi r_0^2 + 2\pi r_0 a_0 + \pi \sum_k |a_k|^2. \tag{3.3}$$

Note that there are no stable solutions using these terms alone. For the circle to be an extremum, we require $\lambda_C 2\pi + \alpha_C 2\pi r_0 = 0$, which tells us that $\alpha_C = -\lambda_C/r_0$. The criterion for a minimum is, for each k, $\lambda_C r_0 k^2 + \alpha_C \geqslant 0$. We must have $\lambda_C > 0$ for stability at high frequencies. Substituting for α_C, the condition becomes $\lambda_C(r_0 k^2 - r_0^{-1}) \geqslant 0$. Substituting $k = m/r_0$, gives the condition $m^2 - 1 \geqslant 0$: the zero frequency perturbation is never stable.

The quadratic term can be expressed to second order as

$$\begin{aligned} \iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\ G(t, t') = 2\pi \int_{-\pi}^{\pi} \mathrm{d}p\ F_{00}(p) + 4\pi a_0 \int_{-\pi}^{\pi} \mathrm{d}p\ F_{10}(p) \\ + \sum_k 2\pi |a_k|^2 \left\{ \left[2 \int_{-\pi}^{\pi} \mathrm{d}p\ F_{20}(p) \right.\right. \\ \left. + \int_{-\pi}^{\pi} \mathrm{d}p\ \mathrm{e}^{-\mathrm{i} r_0 k p} F_{21}(p) \right] \\ - \left[2 i r_0 k \int_{-\pi}^{\pi} \mathrm{d}p\ \mathrm{e}^{-\mathrm{i} r_0 k p} F_{23}(p) \right] \\ \left. + \left[r_0^2 k^2 \int_{-\pi}^{\pi} \mathrm{d}p\ \mathrm{e}^{-\mathrm{i} r_0 k p} F_{24}(p) \right] \right\}, \end{aligned} \tag{3.4}$$

where $G(t', t') = \tau(t') \cdot \tau(t)\ \Phi(R(t, t'))$. The F_{ij} are functionals of Φ (and hence of d), and functions of r_0, as well as of p.

Combining Eqs. (3.2)–(3.4), we find, up to second order:

$$E\mathrm{g}(\gamma_0 + \delta\gamma) \simeq e_0(r_0) + a_0 e_1(r_0) + \frac{1}{2} \sum_k |a_k|^2 e_2(k, r_0), \tag{3.5}$$

where

$$\begin{aligned} e_0(r_0) &= 2\pi\lambda_C r_0 + \pi\alpha_C r_0^2 - \pi\beta_C G_{00}(r_0) \\ e_1(r_0) &= 2\pi\lambda_C + 2\pi\alpha_C r_0 - 2\pi\beta_C G_{10}(r_0) \\ e_2(k, r_0) &= 2\pi\lambda_C r_0 k^2 + 2\pi\alpha_C - 2\pi\beta_C [2G_{20}(r_0) + G_{21}(k, r_0) \\ &\quad - 2 i r_0 k G_{23}(k, r_0) + r_0^2 k^2 G_{24}(k, r_0)], \end{aligned}$$

where $G_{ij} = \int_{-\pi}^{\pi} \mathrm{d}p\ \mathrm{e}^{-\mathrm{i} r_0 (1 - \delta(j)) k p} F_{ij}(p)$. Note that there are no off-diagonal terms linking a_k and $a_{k'}$ for $k \neq k'$: the Fourier basis diagonalizes the second order term.

3.1.1. Parameter constraints

Note that a circle of any radius is always an extremum for non-zero frequency perturbations (a_k for $k \neq 0$), as these Fourier coefficients do not appear in the first order term (this is also a consequence

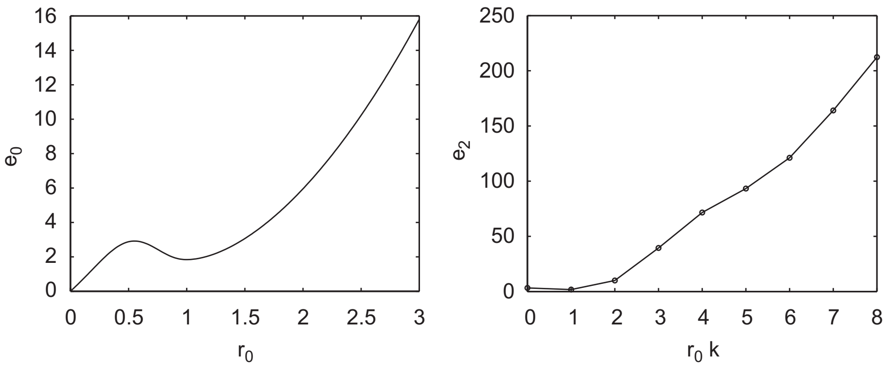


Fig. 1. Plots of e_0 against r_0 and e_2 against $\hat{r}_0 k$. Left: the energy of a circle e_0 plotted against radius r_0 for $\lambda_C = 1.0$, $\alpha = 0.8$, and $\beta_C = 1.39$ calculated from Eq. (3.6) with $\hat{r}_0 = 1.0$. (The parameters of Φ are $d = 1.0$ and $\varepsilon = 1.0$, but note that it is not necessary in general that $d = \hat{r}_0$.) The function has a minimum at $r_0 = \hat{r}_0$ as desired. Right: the second derivative of E_g, e_2, plotted against $\hat{r}_0 k$ for the same parameter values. The function is non-negative for all frequencies.

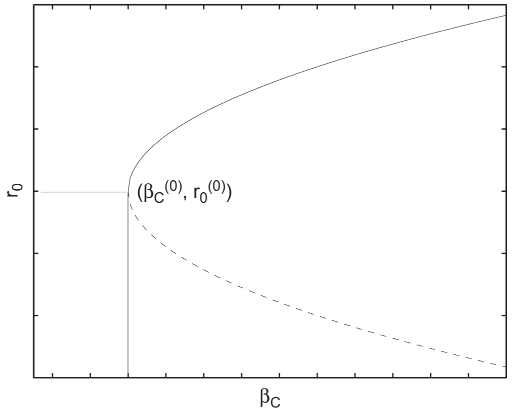


Fig. 2. Schematic plot of the positions of the extrema of the energy of a circle versus β_C.

of invariance to translations around the circle). The condition that a circle be an extremum for a_0 as well ($e_1 = 0$) gives rise to a relation between the parameters:

$$\beta_C(\lambda_C, \alpha_C, \hat{r}_0) = \frac{\lambda_C + \alpha_C \hat{r}_0}{G_{10}(\hat{r}_0)}, \tag{3.6}$$

where we have introduced $\hat{r}_0$ to indicate the radius at which there is an extremum, to distinguish it from r_0, the radius of the circle about which we are calculating the expansion (3.1). The left-hand side of Fig. 1 shows a typical plot of the energy e_0 of a circle versus its radius r_0, with the β_C parameter fixed using Eq. (3.6) with $\lambda_C = 1.0$, $\alpha = 0.8$, and $\hat{r}_0 = 1.0$. The energy has a minimum at $r_0 = \hat{r}_0$ as desired. The relationship between $\hat{r}_0$ and β_C is not quite as straightforward as it might seem though. As can be seen, the energy also has a maximum at some radius. It is not *a priori* clear whether it will be the maximum or the minimum that appears at $\hat{r}_0$. If we graph the positions of the extrema of the energy of a circle against β_C for fixed α_C, we find a curve qualitatively similar to that shown in Fig. 2 (this is an example of a fold catastrophe). The solid curve represents the minimum, the dashed the maximum. Note that there is indeed a unique β_C for a given choice of $\hat{r}_0$. Denote the point at the bottom of the curve by $(\beta_C^{(0)}, \hat{r}_0^{(0)})$. Note that at $\beta_C = \beta_C^{(0)}$, the extrema merge and for $\beta_C < \beta_C^{(0)}$, there are no extrema: the energy curve is monotonic because the quadratic term is not strong enough to overcome the shrinking effect of the length and area terms. Note also that the minimum cannot move below $r_0 = r_0^{(0)}$. This behaviour is easily understood qualitatively in terms of the interaction function in Eq. (2.3). If $2r_0 < d - \varepsilon$, the quadratic term will be constant, and no force will exist to stabilize the circle. In order to use Eq. (3.6) then, we have to ensure that we are on the upper branch of Fig. 2.

Eq. (3.6) gives the value of β_C that provides an extremum of e_0 with respect to changes of radius a_0 at a given $\hat{r}_0$ ($e_1(\hat{r}_0) = 0$), but we still need to check that the circle of radius $\hat{r}_0$ is indeed stable to perturbations with non-zero frequency, i.e. that $e_2(k, \hat{r}_0)$ is non-negative for all k. Scaling arguments mean that in fact the sign of e_2 depends only on the combinations $\tilde{r}_0 = r_0/d$ and $\tilde{\alpha}_C = (d/\lambda_C)\alpha_C$. The equation for e_2 can then be used to obtain bounds on $\tilde{\alpha}_C$ in terms of $\tilde{r}_0$. (Details of these calculations and bounds can be found in [17].) The right-hand side of Fig. 1 shows a plot of $e_2(k, \hat{r}_0)$ against $\hat{r}_0 k$ for the same parameter values used for the left-hand side, showing that it is non-negative for all $\hat{r}_0 k$.

We call the resulting model, the energy E_g with parameters chosen according to the above criteria, the 'gas of circles' model.

3.2. Geometric experiments

To illustrate the behaviour of the 'gas of circles' model, in this section we show the results of some experiments using E_g (there are no image terms). Fig. 3 shows the result of gradient descent using E_g starting from various different initial regions. (For details of the implementation of gradient descent for higher-order active contour energies using level set methods, see [14].) In the first column, four different initial regions are shown. The other three columns show the final regions, at convergence, for three different sets of parameters. In particular, the three columns have $\hat{r}_0 = 15.0$, 10.0, and 5.0, respectively.

In the first row, the initial shape is a circle of radius 32 pixels. The stable states, which can be seen in the other three columns, are circles with the desired radii in every case. In the second row, the initial region is composed of four circles of different radii. Depending on the value of $\hat{r}_0$, some of these circles shrink and disappear. This behaviour can be explained by looking at Fig. 1. As already noted, the energy of a circle e_0 has a maximum at some radius $r_{\max}$. If an initial circle has a radius less than $r_{\max}$, it will 'slide down the energy slope' towards $r_0 = 0$, and disappear. If its radius is larger than $r_{\max}$, it will finish in the minimum, with radius $\hat{r}_0$. This is precisely what is observed in this second experiment. In the third row, the initial condition is composed of four squares. The squares evolve to circles of the appropriate radii. The fourth row has an initial condition composed of four differing shapes. The nature of the stable states depends on the relation between the stable radius, $\hat{r}_0$, and the size of the initial shapes. If $\hat{r}_0$ is much smaller than an initial shape, this shape will 'decay' into several circles of radius $\hat{r}_0$.

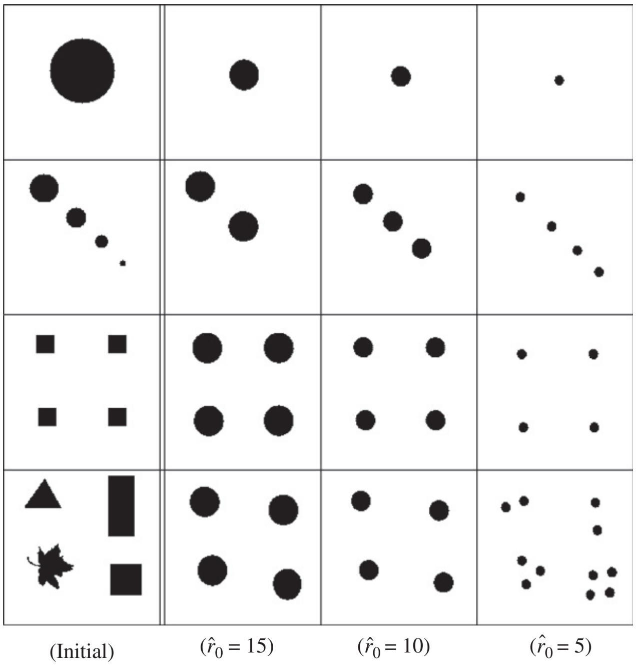


Fig. 3. Experimental results using the geometric term: the first column shows the initial conditions; the other columns show the stable states for various choices of the radius.

4. Likelihood energy and experiments

In this section, we apply the 'gas of circles' model to the extraction of trees from aerial images. We give a brief state of the art for tree crown extraction, and then present the likelihood energy we use in Section 4.2. In Section 4.3, we describe tree crown extraction experiments on aerial images and compare the results to those found using a classical active contour model. In Section 4.4, we examine the robustness of the method to noise using synthetic images. This illuminates the principal failure modes of the model, which will be further discussed in Section 5, and which point the way for future work. In Section 4.5, we illustrate the importance of prior information via tree crown separation experiments on synthetic images, and compare the results to those obtained using a classical active contour model.

4.1. Previous work

The problem of locating, counting, or delineating individual trees in high resolution aerial images has been studied in a number of papers. For example, Gougeon [18] observes that trees are brighter than the areas separating them. Local minima of the image are found using a 3 × 3 filter, and the 'valleys' connecting them are then found using a 5 × 5 filter. The tree crowns are subsequently delineated using a five-level rule-based method designed to find circular shapes, but with some small variations permitted. While the method is quite effective in separating trees, the size of the filters results in significant overestimation of the size of the trees. Larsen [19] concentrates on spruce tree detection using a template matching method. The 3D shape of the tree is modelled using a generalized ellipsoid, while illumination is modelled using the position of the sun and a clear-sky model. Template matching is used to calculate a correlation measure between the tree image predicted by the model and the image data. The local maxima of this measure are treated as tree candidates, and various strategies are then used to eliminate false positives. This method provides 3D information about the trees, but requires specific models for each species of tree, as well as knowledge of a number of extraneous parameters, for example, illumination. Brandtberg and Walter [20] decompose an image into multiple scales, and then define tree crown boundary candidates at each scale as zero crossings with convex grey-scale curvature. Edge segment centres of curvature are then used to construct a candidate tree crown region at each scale. These are then combined over different scales and a final tree crown region is grown.

The above methods use a series of *ad hoc* steps rather than a single unified model, which makes identifying the assumptions behind the methods difficult. Closer in spirit to the present work is that of [21], which models the collection of tree crowns by a marked point process, where the marks are circles or ellipses. An energy is defined that penalizes, for example, overlapping shapes, and controls the parameters of the individual shapes. Compared to the work described in this paper, the method has the advantage that overlapping trees can be represented as two separate objects, but the disadvantage that the tree crowns are not precisely delineated due to the small number of degrees of freedom for each mark.

4.2. Likelihood energy and gradient descent

In order to couple the region model E_{g} to image data, we need a likelihood, $\mathrm{P}(I|R,K)$. The images we use for the experiments are coloured infrared (CIR) images. Originally they are composed of three bands, corresponding roughly to green, red, and near infrared (NIR). Analysis of the one-point statistics of the image in the region corresponding to trees and the image in the background, shows that the 'colour' information does not add a great deal of discriminating power compared to a 'greyscale' combination of the three bands, or indeed the NIR band on its own. We therefore model the latter.

The image resolution is ~ 0.5 m/pixel, and tree crowns have diameters of the order of 10 pixels. Little dependence remains between the pixels at this resolution, which means, when combined with the paucity of statistics within each tree crown, that pixel dependencies (i.e. texture) are very hard to use for modelling purposes. We therefore model the interior of tree crowns using a Gaussian distribution with mean μ and covariance $\sigma^2\delta_R$, where δ_A is the identity operator on images on $A \subset \Omega$.

The background is very varied, and thus hard to model in a precise way. We use a Gaussian distribution with mean $\bar{\mu}$ and variance $\bar{\sigma}^2\delta_{\bar{R}}$. In general, $\mu > \bar{\mu}$, and $\sigma < \bar{\sigma}$; trees are brighter and more constant in intensity than the background. The boundary of each tree crown has significant inward-pointing image gradient, and although the Gaussian models should in principle take care of this, we have found in practice that it is useful to add a gradient term to the likelihood energy. Our likelihood thus has three factors:

$$\mathrm{P}(I|R,K) = Z^{-1}\, g_R(I_R)\, g_{\bar{R}}\,(I_{\bar{R}})\, f_{\partial R}(I_{\partial R}),$$

where I_R and $I_{\bar{R}}$ are the images restricted to R and $\bar{R}$, respectively, and g_R and $g_{\bar{R}}$ are proportional to the Gaussian distributions already described, i.e.

$$-\ln g_R(I_R) = \int_R d^2x\, \frac{1}{2\sigma^2}(I_R(x) - \mu)^2 \tag{4.1}$$

and similarly for $g_{\bar{R}}$. The function $f_{\partial R}$ depends on the gradient of the image ∂I on the boundary ∂R:

$$-\ln f_{\partial R}(I_{\partial R}) = \lambda_i \int_{\mathrm{dom}\,\gamma} \mathrm{d}t\, n(t) \cdot \partial I(t), \tag{4.2}$$

where n is the unnormalized outward normal to γ. The normalization constant Z is thus a function of μ, σ, $\bar{\mu}$, $\bar{\sigma}$, and λ_i. Z is also a functional

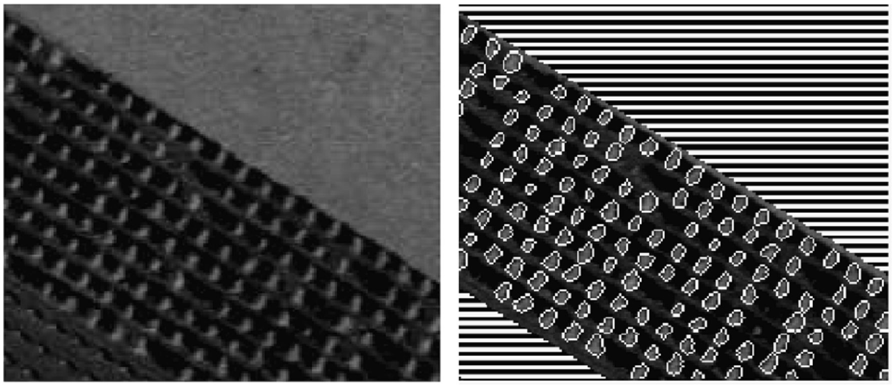


Fig. 4. Left: real image with a planted forest ©IFN (0.3, 0.06, 0.05, 0.05). Right: the result obtained using the 'gas of circles' model (529, 5.88, 5.88, 5.64, 4, 4).

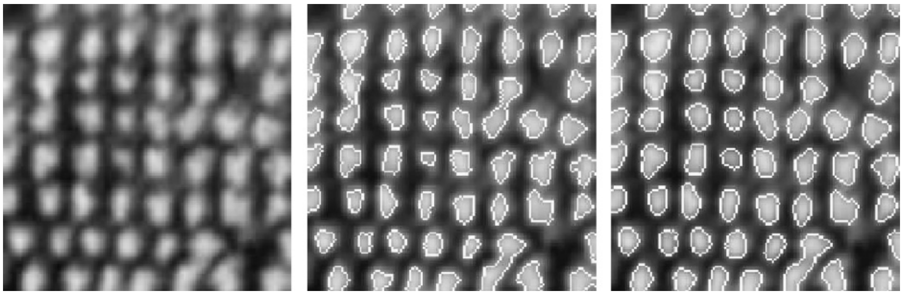


Fig. 5. From left to right: image of poplars ©IFN (0.73, 0.11, 0.23, 0.094); the best result with a classical active contour (880, 13, 73); result with the 'gas of circles' model (100, 6.7, 39, 31, 4.2, 4.2).

of the region R. To a first approximation, it is a linear combination of $L(\partial R)$ and $A(R)$. It thus has the effect of changing the parameters λ_C and α_C in E_{g}. However, since these parameters are essentially fixed by hand (the criteria described in Section 3.1.1 only allow us to fix β_C and constrain α_C), knowledge of the normalization constant does not change their values, and we ignore it once the likelihood parameters have been learnt.

The full model is then given by $E(R)=E_{\mathrm{i}}(I,R)+E_{\mathrm{g}}(R)$, where

$$E_{\mathrm{i}}(I,R)=-\ln g_R(I_R)-\ln g_{\bar{R}}(I_{\bar{R}})-\ln f_{\partial R}(I_{\partial R}).$$

The energy is minimized by gradient descent. The functional derivatives of all terms except the quadratic term in E_{g} are standard. The functional derivative of the quadratic term gives rise to a gradient descent force given by

$$\hat{n}\cdot\frac{\partial\gamma}{\partial\tau}(t)=\beta\int_{\mathrm{dom}\,\gamma}\mathrm{d}t'\,\hat{R}(t,t')\cdot n(t')\dot{\Phi}(R(t,t')), \tag{4.3}$$

where $\hat{R}(t,t')=(\gamma(t)-\gamma(t'))/|\gamma(t)-\gamma(t')|$. To evolve the region we use the level set framework of [22] extended to the demands of nonlocal forces such as Eq. (4.3) [14].

The computational complexity of the algorithm is unknown without a bound on the number of iterations. However, the complexity of one iteration is easily analysed. Eq. (4.3) involves an integration over the contour for each contour point. The worst case complexity is thus $O(L^2)$, where L is the length of the contour. The implementation however, only integrates over those points within interaction range (i.e. $d+\varepsilon$), and so the complexity depends on the average length l of contour within interaction range of a point, becoming $O(Ll)$. Typically l is a local quantity that does not depend on the size of the image. In our application, L, on the other hand, is proportional to the number of trees, which is in turn proportional to the size of the image, n. So the complexity of one iteration is $O(n)$.

4.3. *Tree crown extraction from aerial images*

In this section, we present the results of the application of the above model to 50 cm/pixel colour infrared aerial images of poplar stands located in the 'Saône et Loire' region in France. The images were provided by the French National Forest Inventory (IFN). As stated in Section 4.2, we model only the NIR band of these images, as adding the other two bands does not increase discriminating power. The tree crowns in the images are ~8–10 pixels in diameter, i.e. ~4–5 m.

In the experiments, we compare our model to a classical active contour model ($\beta_C=0$). The parameters μ, σ, $\bar{\mu}$, and $\bar{\sigma}$ were the same for both models, and were learned from hand-labelled examples in advance. The classical active contour prior model thus has three free parameters (λ_i, λ_C and α_C), while the 'gas of circles' model has six (λ_i, λ_C, α_C, β_C, d and r_0). We fixed r_0 based on our prior knowledge of tree crown size in the images, and d was then set equal to r_0. Once α_C and λ_C have been fixed, β_C is determined by Eq. (3.6). There are thus three effective parameters for the HOAC model. In the absence of any method to learn λ_i, α_C and λ_C, they were fixed by hand to give the best results, as with most applications of active contour models. The values of λ_i, λ_C and α_C were not the same for the classical active contour and HOAC models; they were chosen to give the best possible result for each model separately. The initial region in all experiments was a rounded rectangle slightly bigger than the image domain. The image values in the region exterior to the image domain were set to $\bar{\mu}$ to ensure that the region would shrink inwards.

Fig. 4 illustrates the first experiment. On the left are the data, showing a regularly planted poplar stand. The result is shown on the right. We applied the algorithm to the central part of the image only, for reasons that will be explained in Section 5.

Fig. 5 illustrates a second experiment. On the left are the data. The image shows a small piece of an irregularly planted poplar forest.

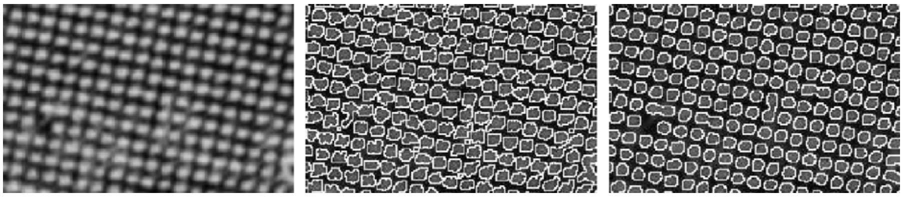


Fig. 6. From left to right: image of poplars ©IFN (0.71, 0.075, 0.18, 0.075); the best result with a classical active contour (24000, 100, 500); result with the 'gas of circles' model (1500, 25, 130, 100, 3.5, 3.5).

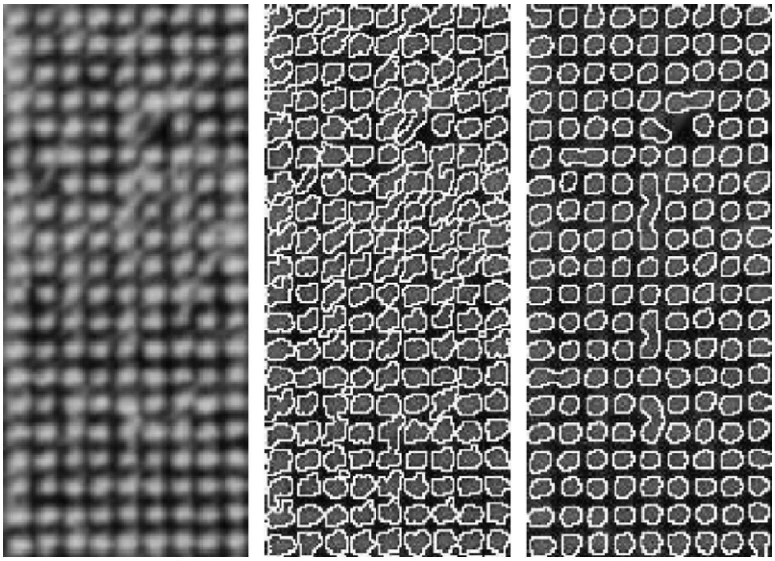


Fig. 7. From left to right: image of poplars ©IFN (0.71, 0.075, 0.18, 0.075); the best result with a classical active contour (35000, 100, 500); result with the 'gas of circles' model (1200, 20, 100, 82, 3.5, 3.5).

The image is difficult because the intensities of the crowns are varied and the gradients are blurred. In the middle is the best result we could obtain using a classical active contour. On the right is the result we obtain with the 'gas of circles' model.[1] Note that in the classical active contour result several trees that are in reality separate are merged into single connected components, and the shapes of trees are often rather distorted, whereas the prior geometric knowledge included when $\beta \neq 0$ allows the separation of almost all the trees and the regularization of their shapes.

Fig. 6 illustrates a third experiment. Again the data is on the left, the best result obtained with a classical active contour model is in the middle, and the result with the 'gas of circles' model is on the right. The trees are closer together than in the previous experiment. Using the classical active contour, the result is that the tree crown boundaries touch in the majority of cases, despite their separation in the image. Many of the connected components are malformed due to background features. The HOAC model produces more clearly delineated tree crowns, but there are still some joined trees. We will discuss this further in Section 5.

Fig. 7 shows a fourth experiment. Again the data is on the left, the best result obtained with a classical active contour model is in the middle, and the result with the 'gas of circles' model is on the right. Again, the 'gas of circles' model better delineates the tree crowns and separates more trees, but some joined trees remain also. The HOAC model selects only objects of the size chosen, so that false positives involving small objects do not occur.

Table 1
Results on real images using a classical active contour model (left) and the 'gas of circles' model (right)

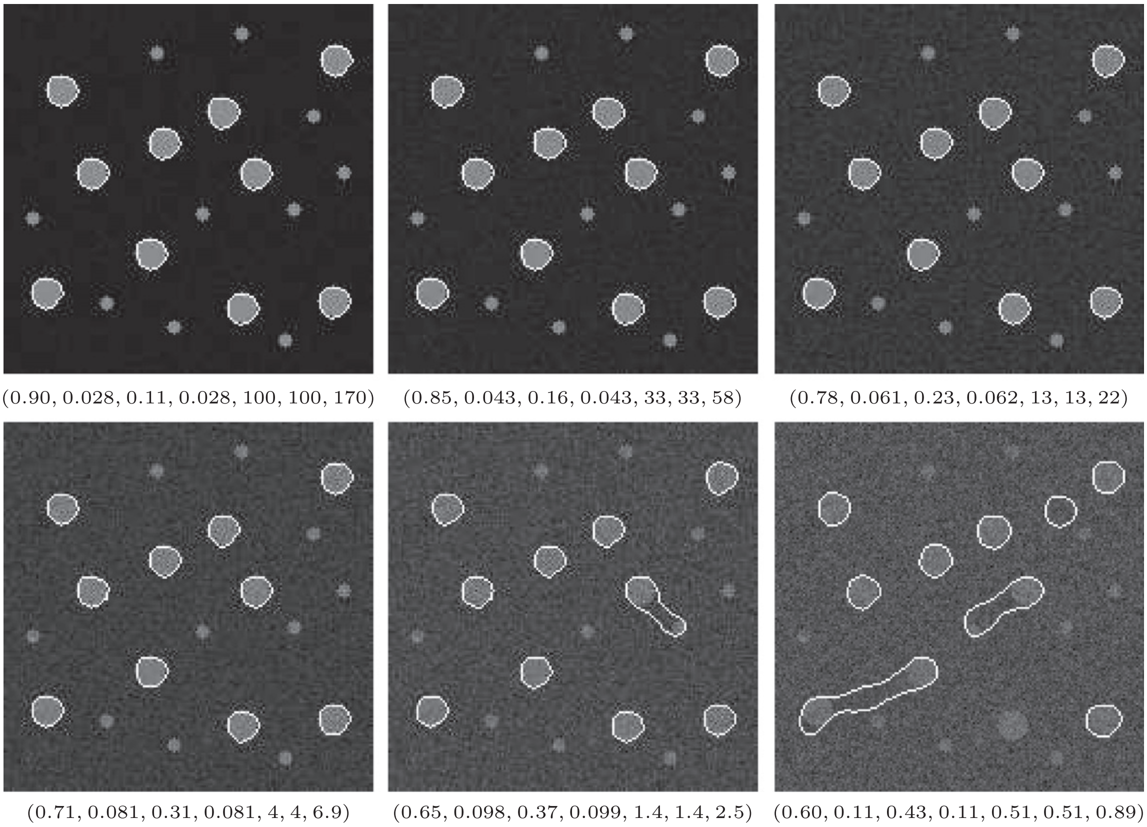
Figure	CD %	FP %	FN %	CD %	FP %	FN %
Fig. 5	85	0	15	97	0	3
Fig. 6	96.2	2.8	1.9	97.7	0	2.3
Fig. 7	89.4	5	5.6	95.5	0.6	3.9

CD: correct detections; FP: false positives; FN: false negatives (two joined trees count as one false negative).

Table 1 shows the percentages of correct tree detections, false positives and false negatives (two joined trees count as one false negative), obtained with the classical active contour model and the 'gas of circles' model in the experiments shown in Figs. 5–7. The 'gas of circles' model outperforms the classical active contour in all measures, except in the number of false negatives in the experiment in Fig. 6.

The typical runtime of the 'gas of circles' model in these experiments (image size $O(100)$ pixels) is of the order of 10 minutes on a normal personal computer. In our implementation, this is approximately ten times slower than using classical active contours.

[1] Unless otherwise specified, in the figure captions the values of the parameters learned from the image are shown when the data is mentioned, in the form $(\mu, \sigma, \bar{\mu}, \bar{\sigma})$. The other parameter values are shown when each result is mentioned, in the form $(\lambda_i, \lambda_C, \alpha_C, \beta_C, d, r_0)$, truncated if the parameters are not present. All parameter values are truncated to two significant figures. Unless otherwise specified, images were scaled to take values in $[0, 1]$. The region boundary is shown in white.



$(0.90, 0.028, 0.11, 0.028, 100, 100, 170)$ $(0.85, 0.043, 0.16, 0.043, 33, 33, 58)$ $(0.78, 0.061, 0.23, 0.062, 13, 13, 22)$

$(0.71, 0.081, 0.31, 0.081, 4, 4, 6.9)$ $(0.65, 0.098, 0.37, 0.099, 1.4, 1.4, 2.5)$ $(0.60, 0.11, 0.43, 0.11, 0.51, 0.51, 0.89)$

Fig. 8. One of the synthesized images, with six different levels of added white Gaussian noise. Reading from left to right, top to bottom, the image variance to noise power ratios are 20, 15, 10, 5, 0, −5 dB. Parameter values in the form $(\mu, \sigma, \bar{\mu}, \bar{\sigma}, \lambda_C, \alpha_C, \beta_C)$ are shown under the six images. The parameters d and r_0 were fixed to 8 throughout.

Table 2
Results on synthetic noisy images

Noise (dB)	20	15	10	5	0	−5
FP %	0	0	0	2	6.4	27.6
FN %	0	0	0	0	4	3.6
J %	0	0	0	0	0	23

FP, FN, J: percentages of false positive, false negative, and joined circle detections, respectively, with respect to the potential total number of correct detections.

4.4. Noisy synthetic images

In this section, we present the results of tests of the sensitivity of the model to noise in the image. Fifty synthetic images were created, each with ten circles with radius 8 pixels and 10 circles with radius 3.5 pixels, placed at random but with overlaps rejected. Six different levels of white Gaussian noise, with image variance to noise power ratios from −5 to 20 dB, were then added to the images to generate 300 noisy images. Six of these, corresponding to noisy versions of the same original image, were used to learn μ, σ, $\bar{\mu}$, and $\bar{\sigma}$. The model used was the same as that used for the aerial images, except that λ_i was set equal to zero. The parameters were adjusted to give a stable radius of 8 pixels.

The results obtained on the noisy versions of one of the 50 images are shown in Fig. 8. Table 2 shows the proportion of false negative and false positive circle detections with respect to the total number of potentially correctly detectable circles ($500 = 50 \times 10$), as well as the proportion of 'joined circles', when two circles are grouped together (an example can be seen in the bottom right image of Fig. 8). Detections of one of the smaller circles (which only occurred a few times even at the highest noise level) were counted as false positives. The method is very robust with respect to all but the highest levels of noise. The first errors occur at 5 dB, where there is a 2% false positive rate. At 0 dB, the error rate is $\sim$ 10%, i.e. one of the 10 circles in each image was misidentified on average. At −5 dB, the total error rate increases to $\sim$ 30%, rendering the method not very useful.

Note that the principal error modes of the model are false positives and joined circles. There are good reasons why these two types of error dominate. We will discuss them further in Section 5.

4.5. Circle separation: comparison to classical active contours

In a final experiment, we simulated one of the most important causes of error in tree crown extraction, and examined the response of classical active contour and HOAC models to this situation. The errors, which involve joined circles similar to those found in the previous experiment, are caused by the fact that in many cases nearby tree crowns in an image are connected by regions of significant intensity with significant gradient with respect to the background, thus forming a dumbbell shape. Calling the bulbous extremities, the 'bells', and the join between them, the 'bar', the situation arises when the bells are brighter than the bar, while the bar is in turn brighter than the background, and most importantly, the gradient between the background and the bar is greater than that between the bar and the bells.

The first row of Fig. 9 shows a sequence of bells connected by bars. The intensity of the bar varies along the sequence, resulting in different gradient values. We applied the classical active contour and 'gas of circles' models to these images.

The middle row of Fig. 9 shows the best results obtained using the classical active contour model. The model was either unable to

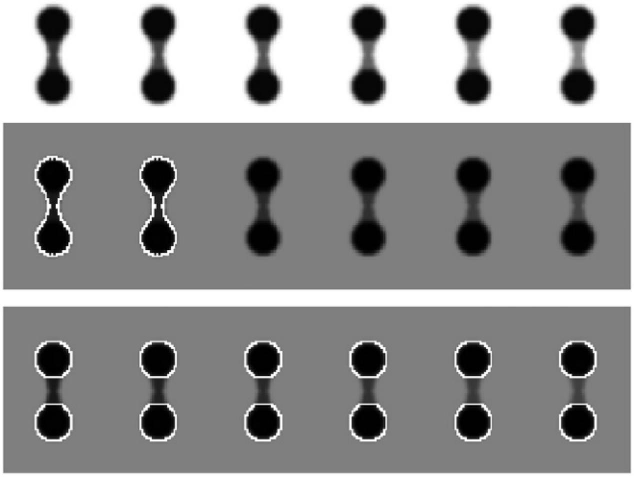


Fig. 9. Results on circle separation comparing the HOAC 'gas of circles' model to the classical active contour model. Top: original images. The intensity of the bar takes values equally spaced between 48 and 128 from left to right; the background is 255; the bells are 0. In the middle: the best results obtained using the classical active contour model (8, 1, 1). Either the circles are not separated or the region vanishes. Bottom: the results using the 'gas of circles' model (2, 1, 5, 4.0, 8, 8). All the circles are segmented correctly.

separate the individual circles, or the region completely vanished. The intuition is that if there is insufficient gradient to stop the region at the sides of the bar, then there will also be insufficient gradient to stop the region at the boundary between the bar and the bells, so that the region will vanish. On the other hand, if there is sufficient gradient between the bar and the background to stop the region, the circles will not be separated, and a 'bridge' will remain between the two circles.[2]

The corresponding results using the 'gas of circles' model are shown in the bottom row of Fig. 9. All the circles were segmented correctly, independent of the grey level of the bar. Encouraging as this is, it is not the whole story, as we indicated in Section 4.4. We make a further comment on this issue in Section 5.

5. Conclusion

Higher-order active contours allow the inclusion of sophisticated prior information in active contour models. HOACs are particularly well adapted to cases in which the topology is unknown *a priori*. In this paper, we have shown via a stability analysis that a HOAC energy can be constructed that describes a 'gas of circles', that is, it favours regions composed of an *a priori* unknown number of circles of a certain radius. The requirement that circles be stable, i.e. local minima of the energy, fixes one of the prior parameters and constrains the others.

The 'gas of circles' model has many uses in computer vision and image processing. Combined with a suitable likelihood, we have applied it to the problem of tree crown extraction from aerial images of plantations. It performs better than simpler techniques such as maximum likelihood and classical active contours. In particular, it is better able to separate trees that appear joined in the data than is a classical active contour model.

The model is not without its issues, however. First, the computation time is too long. We are currently working on a phase field HOAC [23] version of the 'gas of circles' model that we hope will significantly reduce this time. Second, there are two significant error modes, as shown in the noise experiments of Section 4.4: circles are found where the data do not ostensibly support them ('phantom circles'), and two circles may be joined into a dumbbell shape and never separated. We discuss these in turn.

The first issue is that of 'phantom' circles. Circles of radius $\hat{r}_0$ are local minima of the prior energy. It is the effect of the data that converts such configurations into global minima. Were we able to find the global minimum of the energy, this would be fine. However, gradient descent finds only a local minimum. This can create problems in areas where the data do not support the existence of circles because a circle, once formed during gradient descent, cannot disappear unless there is an image force acting on it. We thus find that circles can appear and remain even though there is no data to support them.

The second issue is that of joined circles, discussed in Section 4.5. Although the current HOAC model is better able to separate circles than a classical active contour, it still fails to do so in a number of cases, leaving a bridge between the circles. The issue here is a delicate balance between the parameters, which must be adjusted so that the sides of the bridge attract one another, thus breaking the bridge, and so that nearby circles repel one another at close range, so that the bridge does not re-form. Again, this is at least in part an algorithmic issue. Even if the two separated circles have a lower energy than the joined circles, separation may never be achieved due to a local minimum caused by the bridge.

We propose to solve the first problem via a more detailed theoretical analysis of the circle energy that will allow us to remove the local minima causing the problem, and the second via an in-depth analysis of the energy of the dumbbell configuration. Both these studies should lead to further constraints on the parameters, which is a desirable goal in itself.

Acknowledgements

This work was partially supported by EU project IMAVIS (FP5 IHP-MCHT99), EU project MUSCLE (FP6-507752), Egide PAI Balaton (09020XJ), OTKA T-046805, and a Janos Bolyai Research Fellowship of HAS. The authors would like to thank the French National Forest Inventory (IFN) for the aerial images.

Appendix A. Details of stability computations

In this appendix, we give most of the steps involved in reaching Eq. (3.5). The equation of the region boundary is

$$\gamma(t) = \gamma_0(t) + \delta\gamma(t) = (r(t), \theta(t)) = (r_0(t) + \delta r(t), \theta_0(t)), \tag{A.1}$$

where $\gamma_0(t) = (r_0(t)), \theta_0(t)) = (r_0, t)$, $\delta r(t) = \sum_k a_k \mathrm{e}^{\mathrm{i} r_0 \mathrm{kt}}$, and $k \in \{m/r_0 : m \in \mathbb{Z}\}$. The components of $\dot{\gamma}$ are

$$\dot{\theta}(t) = 1 \quad \text{and} \quad \dot{r}(t) = \dot{\delta r}(t) = \sum_k a_k i r_0 k \mathrm{e}^{\mathrm{i} r_0 \mathrm{kt}}. \tag{A.2}$$

The tangent vector field is given by

$$\tau(t) = \dot{r}(t)\partial_r + \dot{\theta}(t)\partial_\theta. \tag{A.3}$$

A.1. Linear terms

To compute the length, we need the magnitude of τ to second order. The metric in polar coordinates is $\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2$, so we have that $|\tau(t)|^2 = \dot{r}(t)^2 + r(t)^2$ by Eqs. (A.2). Substituting from Eqs. (A.1) and (A.2) gives

$$|\tau(t)|^2 = r_0^2 + 2r_0 \sum_k a_k \mathrm{e}^{\mathrm{i} r_0 \mathrm{kt}} + \sum_{k,k'} a_k a_{k'} \mathrm{e}^{\mathrm{i} r_0 (\mathrm{k}+\mathrm{k}')\mathrm{t}} (1 - r_0^2 k k'). \tag{A.4}$$

[2] 'Bar' and 'bell' refer to image properties; we use 'bridge' and 'circle' to refer to the corresponding pieces of a dumbbell-shaped region.

Taking the square root, expanding it as $\sqrt{1+x} \approx 1 + \frac{1}{2}x - \frac{1}{8}x^2$, and keeping terms to second order in the a_k then gives

$$|\tau(t)| = r_0 \left\{ 1 + \sum_k \frac{a_k}{r_0} \mathrm{e}^{\mathrm{i} r_0 \mathrm{kt}} - \frac{1}{2} \sum_{k,k'} a_k a_{k'} k k' \mathrm{e}^{\mathrm{i} r_0 (\mathrm{k}+\mathrm{k}')\mathrm{t}} \right\}. \tag{A.5}$$

Using Eq. (A.5), the boundary length is then given to second order by

$$L(\gamma) = \int_{-\pi}^{\pi} \mathrm{d}t \, |\tau(t)| = 2\pi r_0 \left\{ 1 + \frac{a_0}{r_0} + \frac{1}{2} \sum_k k^2 |a_k|^2 \right\},$$

where we have used the reality of δr to set $a_{-k} = a_k^*$, where * indicates complex conjugation, and orthonormality of the Fourier basis elements.

We can write the interior area of the region as

$$A(\gamma) = \int_{-\pi}^{\pi} \mathrm{d}\theta \int_0^{r(\theta)} \mathrm{d}r' \, r' = \int_{-\pi}^{\pi} \mathrm{d}\theta \, \frac{1}{2} r^2(\theta)$$

Thus, using Eq. (A.1), and again using orthonormality, we have that

$$A(\gamma) = \pi r_0^2 + 2\pi r_0 a_0 + \pi \sum_k |a_k|^2. \tag{A.6}$$

A.2. Quadratic terms

To compute the expansion of the quadratic term in Eq. (2.2) for E_g, we need the expansions of $\tau(t) \cdot \tau(t')$ and $\Phi(R(t,t'))$.

A.2.1. Inner product of tangent vectors

The tangent vector is given by Eq. (A.3), but we must take care as $\tau(t)$ and $\tau(t')$ live in different tangent spaces, at $\gamma(t)$ and $\gamma(t')$, respectively. It is easiest to convert the tangent vectors to the Euclidean coordinate basis, $\partial_r = \cos(\theta)\partial_x + \sin(\theta)\partial_y$ and $\partial_\theta = -r\sin(\theta)\partial_x + r\cos(\theta)\partial_y$ as these basis vectors are preserved by parallel transport. Taking the inner product then gives

$$\begin{aligned} \tau \cdot \tau' = {} & \cos(\theta' - \theta)[r_0^2 + r_0 \delta r + r_0 \delta r' + \delta r \delta r' + \dot{\delta r} \dot{\delta r}'] \\ & + \sin(\theta' - \theta)[r_0 \dot{\delta r}' - r_0 \dot{\delta r} + \delta r \dot{\delta r}' - \dot{\delta r} \delta r'], \end{aligned}$$

where unprimed quantities are evaluated at t and primed quantities at t'.

A.2.2. Interaction function

First, we expand $R(t,t')$. The squared distance between $\gamma(t')$ and $\gamma(t)$ is given by

$$\begin{aligned} |\gamma(t') - \gamma(t)|^2 = {} & [(r_0 + \delta r')\cos(\theta') - (r_0 + \delta r)\cos(\theta)]^2 \\ & + [(r_0 + \delta r')\sin(\theta') - (r_0 + \delta r)\sin(\theta)]^2, \end{aligned}$$

which after expansion gives

$$\begin{aligned} R^2(t,t') = 2r_0^2(1 - \cos(\Delta t)) \Bigg\{ & 1 + \frac{1}{r_0}(\delta r + \delta r') \\ & + \frac{\delta r^2 + \delta r'^2 - 2\cos(\Delta t)\delta r \delta r'}{2r_0^2(1 - \cos(\Delta t))} \Bigg\}, \end{aligned}$$

where $\Delta t = \theta' - \theta = t' - t$. Expanding $\sqrt{1+x} \approx 1 + \frac{1}{2}x - \frac{1}{8}x^2$ to second order and collecting terms, we then find

$$R(t,t') = 2r_0 |\sin(\Delta t/2)| + |\sin(\Delta t/2)|(\delta r + \delta r') + \frac{A(\Delta t)}{4r_0}(\delta r - \delta r')^2, \tag{A.7}$$

where $A(z) = \cos^2(z/2)|\sin(z/2)|^{-1}$.

Expanding $\Phi(z)$ in a Taylor series to second order, and then substituting $R(t,t')$ for z using the approximation in Eq. (A.7), and keeping only terms up to second order in $\delta\gamma$ then gives

$$\begin{aligned} \Phi(R(t,t')) = {} & \Phi(X_0) + \left|\sin \frac{\Delta t}{2}\right| \Phi'(X_0)(\delta r + \delta r') \\ & + \frac{1}{4r_0} A(\Delta t) \Phi'(X_0)(\delta r - \delta r')^2 \\ & + \frac{1}{2}\sin^2\left(\frac{\Delta t}{2}\right) \Phi''(X_0)(\delta r + \delta r')^2, \end{aligned} \tag{A.8}$$

where $X_0 = 2r_0 |\sin(\Delta t/2)|$.

A.3. Combining terms

Using the above equations gives

$$\begin{aligned} & G(t,,t') \\ & = \underbrace{r_0^2 \cos(\Delta t)\Phi(X_0)}_{F_{00},\text{even}} \\ & \quad + (\delta r + \delta r') \underbrace{r_0 \cos(\Delta t) \left\{ \Phi(X_0) + r_0 \left|\sin\frac{\Delta t}{2}\right| \Phi'(X_0) \right\}}_{F_{10},\text{even}} \\ & \quad + (\dot{\delta r}' - \dot{\delta r}) \underbrace{r_0 \sin(\Delta t) \Phi(X_0)}_{F_{11},\text{odd}} \\ & \quad + (\delta r^2 + \delta r'^2) \underbrace{r_0 \cos(\Delta t) \left\{ \frac{1}{4} A(\Delta t)\Phi'(X_0) + \frac{1}{2} r_0 \sin^2\left(\frac{\Delta t}{2}\right) \Phi''(X_0) + \left|\sin\frac{\Delta t}{2}\right| \Phi'(X_0) \right\}}_{F_{20},\text{even}} \\ & \quad + (\delta r \delta r') \underbrace{\cos(\Delta t) \left\{ \Phi(X_0) + 2r_0 \left|\sin\frac{\Delta t}{2}\right| \Phi'(X_0) - \frac{1}{2} r_0 A(\Delta t)\Phi'(X_0) + r_0^2 \sin^2\left(\frac{\Delta t}{2}\right) \Phi''(X_0) \right\}}_{F_{21},\text{even}} \\ & \quad + (\delta r' \dot{\delta r}' - \delta r \dot{\delta r}) \underbrace{r_0 \left|\sin\frac{\Delta t}{2}\right| \sin(\Delta t)\Phi'(X_0)}_{F_{22},\text{odd}} \\ & \quad + (\delta r \dot{\delta r}' - \delta r' \dot{\delta r}) \underbrace{\sin(\Delta t) \left\{ \Phi(X_0) + r_0 \left|\sin\frac{\Delta t}{2}\right| \Phi'(X_0) \right\}}_{F_{23},\text{odd}} \\ & \quad + (\dot{\delta r} \dot{\delta r}') \underbrace{\cos(\Delta t)\Phi(X_0)}_{F_{24},\text{even}}, \end{aligned}$$

where the F_{ij} denote the functions appearing in the terms of G, and 'odd' and 'even' refer to parity under exchange of t and t'. Each line, and hence G, is symmetric in t and t', as it should be. We can now substitute the expressions for δr and $\dot{\delta r}$ in terms of their Fourier coefficients, and calculate $\iint_{-\pi}^{\pi} \mathrm{d}t \, \mathrm{d}t' \, G(t,t')$. We note that in the terms involving F_{10}, F_{11}, F_{20}, F_{22}, and F_{23}, the presence of the symmetric or antisymmetric factors in δr and $\delta r'$ simply leads to a doubling of the value of the integral for one of the terms in these factors, due to the corresponding symmetry or antisymmetry of the F functions. We therefore only need to evaluate one of these integrals for the relevant terms.

Because the F's depend only on Δt, the resulting integrals can be reduced, via a change of variables $p = \Delta t$, to integrals over p. For F_{00} and F_{10}, we have

$$\iint_{-\pi}^{\pi} \mathrm{d}t \, \mathrm{d}t' \, F_{00}(t' - t) = \iint_{-\pi}^{\pi} \mathrm{d}p \, \mathrm{d}t' \, F_{00}(p) = 2\pi \int_{-\pi}^{\pi} \mathrm{d}p \, F_{00}(p),$$

and

$$\iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \delta r(t)\, F_{10}(t'-t) = \iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \sum_k a_k \mathrm{e}^{\mathrm{i}r_0 \mathrm{kt}}\, F_{10}(t'-t)$$

$$= \sum_k a_k \iint_{-\pi}^{\pi} \mathrm{d}p\, \mathrm{d}t'\; \mathrm{e}^{\mathrm{i}r_0 \mathrm{k}(-\mathrm{p}+\mathrm{t}')}\, F_{10}(p)$$

$$= \sum_k a_k \int_{-\pi}^{\pi} \mathrm{d}t' \mathrm{e}^{\mathrm{i}r_0 \mathrm{kt}'} \int_{-\pi}^{\pi} \mathrm{d}p \mathrm{e}^{-\mathrm{i}r_0 kp} F_{10}(p)$$

$$= \sum_k a_k 2\pi \delta(k) \int_{-\pi}^{\pi} \mathrm{d}p\; \mathrm{e}^{-\mathrm{i}r_0 kp}\, F_{10}(p)$$

$$= 2\pi a_0 \int_{-\pi}^{\pi} \mathrm{d}p\, F_{10}(p).$$

The calculations for the other terms proceed in a very similar fashion, using the same change of variable and the orthonormality of the Fourier basis. We merely list the results (full details may be found in [24]):

$$\iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \dot{\delta} r(t)\, F_{11}(t'-t) = 0$$

$$\iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \delta r^2(t)\, F_{20}(t'-t) = 2\pi \sum_k |a_k|^2 \int_{-\pi}^{\pi} \mathrm{d}p\, F_{20}(p)$$

$$\iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \delta r(t) \delta r(t')\, F_{21}(t'-t) = 2\pi \sum_k |a_k|^2 \int_{-\pi}^{\pi} \mathrm{d}p\; \mathrm{e}^{-\mathrm{i}r_0 kp}\, F_{21}(p)$$

$$\iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \delta r(t) \dot{\delta} r(t)\, F_{22}(t'-t) = 0$$

$$\iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \delta r(t) \dot{\delta} r(t') F_{23}(t'-t) = -2\pi \sum_k |a_k|^2 \mathrm{i}r_0 k \int_{-\pi}^{\pi} \mathrm{d}p\, \mathrm{e}^{-\mathrm{i}r_0 kp} F_{23}(p)$$

$$\iint_{-\pi}^{\pi} \mathrm{d}t\, \mathrm{d}t'\; \dot{\delta} r(t) \dot{\delta} r(t') F_{24}(t'-t) = 2\pi \sum_k |a_k|^2 r_0^2 k^2 \int_{-\pi}^{\pi} \mathrm{d}p\; \mathrm{e}^{-\mathrm{i}r_0 kp}\, F_{24}(p).$$

Using these results then gives Eq. (3.4), which in combination with Eqs. (3.2) and (3.3), gives Eq. (3.5).

References

[1] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, Int. J. Comput. Vision 1 (4) (1988) 321–331.

[2] L. Cohen, On active contours and balloons, CVGIP: Image Understanding 53 (1991) 211–218.

[3] V. Caselles, R. Kimmel, G. Sapiro, Geodesic active contours, Int. J. Comput. Vision 22 (1) (1997) 61–79.

[4] Y. Chen, H. Tagare, S. Thiruvenkadam, F. Huang, D. Wilson, K. Gopinath, R. Briggs, E. Geiser, Using prior shapes in geometric active contours in a variational framework, Int. J. Comput. Vision 50 (3) (2002) 315–328.

[5] D. Cremers, F. Tischhäuser, J. Weickert, C. Schnörr, Diffusion snakes: introducing statistical shape knowledge into the Mumford–Shah functional, Int. J. Comput. Vision 50 (3) (2002) 295–313.

[6] M. Leventon, W. Grimson, O. Faugeras, Statistical shape influence in geodesic active contours, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hilton Head Island, SC, USA, 2000, pp. 316–322.

[7] N. Paragios, M. Rousson, Shape priors for level set representations, in: Proceedings of the European Conference on Computer Vision (ECCV), Copenhagen, Denmark, 2002, pp. 78–92.

[8] A. Srivastava, S. Joshi, W. Mio, X. Liu, Statistical shape analysis: clustering, learning, and testing, IEEE Trans. Pattern Anal. Mach. Intell. 27 (4) (2005) 590–602.

[9] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, E. Grimson, A. Willsky, Model-based curve evolution technique for image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2001.

[10] A. Foulonneau, P. Charbonnier, F. Heitz, Geometric shape priors for region-based active contours, Proceedings of the IEEE International Conference on Image Processing (ICIP) 3 (2003) 413–416.

[11] D. Cremers, S. Soatto, A pseudo-distance for shape priors in level set segmentation, in: Proceedings of the 2nd IEEE Workshop on Variational, Geometric and Level Set Methods, Nice, France, 2003, pp. 169–176.

[12] D. Cremers, T. Kohlberger, C. Schnörr, Shape statistics in kernel space for variational image segmentation, Pattern Recognition 36 (9) (2003) 1929–1943.

[13] D. Cremers, N. Sochen, C. Schnörr, A multiphase dynamic labeling model for variational recognition-driven image segmentation, Int. J. Comput. Vision 66 (1) (2006) 67–81.

[14] M. Rochery, I.H. Jermyn, J. Zerubia, Higher-order active contours, Int. J. Comput. Vision 69 (1) (2006) 27–42.

[15] Y. Choquet-Bruhat, C. DeWitt-Morette, M. Dillard-Bleick, Analysis, Manifolds and Physics, Elsevier Science, Amsterdam, The Netherlands, 1996.

[16] G. Sundaramoorthi, A. Yezzi, More-than-topology-preserving flows for active contours and polygons, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Washington DC, USA, 2005, pp. 1276–1283.

[17] P. Horvath, I. H. Jermyn, J. Zerubia, Z. Kato, An improved model for tree crown extraction using higher order active contours, Research report, INRIA, France, November 2007, to appear.

[18] F.A. Gougeon, Automatic individual tree crown delineation using a valley-following algorithm and rule-based system, in: D. Hill, D. Leckie (Eds.), in: Proceedings of the International Forum on Automated Interpretation of High Spatial Resolution Digital Imagery for Forestry, Victoria, British Columbia, Canada, 1998, pp. 11–23.

[19] M. Larsen, Finding an optimal match window for Spruce top detection based on an optical tree model, in: D. Hill, D. Leckie (Eds.), Proceedings of the International Forum on Automated interpretation of High Spatial Resolution Digital Imagery for Forestry, Victoria, British Columbia, Canada, 1998, pp. 55–66.

[20] T. Brandtberg, F. Walter, Automated delineation of individual tree crowns in high spatial resolution aerial images by multiple-scale analysis, Mach. Vision Appl. 11 (1998) 64–73.

[21] G. Perrin, X. Descombes, J. Zerubia, A marked point process model for tree crown extraction in plantations, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), Genova, Italy, 2005.

[22] S. Osher, J.A. Sethian, Fronts propagating with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations, J. Comput. Phys. 79 (1) (1988) 12–49.

[23] M. Rochery, I.H. Jermyn, J. Zerubia, Phase field models and higher-order active contours, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Beijing, China, 2005.

[24] P. Horvath, I.H. Jermyn, Z. Kato, J. Zerubia, A higher-order active contour model of a 'gas of circles' and its application to tree crown extraction, Research Report 6026, INRIA, France, November 2006.

About the Author—PÉTER HORVÁTH received an M.S. (Computer Science) from the University of Szeged, Hungary (2004), a joint Ph.D. (Image Segmentation) from the University of Nice Sophia Antipolis, doing his research at INRIA Sophia Antipolis, France, and from the University of Szeged, Hungary (2007). He is currently a Staff Scientist at ETH Zurich. His research interests include image segmentation, motion detection, and image based classification.

About the Author—IAN JERMYN received a B.A. (Physics) from Oxford University (1986), a Ph.D. (Theoretical Physics) from Manchester University (1991), and a Ph.D. (Computer Science) from the Courant Institute, New York University (2000). He is currently a Senior Research Scientist in the Ariana group at INRIA. His research interests include shape and texture modelling.

About the Author—ZOLTAN KATO received the M.S. degree in Computer Science from the University of Szeged, Hungary in 1990, and the Ph.D. degree from the University of Nice doing his research at INRIA Sophia Antipolis, France in 1994. Currently, he is head of the Department of Image Processing and Computer Graphics at the University of Szeged. His research interests include statistical image models, MCMC methods, and shape modeling.

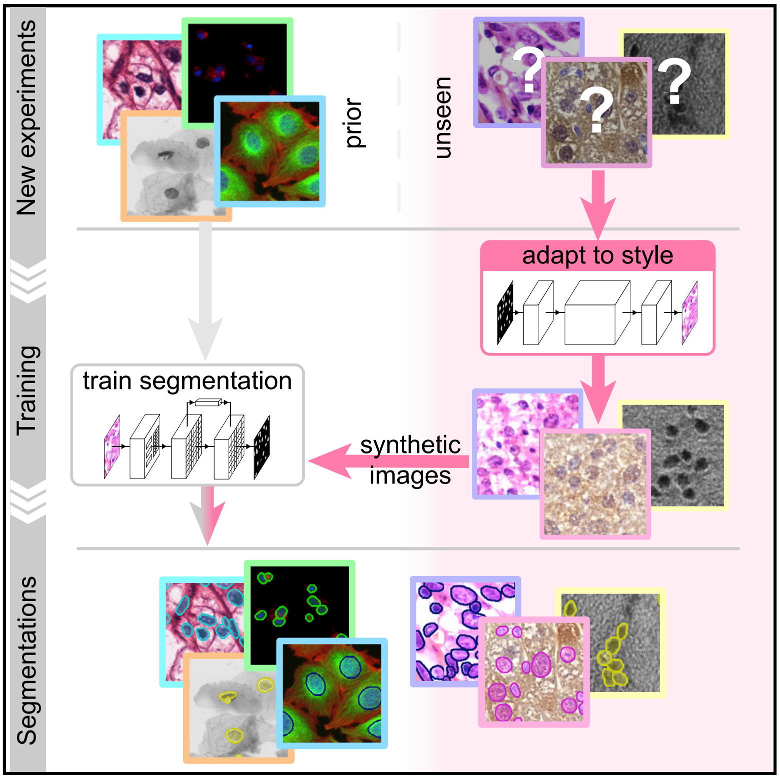
About the Author—JOSIANE ZERUBIA received an MSc (Electrical Engineering) from ENSIEG (1981), a Doctor of Engineering (1986), a Ph.D. (1988), and an 'Habilitation' (1994) all from the University of Nice Sophia Antipolis. She is currently a director of research and head of the Ariana group at INRIA. She is an IEEE Fellow. Her research interests are in image processing using probabilistic models and variational methods.

Methods in Brief

Cell Systems

nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer

Graphical Abstract



Highlights

- Robust method automatically adapting to various unseen experimental scenarios
- Deep learning solution for accurate nucleus segmentation without user interaction
- Accelerates, improves quality, and reduces complexity of bioimage analysis tasks
- Easy-to-use online tool provided for the method

Authors

Reka Hollandi, Abel Szkalisity,
Timea Toth, ...,
Anne Elizabeth Carpenter, Kevin Smith,
Peter Horvath

Correspondence

horvath.peter@brc.hu

In Brief

Microscopy image analysis of single cells can be challenging but also eased and improved. We developed a deep learning method to segment cell nuclei. Our strategy is adapting to unexpected circumstances automatically by synthesizing artificial microscopy images in such a domain as training samples.

Hollandi et al., 2020, Cell Systems *10*, 453–458
May 20, 2020 © 2020 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.04.003

Cell Systems

CellPress
OPEN ACCESS

Methods in Brief

nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer

Reka Hollandi,[1] Abel Szkalisity,[1] Timea Toth,[1,2] Ervin Tasnadi,[1,3] Csaba Molnar,[1,3] Botond Mathe,[1] Istvan Grexa,[1,4] Jozsef Molnar,[1] Arpad Balind,[1] Mate Gorbe,[1] Maria Kovacs,[1] Ede Migh,[1] Allen Goodman,[6] Tamas Balassa,[1,5] Krisztian Koos,[1] Wenyu Wang,[7] Juan Carlos Caicedo,[6] Norbert Bara,[1,8] Ferenc Kovacs,[1,8] Lassi Paavolainen,[7] Tivadar Danka,[1] Andras Kriston,[1,8] Anne Elizabeth Carpenter,[6] Kevin Smith,[9,10] and Peter Horvath[1,7,8,11,*]

[1]Synthetic and Systems Biology Unit, Hungarian Academy of Sciences, Biological Research Center (BRC), Temesvári körút 62, Szeged 6726, Hungary
[2]Doctoral School of Biology, University of Szeged, Közép fasor 52, Szeged 6726, Hungary
[3]Doctoral School of Computer Science, University of Szeged, Árpád tér 2, Szeged 6720, Hungary
[4]Doctoral School of Interdisciplinary Medicine, University of Szeged, Koranyi fasor 10, Szeged 6720, Hungary
[5]Doctoral School of Informatics, Eötvös Loránd University, Pázmány Péter sétány 1/C, Room 2.317, Budapest 1117, Hungary
[6]Imaging Platform, Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA
[7]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Tukholmankatu 8, Helsinki 00014, Finland
[8]Single-Cell Technologies Ltd, Szeged 6726, Hungary
[9]KTH Royal Institute of Technology, School of Computer Science and Communication, Lindstedtsvägen 3, Stockholm 10044, Sweden
[10]Science for Life Laboratory, Solna, Sweden
[11]Lead Contact
*Correspondence: horvath.peter@brc.hu

https://doi.org/10.1016/j.cels.2020.04.003

SUMMARY

Single-cell segmentation is typically a crucial task of image-based cellular analysis. We present nucleAIzer, a deep-learning approach aiming toward a truly general method for localizing 2D cell nuclei across a diverse range of assays and light microscopy modalities. We outperform the 739 methods submitted to the 2018 Data Science Bowl on images representing a variety of realistic conditions, some of which were not represented in the training data. The key to our approach is that during training nucleAIzer automatically adapts its nucleus-style model to unseen and unlabeled data using image style transfer to automatically generate augmented training samples. This allows the model to recognize nuclei in new and different experiments efficiently without requiring expert annotations, making deep learning for nucleus segmentation fairly simple and labor free for most biological light microscopy experiments. It can also be used online, integrated into CellProfiler and freely downloaded at www.nucleaizer.org.
A record of this paper's transparent peer review process is included in the Supplemental Information.

INTRODUCTION

Identifying nuclei is the starting point for many microscopy-based cellular analyses, which are widespread in biomedical research. Accurate localization of the nucleus is the basis of a variety of quantitative measurements of important cell functions but is also a first step for identifying individual cell borders, which enables a multitude of further analyses. Until recently, the dominant approaches for this task have been based on classic image processing algorithms (e.g., thresholding and seeded watershed; Carpenter et al., 2006), guided by shape and spatial priors (Molnar et al., 2016). These methods require expert knowledge to properly adjust the parameters, which typically must be retuned when experimental conditions change.

Recently, deep learning has revolutionized an assortment of tasks in image analysis, from image classification (Krizhevsky et al., 2017) to face recognition (Taigman et al., 2014) and scene segmentation (Badrinarayanan et al., 2017). It is also responsible for breakthroughs in diagnosing retinal images (De Fauw et al., 2018), classifying skin lesions with superhuman performance (Esteva et al., 2017), and correcting artifacts in fluorescence images (Weigert et al., 2017). Initial work (reviewed in Moen et al., 2019) indicates that deep learning is effective for nucleus segmentation (Falk et al., 2019; Van Valen et al., 2016; Cui et al., 2018); however, these methods often fail to properly separate touching nuclei well and most importantly lack robustness to unseen domains.

The 2018 Data Science Bowl (DSB) organized by Kaggle, Booz Allen Hamilton, and the Broad Institute challenged participants to push the state of the art in nucleus segmentation. The goal of the challenge was to develop fully automated and robust methods effective in a variety of conditions, including differing cell lines, treatments, and types of light microscopy. The challenge attracted thousands of data scientists from around the world. Approaches using deep learning dominated the competition, achieving scores that shattered what was previously possible: the best performing traditional methods we submitted ranked no higher than 1,000 out of 3,891 submissions in stage 1 (data not shown); even classical methods hand-tuned to five subsets of the testing data were beaten by 85 out of 739 submissions in stage 2 testing (Caicedo et al., 2019b). The top deep-learning-based methods relied on only a handful of different architectures, namely Mask R-CNN, U-Net, and feature-pyramid networks; the factors that participants commonly believed had most influence over their method's ranking were the amount of data, the pre-processing, and methods used to augment the data.

We present here a superior approach we named nucleAIzer, which, unlike the previous best submissions, applies image style transfer (Isola et al., 2017): an image-to-image translation using a pixel-wise mapping from one image to the other that ensures the generated synthetic output image resembles the original as closely as possible. It aims to overcome one of the greatest challenges of deep learning, the extent of the annotated training set. In particular, we address the unsupervised domain adaptation problem in which the target (test) samples are drawn from a different distribution than the labeled training samples, but we have access to some unlabeled samples from the target distribution. We augment the training samples by creating realistic-looking artificial sample images with the texture, coloration, and pattern elements from source images not included in the training set using image style transfer (Figure 1). Combining this with a segmentation network based on Mask R-CNN (He et al., 2017), an instance segmentation and classification network, along with boundary correction using U-Net (Ronneberger et al., 2015), a semantic segmentation network for biomedical images, (Figure S4) and mathematical morphology, our method outperforms all other methods reported on the final DSB leaderboard (post-competition) (Our method achieved the top-score after the competition ended. An early version of our approach placed 27th out of 739 submissions in round 2 of the competition). We also demonstrate that our method outperforms similar baselines on public fluorescent and histology datasets. Our trained model does not require parameter tuning or specialized knowledge to use and can be applied on a wide variety of conditions and imaging modalities.

Our software is open source and freely available (Data S1 at https://github.com/spreka/biomagdsb). Pre-trained networks for DSB, fluorescent, and histology data can be applied to new images via CellProfiler (Data S2 and at https://github.com/CellProfiler/CellProfiler-plugins/blob/master/nucleaizer.py) or through an online interface at www.nucleaizer.org.

Our approach (Figures 1A and S1; STAR Methods) begins by automatically rescaling the images such that nucleus size is approximately uniform, as the performance of the network is improved if the nucleus size is fixed during training and inference (see STAR Methods; Figures S3 and S6). To do this, we estimate the typical nucleus size in the provided images with a Mask R-CNN-based network pretrained on a large set of diverse images with nucleus segmentations and fine-tuned using the provided training data and label masks. The output of this network is an initial segmentation we use to estimate the typical nucleus size. Alternatively, if the typical nucleus size is known a priori, it can be provided manually and the images rescaled accordingly.

Next, to adapt our model to handle a wide variety of cell types, staining methods, and imaging modalities, even those for which no segmentation annotations are available, we augment the training set with an artificially generated set of representative image-label pairs. This is accomplished using image style transfer. Training and inference both begin by automatically clustering training images into similar styles based on their appearance, using *k*-means (see STAR Methods; Figure 1B). For each cluster of similar image types, a style transfer network (Isola et al., 2017) is trained to generate synthetic images of the desired style with nuclei at specified locations. During training, nucleus annotations are used to train the style transfer network; during inference on out-of-domain target images, we use nucleus masks output from the initial segmentation network. After a style transfer network is trained for each image style, we generate a set of artificial nucleus masks representative of the shape, size, and spatial distribution of nuclei belonging to that style. For this, we used ~100,000 manually labeled single nucleus masks from the DSB set. A subset of these nuclei is selected that represent the shape distribution of the original morphologies, and they are placed such that they follow the spatial distribution of the image style (see STAR Methods). With trained style networks and representative nucleus masks in hand, we generate synthetic images in the desired style nearly indistinguishable from real microscopy images (see STAR Methods) with nuclei in locations defined by the artificial masks. The synthetic image-mask pairs make up the augmented dataset; samples are shown in Figures 1B and S7A. The augmented data are added to the training data for the segmentation network and further extended with conventional augmentations (rotation, cropping, intensity stretching, etc., see STAR Methods). For this experiment, we generated 20 synthetic image/mask pairs for each of the 134 style clusters we identified in the final round data.

Finally, the ultimate Mask R-CNN segmentation model is trained on the combined augmented and rescaled training data. All images are adjusted such that the estimated nuclei size is uniform. To refine the segmentations for high pixel-level accuracy, the edges of each detected nucleus are corrected using a U-Net-based model trained on the same data, followed by some mathematical morphology-based post-processing (see STAR Methods). This step may be skipped if such accuracy is unnecessary for the application, for example, if simply counting nuclei.

RESULTS

We evaluated our approach on four different datasets: DSB stage 1, DSB stage 2, our own set of fluorescence microscopy images, and our own set of histology images from various sources (*DSB1, DSB2, fluo*, and *hist*, respectively, details in Table S2). We compare our approach against submissions from other teams on *DSB1* and *DSB2* (nearly 3,000 in stage 1 and 739 in stage 2). As benchmarks, we include the results reported in the first and second positions of the leaderboard, which was frozen

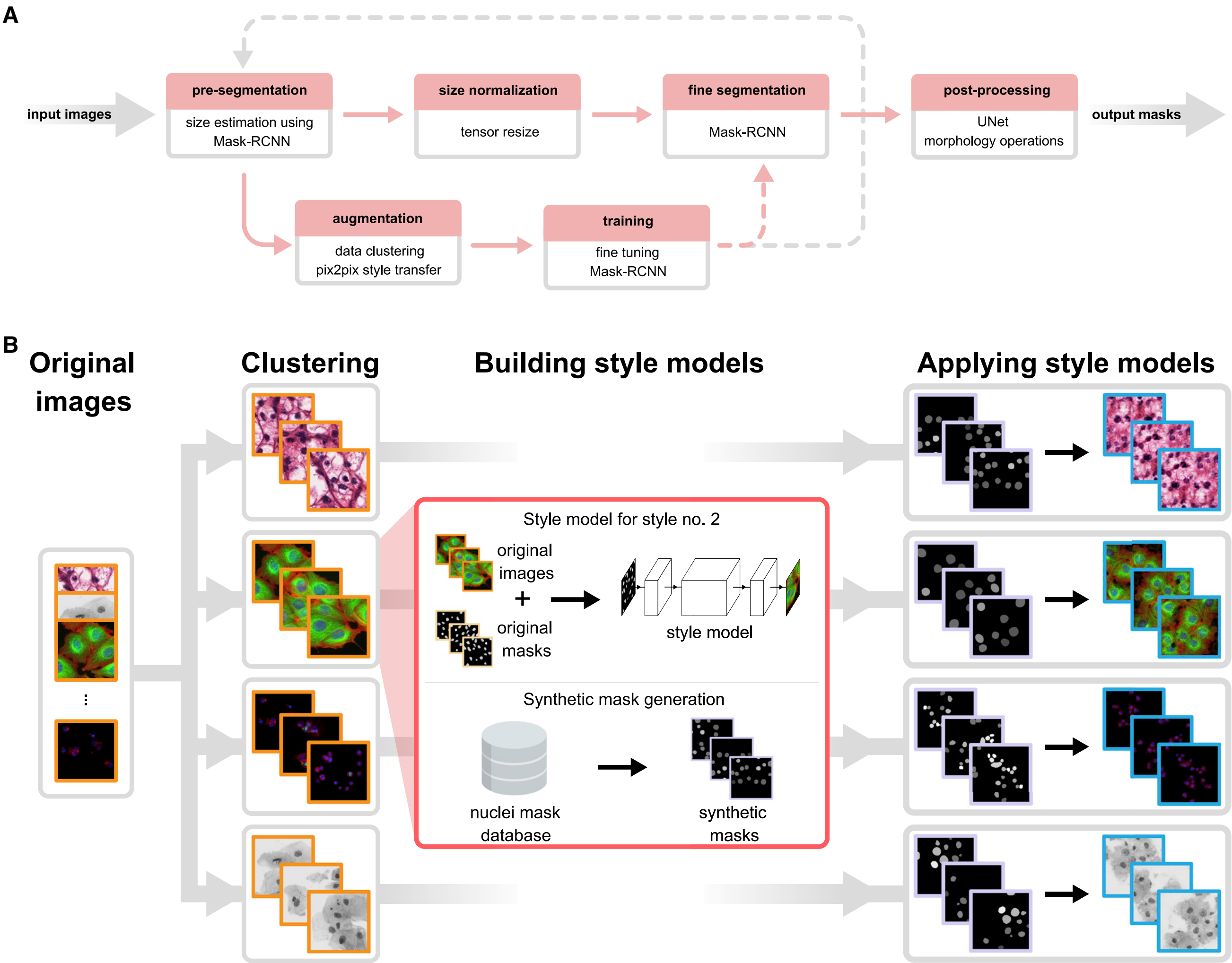


Figure 1. Overview of Our Approach

(A) Upper row of boxes presents the nucleus segmentation and pre-processing; an initial Mask R-CNN network estimates typical nucleus sizes, then images are rescaled such that mean nucleus size is uniform and a Mask R-CNN network trained on images with uniform nucleus size predicts segmentations. A contour refinement step using a U-Net-based network with a morphology operation is applied to obtain the final segmentation result. The data augmentation pipeline is depicted in the bottom row, the training set is augmented with an artificially generated set of image/label pairs in the target domain(s), and a pre-trained Mask R-CNN method is fine-tuned using the augmented images. Augmentation and training steps may be iteratively repeated as the gray dashed line suggests. Upper row depicts the inference pipeline; bottom row, training. Solid lines indicate data flow; dashed lines indicate transfer of a trained model.

(B) Image style-transfer-based data augmentation. To adapt our model to handle out-of-domain image types for which we have no segmentation labels, we synthesize new training data by first clustering images into similar groups, then learn a style transfer model. The style transfer model is provided with simulated nucleus masks, which mimic the number, shape, and size of the unseen nuclei, and then synthetic training image/label pairs are generated using the masks and the style transfer models. These data are added to the standard training data provided to Mask R-CNN, and the network learns to segment nuclei in the new domain. See also Figure S1.

at the close of the competition (https://www.kaggle.com/c/data-science-bowl-2018/leaderboard), a recent deep learning method, unet4nuclei (Caicedo et al., 2019a), which is based on a U-Net (Ronneberger et al., 2015) structure, a widely used Otsu threshold and seeded watershed method with object splitting (Carpenter et al., 2006), the pixel-based classification software ilastik (Sommer et al., 2011), and a more sophisticated but still classical gradient vector flow (GVF) based method, where an active contour is driven to edges using gradient vectors pointing to bright regions (Li et al., 2008) (Figure 2; Table S1; Data S2). Notably, the DSB stage 2 evaluation is performed on an unknown subset of the provided test images, many of which are outside the domain of the training images, truly challenging the ability of the model to generalize. We provide additional benchmarks and variations of our approach for comparison—including how our proposed style transfer learning step improves performance—in STAR Methods and Figure S2. Training a model on the same data with and without style transfer augmentation showed increased accuracy with style.

Our method scores higher (DSB-score, 0.633) than the top ranked deep learning approach (0.631, the highest of 739 teams) on the DSB stage 2 test set and has a simpler

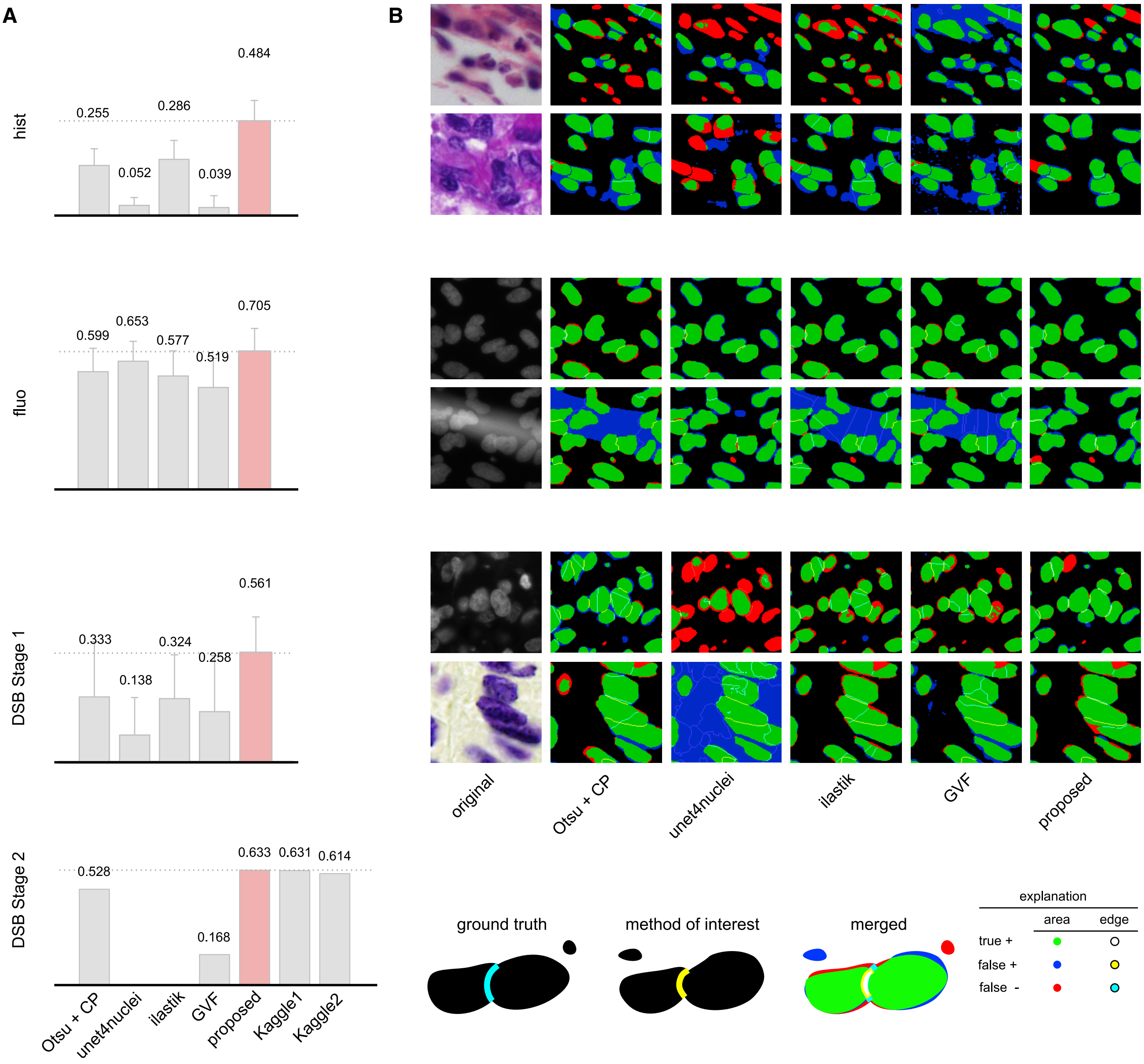


Figure 2. Results

(A) DSB-scores with error bars (standard deviation) for four image sets: hist, fluo, DSB stage 1, and DSB stage 2 (see details in STAR Methods). DSB-score is a modified mean average precision of segmented nuclei (see STAR Methods). Highest scores are marked with dashed lines and red color.

(B) Segmentation results for various methods on sample image crops with difficult cases (two example images of each); rows match those of (A) (note: ground truth is not public for DSB stage 2). A crop of the original image is provided in the first column, followed by segmentation results predicted by various methods. The color coding of the results is explained in the legend at the bottom. See also Figures S2, S5, and S8; Table S1.

architecture with fewer parameters. Our method outperforms all other tested methods, too, including a classical baseline (0.528) (Caicedo et al., 2019b; Carpenter et al., 2006) (Figure 2A). In addition, our proposed method outperformed all prior published results on *hist*, a diverse set of histology images and on *fluo*, a fluorescent image set (BBBC039; Caicedo et al., 2019a) (see Data S1 and S2 for details). A detailed comparison of our results against six other methods evaluated with additional metrics is provided in Table S1; Figures S5 and S8 (see details in STAR Methods).

DISCUSSION

We proposed a deep-learning-based nucleus segmentation approach designed for robustness to new experimental settings, using image style transfer to augment our training data with valuable out-of-domain samples. Our segmentation network learned from these artificially generated image/mask pairs, which mimic the patterns of new data types. This approach helped the network adapt to a diverse set of test data outside the domain of the training data, outperforming every other deep learning and classical method tested. Our generalized models successfully segment images across several domains, achieving performance close to or matching that achieved by models derived from and applied to a specific domain. The idea of augmenting difficult-to-obtain data using style transfer has enormous potential not only for nucleus detection but also more broadly in applications requiring some form of image understanding.

Key Changes Prompted by Reviewer Comments

The manuscript was extended with the section Segmentation Error Analysis describing both advantages and limitations of our approach compared with other methods, while practical runtime and resource details were also given in section Methods Used for Comparison for training and inference so that the reader might have a better overview of applicability. Specific algorithmic considerations were clarified more extensively, e.g., clustering and image style transfer or post-processing. For context, the complete transparent peer review record is included within the Supplemental Information.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Kaggle Competition
  - Data
  - Computational environment
  - Related work
- METHOD DETAILS
  - Overview of the Pipeline
  - Training and Style Transfer Data Augmentation
  - Clustering for Style Transfer Learning
  - Inference
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Evaluation Metrics
  - Methods Used for Comparison
  - Detailed Results

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.04.003.

ACKNOWLEDGMENTS

We acknowledge support from the LENDULET-BIOMAG grant (2018-342) and from the European Regional Development Funds (GINOP-2.3.2-15-2016-00006, GINOP-2.3.2-15-2016-00026, and GINOP-2.3.2-15-2016-00037), the Academy of Finland grant 310552 to L.P. Funding was provided by the National Institute of General Medical Sciences of the National Institutes of Health under MIRA R35 GM122547 to A.E.C. Funding was provided by the Swedish Research Council (Vetenskapsrädet) under research project grant 2017-040609_3 to K.S. We gratefully acknowledge the NVIDIA Corporation for the donation of NVIDIA TITAN Xp GPUs for our research.

AUTHOR CONTRIBUTIONS

R.H., A.S., E.T., C.M., F.K., T.D., A.K., K.S., and P.H. designed the method. R.H., A.S., T.T., E.T., C.M., B.M., I.G., J.M., A.B., M.G., M.K., E.M., T.B., K.K., W.W., J.C.C., N.B., F.K., L.P., T.D., A.K., and P.H. performed annotation, testing, and benchmarking. C.M., K.K., E.T., A.K., R.H., and P.H. designed the online tool. R.H., A.G., K.K., L.P., A.E.C., K.S., and P.H. wrote the manuscript. A.G. and A.E.C. enabled CellProfiler connection and co-organized the 2018 Data Science Bowl.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 5, 2019
Revised: January 22, 2020
Accepted: April 13, 2020
Published: May 7, 2020

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. *39*, 2481–2495.

Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al. (2019a). Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nat. Methods. *16*, 1247–1253.

Caicedo, J.C., Roth, J., Goodman, A., Becker, T., Karhohs, K.W., Broisin, M., Molnar, C., McQuin, C., Singh, S., Theis, F.J., and Carpenter, A.E. (2019b). Evaluation of deep learning strategies for nucleus sgmentation in fluorescence images. Cytometry A *95*, 952–965.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol *7*, R100.

Cui, Y., Zhang, G., Liu, Z., Xiong, Z., and Hu, J. (2018). A deep learning algorithm for one-step contour aware nuclei segmentation of histopathological images. arXiv http://arxiv.org/abs/1803.02786.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. *24*, 1342–1350.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al. (2019). U-net: deep learning for cell counting, detection, and morphometry. Nat. Methods *16*, 67–70.

Frank, E., Hall, M.A., and Witten, I.H. (2016). https://waikato.github.io/weka-wiki/citing_weka/.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask R-CNN. IEEE International Conference on Computer Vision (ICCV) *2017*, 2980–2988.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. arXiv https://arxiv.org/abs/1712.00409.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks IEEE Conference on Computer Vision and Pattern Recognition, *2017* (CVPR)), pp. 5967–5976.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90.

Lehmussola, A., Ruusuvuori, P., Selinummi, J., Huttunen, H., and Yli-Harja, O. (2007). Computational framework for simulating fluorescence microscope images with cell populations. IEEE Trans. Med. Imaging *26*, 1010–1016.

Li, G., Liu, T., Nie, J., Guo, L., Chen, J., Zhu, J., Xia, W., Mara, A., Holley, S., and Wong, S.T. (2008). Segmentation of touching cell nuclei using gradient flow tracking. J. Microsc. *231*, 47–58.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014). Microsoft COCO: common objects in context. Lect. Notes Comput. Sci. 740–755.

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., and Van Valen, D. (2019). Deep learning for cellular image analysis. Nat. Methods *16*, 1233–1246.

Molnar, C., Jermyn, I.H., Kato, Z., Rahkama, V., Östling, P., Mikkonen, P., Pietiäinen, V., and Horvath, P. (2016). Accurate morphology preserving segmentation of overlapping cells based on active contours. Sci. Rep. *6*, 32412.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. 234–241.

Van Valen, D.A., Kudo, T., Lane, K.M., Macklin, D.N., Quach, N.T., DeFelice, M.M., Maayan, I., Tanouchi, Y., Ashley, E.A., and Covert, M.W. (2016). Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. PLoS Comput. Biol. *12*, e1005177.

Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., et al. (2017). Content-aware image restoration: pushing the limits of fluorescence microscopy. Nat Methods *15*, 1090–1097.

Weng, L. (2017). Object detection for dummies part 3: R-CNN family. https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html.

Sommer, C., Straehle, C., Kothe, U., and Hamprecht, F.A. (2011). Ilastik: interactive learning and segmentation toolkit IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2011, 230–233.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: closing the gap to human-level performance in face verification IEEE Conference on Computer Vision and Pattern Recognition, 2014, 1701–1708.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Code repository	This manuscript	https://github.com/spreka/biomagdsb
NucleAIzer online tool	This manuscript	www.nucleaizer.org
CellProfiler plugin	This manuscript	https://github.com/CellProfiler/CellProfiler-plugins/blob/master/nucleaizer.py

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Peter Horvath (horvath.peter@brc.hu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The authors declare that the data supporting the findings of this study are available within the paper and its Supplemental Information files.

The authors also declare that the software supporting the findings of this study are available within the paper, its Supplemental Information files, under www.nucleaizer.org, and https://github.com/spreka/biomagdsb.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Kaggle Competition

We designed our pipeline to recognize nuclei as accurately as possible in a wide variety of images acquired with different microscopes, under varying imaging conditions with different stains for nuclei of various cell types. This was the challenge set forth in the 2018 Data Science Bowl (DSB) by Kaggle, Booz Allen Hamilton and the Broad Institute. The competition included a preparatory stage 1, to which teams could submit their solutions during a four-month period and a 4-day long stage 2 final scoring period.

Existing nucleus segmentation methods do not generalize well, they perform well only on the limited experimental conditions they are designed or tuned for. The Data Science Bowl was highly successful in the sense that many robust solutions were developed that pushed the state-of-the-art in terms of segmentation performance and insensitivity to image type and quality. Solutions such as ours are now being developed into toolkits for biologists that will accelerate science by improving automation in identifying nuclei.

We participated in the competition in both stages, reaching the top 1% in stage1 and top 4% in stage 2. The presented results are based on further improvements post-competition.

Data

The official dataset for the challenge is composed of a training set and two tests sets, one for each stage. The number of images in each set is 670 (training), 65 (stage 1 test), and 3019 (stage 2 test), stage 1 test masks were released in the second stage. The final evaluation of the teams' performance was measured on a subset of the stage 2 test set (the identity of the subset remained hidden to the competitors). Many of the competitors used additional data besides the provided training data, as this was permitted as long as participants shared their sources on the official competition website (https://www.kaggle.com/c/data-science-bowl-2018). Our annotated training data included 12 additional data sources besides the DSB data, including some data sources annotated by experts in our institution. This extended the total number of training image/mask pairs from 735 to 1,102, and the number of annotated nuclei from 33,814 to 80,692 (not including the synthetic data). A summary of the data we used is provided in Table S2.

Using style transfer, we augmented our training data with synthetic image/mask pairs generated in the style of k=134 clusters of images from the DSB Stage 2 set, as described in Sections Clustering to Synthesizing new image/mask pairs. This added 2,680 synthetic image/mask pairs to the training data (approximately 263,701 annotated nuclei).

We tested various versions of our method along with several competing methods on four test datasets: *DSB test1*, *DSB test2*, *fluo*, and *hist*. DSB test1 and DSB test2 are heterogeneous test sets from the Kaggle challenge (stage 1 and stage 2). The *fluo* dataset is fluorescence images of U2OS cells in a chemical screen taken from the Broad Bioimage Benchmark Collection

(BBBC039) (Caicedo et al., 2019a). The *hist* dataset is a mixture of histology images collected from the internet and prostate H&E stained slides collected in-house.

A fraction of the histological images manually annotated in our lab were used as test set *hist* (see Supplemental Data). BBBC039 (Caicedo et al., 2019a) images were used to train a fluorescent segmentation model, we refer as *fluo*. The *hist* and *fluo* test sets are disjoint from the respective training data.

We carefully prepared our test sets for evaluation by automatic clustering as follows. Each test set was split into disjoint parts; one was completely held out of all training procedures and solely used for evaluation, while the remaining part served as out-of-domain unannotated data, was clustered by *k*-means and forwarded to style transfer and subsequent training steps.

We collected histopathology images of test set *hist* intentionally from such experiments that lacked similar instances in our entire training set to test how well our approach would perform on various out-of-domain experiments. Hence, only style transfer learning could be used to input these missing domains' information to our segmentation network.

All input images were initially converted to 8-bit 3-channel RGB images in.png format as well as images produced by our pipeline (except masks).

Computational environment

Software

Our pipeline is implemented using a shell script to allow continuous execution of the entire pipeline. Python 3 scripts execute the training and inference of Mask R-CNN, U-Net, and pix2pix which rely on the TensorFlow, Keras, and PyTorch environments. The clustering, post-processing, and initial steps of style transfer are implemented in Matlab. Our software is available for download at: https://github.com/spreka/biomagdsb where a detailed documentation can also be found discussing the required versions of frameworks and details about the architecture parameters.

The entire pipeline can be run both under Linux and Windows. In a typical use case, it is not necessary to retrain any of the models. Calling the *postComp* method without post processing provides excellent results. For specific experiments with no ground truth annotations, performing the style transfer learning part of our pipeline generates new synthetic training data in the missing domain on which training a new model results in fine segmentation. Alternatively, an online version of our method is available at www.nucleaizer.org.

Hardware

Our methods were trained and tested on a variety of Nvidia graphics cards, including GTX 1070, 1080Ti, and Titan Xp.

Related work

Mask R-CNN

He et al. (2017) published Mask R-CNN as an extension of Faster R-CNN to allow simultaneous instance detection and segmentation. The network architecture is similar to that of Faster R-CNN: feature extraction uses ResNet (50 or 101 layers) or alternatively Feature Pyramid Network (FPN), while head is as in Faster R-CNN extended with a fully convolutional mask prediction branch. A detailed discussion of extended R-CNN versions can be found in Weng, 2017.

We decided to incorporate Mask R-CNN in our pipeline due to its robustness, scalability and instance-awareness. It is currently one of the leading computational architectures in instance segmentation of arbitrary object classes, and its applications dominated the methods submitted to the DSB 2018 competition alongside solutions based on U-Net.

U-Net

U-Net (Ronneberger et al., 2015) was specifically created for bioimage segmentation with an encoder-decoder architecture and skip connections between layers of the encoding branch and decoding branch to provide the decoder with access to spatial information to reason about upsampling the segmentation.

We applied U-Net in our post-processing pipeline as it can efficiently be used to detect subtle differences such as those around the edges of objects. The network structure is straightforward and computationally feasible.

Post-processing the segmented nuclei per se is needed due to the inevitable uncertainty in marginal cases, like relatively small objects most likely corresponding to false detections. We found probability maps predicted by U-Net helpful in such scenarios.

METHOD DETAILS

Overview of the Pipeline

As a first step, pre-segmentation of the input images is performed using a pre-trained deep convolutional model (which we refer as *preseg*) to estimate nuclei sizes as well as to create a mask input for image style transfer learning. Simultaneously, we cluster similar images of the input data into groups, and learn styles on these clusters (see Figure 1B and sections Clustering for Style Transfer Learning and Learning Image Style Transfer Models for details). As a next step, we extend the training data with artificially created style transferred images for fine-tuning a Mask R-CNN (He et al., 2017) pre-trained on our nucleus segmentation dataset. For inference on unseen data, we use the refined Mask R-CNN network incorporating knowledge about estimated cell sizes. The resulting contours are refined with U-Net (Ronneberger et al., 2015) and a morphology step.

The proposed method consists of procedures for training and inference, as shown in Figure S1. Inference merely requires unannotated images as its input – provided the pre-trained models are available. Training the network produces a learned segmentation model, and requires a set of annotated training data and a pre-trained segmentation network (pre-segmentation network), as well as any available unannotated images that can be used for data augmentation. The pretrained segmentation network is crucial to both the training and inference procedures, so we discuss it first and then continue with training and prediction steps.

Training and Style Transfer Data Augmentation

Pre-segmentation

The architecture for the segmentation networks is based on the Mask R-CNN architecture. The pretrained segmentation network (pre-segmentation network) is used to make rough estimates about the nucleus size and shape while being robust to changes in imaging modality or magnification. The network is initialized with pretrained weights from the MS-COCO dataset, which contains images and segmentation masks for 91 object types including people, trucks, sheep, dogs, etc. For details about the original COCO competition see http://cocodataset.org or the corresponding publication (Lin et al., 2014). The network was trained using a diverse set of annotated images containing various imaging modalities, cell lines, magnifications, etc. For more information see Section Data. The annotations consisted of segmentation masks for the nuclei. Augmentation was used during training including geometric transformations, intensity stretching, cropping, noise, and blur (see Data S1 documentation for details).

The resulting network, which we refer to as *preseg*, already performed reasonably well on unannotated images in the test set (Figure S2), although this was not its purpose. The preseg network is used to: estimate properties of nuclei in new unannotated images (size, shape, and area) in clustering, and to generate rough segmentations on unannotated images for the style transfer data augmentation step (see the following two sections for details).

Clustering for Style Transfer Learning

Images without annotations are automatically clustered to define multiple groups with similar properties: textures, imaging modalities, cell lines, sample type (tissue or culture), etc. These groups are used as data sources to learn style transfer models to generate additional synthetic data that mimics the properties of each cluster of unannotated images.

To perform the clustering, we use a pairwise similarity metric between feature vectors describing each unannotated image. Features were extracted using CellProfiler (Carpenter et al., 2006) modules including intensity and texture and a similarity metric was computed by a shallow fully connected neural network (Frank et al., 2016). This similarity network was trained on the DSB train1 data set, where images taken with the same condition are given a label of 1 and images from different conditions are given a label of 0. The output of this network on the unannotated data yielded a similarity matrix which we clustered with k-means. The number of clusters, k=134 for DSB stage 2 test set, was chosen automatically based on the number of images to over-segment the groups to avoid accidental mixing of the true underlying groups. Ideally, each obtained cluster of unannotated images represents a “style” or distribution of data which can be augmented with style transfer (e.g. digital slides of H&E stained breast cancer histology samples at 63x magnification, or fluorescent images of Human MCF7 cell nuclei at 40x).

Learning Image Style Transfer Models

We use the pix2pix (Isola et al., 2017) framework for image style transfer (https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix). The architecture consists of two adversarial networks, a *generator* tasked with synthesizing realistic looking images, and a *discriminator* tasked with identifying real images from synthesized images. This model learns to map one image domain to another through an adversarial loss that encourages the generator to learn to fool the discriminator. The input to the generator in our case is a binary mask containing 1’s at the locations of the desired nuclei, and 0’s elsewhere. The input of the discriminator is an image/mask pair (either a real pair, or a synthetically generated pair). The generator learns to transform the binary mask into the desired style of the real images from the cluster, and the discriminator encourages this by trying to identify real image/mask pairs from fakes. We use the rough segmentations provided by the *preseg* network as masks for the unannotated images in the style cluster during learning. We train a pix2pix style transfer network to synthesize realistic images from masks for each of the style clusters.

Synthesizing New Image/Mask Pairs

Using our set of 134 trained style transfer networks, we synthesized 20 new image/mask pairs for each of the styles in the unannotated data. A crucial step for this task was to generate novel binary masks to provide as input to the style transfer network, which uses the mask to generate a realistic image of the cells with nuclei in the locations defined in the mask. We generated the masks algorithmically as a combination of 1) fetching real nuclei masks from a database, and 2) synthesizing nuclei using software (simcep; Lehmussola et al., 2007). Approximately 50% of the nucleus masks were created using each approach. In this manner, we generated 20 masks for each of the 134 style clusters, and then used the style transfer network to generate the corresponding images.

We assembled our nucleus mask database from images of the official DSB training set and further external datasets (see Table S2) - some of which we corrected for slight contour errors - and added each nucleus mask to the database. We fetched such nuclei masks that follow the features of the desired style and placed them on the synthetic mask images in accordance with the localization properties of the given style.

Training the Mask R-CNN Segmentation Network

The synthetic image/mask pairs generated by the style transfer network were added to the annotated training data to update the Mask R-CNN segmentation network. We used the implementation of Matterport (https://github.com/matterport/Mask_RCNN) and wrote handler scripts in Python to create the appropriate data structures and call functions. Training was performed in 3 steps with decreasing learning rate and targeted different layers of the Mask R-CNN network, as described in the documentation of the aforementioned Matterport repository.

The loss function was as defined in (He et al., 2017): it comprises of classification, localization and segmentation mask losses: $L=L_{cls}+L_{box}+L_{mask}$ by ROIs, and defines mask loss as follows. Given the *k*-th region does belong to ground truth class *k* it takes the average binary cross-entropy loss which is formulated as

$$L_{mask} = -\frac{1}{m^2}\sum_{1\le i,j\le m}\left[y_{ij}\cdot\log\hat{y}_{ij}^{k} + (1-y_{ij})\log\left(1-\hat{y}_{ij}^{k}\right)\right] \quad \text{(Equation 1)}$$

where y_{ij} is the true label of a cell (i,j) from a ROI of *mxm* size on the ground truth mask of class *k* and $\hat{y}_{ij}^{k}$ is the predicted class label of the same cell. The formula only includes masks for ground truth class *k* that are associated with the *k*-th class.

Image Augmentation and Resizing

The performance of deep learning networks is known to scale with the size of the dataset (Hestness et al., 2017). Therefore, we use a number of approaches to augment the training data. The first, as we described above, is to add new synthetic image/mask pairs generated in the style of unseen examples to the existing annotated training data. Each minibatch contained 10-50% synthetic images. We also used standard data augmentation techniques including random cropping, colour channel swapping, intensity modification by histogram stretching or equalization and inversion, rotation to an arbitrary degree and translation as geometric transformations and finally, to better resemble low-quality images, blur and additive noise were used as well. These operations were applied to all the input training data – style transfer results too – with a random probability.

MASK R-CNN is reasonably robust to changes in scale, but superior performance is obtained if the nucleus size is approximately 40 pixels in diameter for the data and parameters we used. Figure S3 shows the results of the robustness of our method with a fixed parameter against different nuclei sizes. Quantitative evaluation is shown in Figure S6.

Another preprocessing step was to resize the images by a scaling factor to obtain a training dataset homogeneous both in cell and image size. The scaling factors were computed from the size estimation of the *preseg* nucleus masks such that the resulting mean cell size is set to 40 pixels diameter. Images were then either cropped or padded so that the resulting image was 512 x 512 pixels.

Inference

Mask R-CNN Prediction

The Mask R-CNN model trained as described above is used to predict segmentation masks when new images are provided as input. The images are resized before they are input to the network as described in the previous section.

Post-processing and U-Net Correction

We found that the segmentations could be further improved by postprocessing and refining nucleus contours using U-Net (Ronneberger et al., 2015). This encouraged better boundary reasoning between adjacent nuclei, and finer segmentations with the background. First, outlier objects were removed or merged as follows: 1) Smaller objects that were entirely within another object were eliminated. 2) objects that were surrounded by another object more than p_1% were merged, and 3) objects smaller than p_2 pixels area were removed. Next, U-Net based correction was performed (Figure S4): 1) an optimal threshold p_3 for U-Net probability values was determined, 2) a soft margin around the Mask R-CNN contour was defined for each object, with an extension of p_4 pixels inwards and p_5 outwards. The contour was extended/shrunk based on the U-Net predictions. 3) objects that had in total less than p_6 mean U-Net probability were removed. Parameters p_1..p_6 were optimized on the training set with a genetic algorithm to the DSB-score function (see formulation in section Evaluation Metrics). Best values were: (0.17, 44, 0.9375, 1, 1, 0.8).

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation Metrics

The evaluation metric used for the DSB competition is based on the mean average precision, as defined on the competition website, at different intersection-over-union (IoU) thresholds. A successful nucleus detection was determined by an IoU test (also known as the Jaccard index):

$$IoU(x,y) = \frac{|x\cap y|}{|x\cup y|} = \frac{|x\cap y|}{|x|+|y|-|x\cap y|} \quad \text{(Equation 2)}$$

which measures the overlap between prediction pixels *x* and the annotation pixels *y* over the intersection of the two areas. Using a threshold ranging from 0.5 to 0.95 with steps of 0.05, true positive (*TP*) detections, false positive (*FP*) detections and false negative (*FN*) detections were identified. For a threshold of 0.5, a predicted object is considered a “hit” if the IoU is greater than 0.5. For each threshold *t*, a modified version of precision was calculated

$$DSB\ score(t) = \frac{TP(t)}{TP(t)+FP(t)+FN(t)+\varepsilon} \quad \text{(Equation 3)}$$

for all thresholds in (0.5, 0.95). These scores were averaged for all thresholds, and then the mean of the average scores is reported over the images in the test dataset. In addition to the DSB-score, we evaluated our results with three additional metrics based on the IoU detection test: mean average precision- (mAP), recall and F1-score. We used the same *t*, TP, FP and FN values as above. We also added a small $\varepsilon = 10^{-40}$ value to the denominators.

$$precision(t) = \frac{TP(t)}{TP(t) + FP(t) + \varepsilon} \quad \text{(Equation 4)}$$

$$recall(t) = \frac{TP(t)}{TP(t) + FN(t) + \varepsilon} \quad \text{(Equation 5)}$$

$$F1\ score(t) = 2 \cdot \frac{precision(t) \cdot recall(t)}{precision(t) + recall(t) + \varepsilon} \quad \text{(Equation 6)}$$

The same strategy was used to calculate mean values for these measures as was for the DSB-score, taking the average over various thresholds *t*, and the mean among the test images. In the following sections, we refer to these measures as mAP (mean average precision), mAR (mean average recall), and mF1 (mean average F1-score).

We also introduce classification accuracy regarding our style-transfer generated image quality evaluation as follows:

$$accuracy = \frac{\sum correctly\ classified\ instances}{\sum instances} \quad \text{(Equation 7)}$$

Methods Used for Comparison

Our tests included several variations of our method along with six competing methods and several variations of our approaching using different style augmentation: *NOstyle* did not contain style augmented images, *AUTOstyle* used nuclei masks generated by the preseg network, and *GTstyle* used hand annotated ground truth to generate nuclei masks. CellProfiler (*CP*) (Carpenter et al., 2006) is a widely-used bioimage analysis software incorporating several methods to segment, measure and analyze cellular compartments. We created multiple pipelines for the different image types of the test sets – except for our fluorescent set which comprised of a single experiment. *Preseg* refers to our general scale-independent pre-segmentation model while *postComp* is our final refined post-competition submission (an *AUTOstyle* model customized for DSB test2).

We compared against several other approaches including *ilastik* (Sommer et al., 2011), which provides a pixel classification setup where users can manually annotate regions of the input images to desired classes and obtain predictions as either probability maps or segmented images. Segmentations were obtained by applying a threshold to probabilities from *ilastik* (with additional object splitting). *Unet4nuclei* (Caicedo et al., 2019a) is an implementation of the popular U-Net deep learning approach to segmentation. *GVF* (Li et al., 2008), or gradient vector flow, is an active contour-based segmentation method suitable if objects are bright regions on a dark background. Pipelines of these compared methods are provided in Data S2. *DSB1* and *DSB2* are the first and second place entries on the final Kaggle leaderboard. The approach from *DSB1* (https://www.kaggle.com/c/data-science-bowl-2018/discussion/54741) uses a very deep U-Net architecture along with prediction of touching borders. *DSB2* also uses a U-Net approach, and forces the network to predict relative locations within each nucleus (https://github.com/jacobkie/2018DSB).

Comparing the complexity as well as the computation time and resources needed to train *DSB1*, we are confident to claim that our method is considerably simpler and much faster. *DSB1* combines a total of 32 trained deep neural networks to achieve their reported score on *DSB test2* set, the training of which can take days even when performed on a high computation-capable GPU (Nvidia GTX 1080Ti). In contrast, in our method only a Mask R-CNN and U-Net models are trained for prediction, taking approximately 10 hours for training on the same GPU. The computation time for image style transfer strictly depends on the number of different styles present in the target data as one style model is trained for each, individually taking about 15 minutes. *DSB2* uses a simpler architecture.

We also investigated computation time regarding inference with our method. Even though inference time is affected by multiple circumstances including image size, number of objects on the image and VRAM of the GPU used, an approximate one image per 2 seconds can be achieved given the following. An image of 520x696 pixels size having about 120 objects of ~20 pixels median diameter size, rescaled to 2x its original size to have ~40 pixels diameter sized objects, i.e. 1040×1396 pixels resized image, on an Nvidia GTX 1080Ti GPU having 11 GB VRAM can be predicted in 2 seconds.

Detailed Results

Style Transfer Increases Performance

We tested the methods outlined in Section Methods Used for Comparison on four test datasets: *DSB test1*, *DSB test2*, *fluo*, and *hist*, described in Section Data. The resulting DSB-scores are presented in Table S1. When running these tests, the test data was never included in the data to train the model, e.g. when testing on *DSB test1*, the *DSB test1* data was held out from the training set. Similarly, when testing on hist, *biomag2* and *biomag6* subsets were held out.

The test image sets were used as style transfer learning input as determined by our automatic clustering method: a portion of the set was left out when the clustering algorithm could not find a sufficient number of images for a cluster. Therefore, we report our results on such fractions of the test sets that none of the deep learning networks have seen prior to inference as follows. 100/200 *fluo*, 21/50 *hist*, 28/65 *DSB test1* images were used for evaluation. None of the final *DSB test2* evaluation image set was used for training.

The results demonstrate that our style transfer approach improves performance in test sets containing data from heterogeneous sources: *hist*, *DSB test1* and *DSB test2*. We also see excellent performance on single domain fluorescence data, *fluo*. Comparing the results of our method with (*AUTOstyle* [postComp is the *AUTOstyle* for DSB stage 2 test] and *GTstyle*) and without style transfer augmentation (*NOstyle*), we see a clear trend towards increased performance with style transfer augmentation. If we have access to ground truth nucleus masks (*GTstyle*) our performance improves, though in many realistic scenarios such masks will not be available. Figure S2 shows the output of the various methods we tested on challenging examples (note that *DSB1* and *DSB2* are not reported because we did not have access to their code). In Figure S5, we present mAP, mAR, mF1 and mIoU metrics for the various methods on each dataset. As expected, there is a strong correlation between the metrics.

Objects of Various Sizes can Be Detected Accurately

In addition to the qualitative demonstration on Figure S3, we provide a quantitative analysis of the range of object sizes correctly detected by two of our compared methods: *preseg* and *postComp*. Note that while *postComp* was trained on fixed sized (40 pixels diameter) nuclei images and is expected to perform best on objects of approximately the fixed size, *preseg* is more flexible as we intentionally included images presenting a wide range of object sizes in its training to prepare it for an initial robustness. Therefore we expect *preseg* to detect objects robustly in a wider size range. We tested both models on DSB stage 2 test set and scaled the images to 0.25-4.0 times relative to our generally expected median 40 pixels diameter objects. Our results confirm our expectations of *preseg* (our scale-independent model) which performs significantly better than *postComp* (scaled model) on shrunk images as presented on Figure S6 below. We found that the accuracy of both models is decreased far less rapidly when enlarging the images.

We also note that the object sizes can vary on individual images (Figure S6B) suggesting the scaling procedure by median object sizes cannot necessarily be optimal for all images; we mark some of the extremes with black arrows.

Synthetic Images Are often Mistaken for Real

We tested how well our style transfer-generated synthetic images compared to real microscopy images by showing a representative selection of both to field experts (pathologists and biologists) and asked them to tell the synthetic images apart from the real ones. The only prior information forwarded to the participating experts was there are fake images in the collection. Their decision accuracy was measured in a binary fashion: whether the expert could identify a truly synthetic image (1) or not (0). We show an example test image montage below (Figure S7) with the average detection of experts and the labels (real or fake). We collected 64 cropped images each for our two test image mosaics comprising of 50% real and fake tiles, respectively.

We report an approximate 57% accuracy (ranging from 42% to 73%) of fake image recognition averaging both our experts and the test cases. Based on the performance of the experts we can conclude the visual quality of the style transfer-generated images is on par with real microscopy images suggesting the advance our approach may bring to cellular compartment segmentation.

Segmentation Error Analysis

We visually compare segmentation errors and improvements on Figures 2 and S2. To better understand the distribution of such common errors in any of the analyzed segmentation methods we compared how well they perform in terms of avoiding the main error types: 1) missing a nucleus, 2) falsely detecting an object as nucleus, 3) splitting a nucleus and 4) merging adjacent nuclei unnecessarily. An example image presented on Figure S8A shows them visually. All existing methods fail to overcome these issues in at least some instances, as they significantly depend on the experimental and imaging conditions used to produce the images. Our method aims to help reduce these issues.

We measured such types of errors as follows. 1) a missed nucleus is a false negative (FN) i.e. present on ground truth (GT) with no corresponding object on the prediction. 2) A falsely detected nucleus is a false positive (FP): a predicted object with no corresponding GT. 3) A split nucleus is identified as two or more predicted objects that overlap with a significant region of the best corresponding GT object, respectively; we considered an overlap of at least 30% as significant in this case if two objects contributed to the overlap, and 15% if more. Splits were only considered if the given GT object did not have a single matching predicted object. 4) A merged nucleus is a single predicted object that has a significant overlap with multiple GT objects each. We calculated merges similarly to splits but swapped the role of GT and predicted objects.

We conducted our evaluation on the same subsets of each test set discussed in the previous sections. Quantitative analysis of segmentation errors support our results: our method (and its modified versions) generally outperform the compared methods. Comparative results are displayed on Figures S8C–S8E. Remarkably, *unet4nuclei* produced in total fewer errors than our methods on test set *fluo* but it has been trained and published on this image set.

Segmentation errors naturally occur in automatic methods. Classical methods (*CP*, *ilastik*, *GVF*) tend to predict a higher frequency of false positive objects, typically on complex background regions similar to e.g. Figure S4C. They are also more prone to merging touching nuclei or background regions around them to the objects (see Figure S2B rows 1–2) and to split larger, irregularly shaped objects. *Unet4nuclei* could not have been trained accurately enough for heterogeneous sets (*hist*, *DSB test1*) due to the inevitable uncertainty of U-Net in complex histological regions while it excelled on the single-domain set *fluo*.

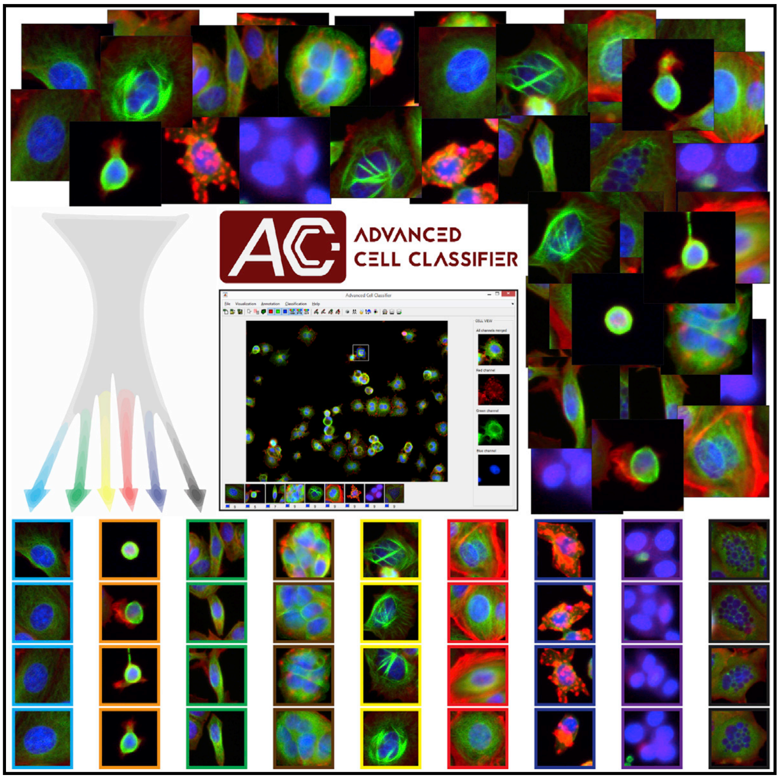
Our method typically failed to split (i.e. merged) very small or elongated adjacent nuclei with weak textural difference from the dividing background region. Similarly, it unnecessarily split nuclei in cases where texture or edge information may suggest multiple nuclei-like structures inside a single nucleus.

Cell Systems

Tool

Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data

Graphical Abstract



Authors

Filippo Piccinini, Tamas Balassa, Abel Szkalisity, ..., Ulrike Kutay, Kevin Smith, Peter Horvath

Correspondence

horvath.peter@brc.mta.hu

In Brief

We have developed new algorithms to mine microscopic image-based screening data, discover new phenotypes, and improve recognition performance. These methods are implemented into a user-friendly, free open-source tool that improves phenotype classification accuracy and helps users to quickly uncover hidden phenotypes from large datasets.

Highlights

- A powerful new suite of algorithms for phenotype discovery in image-based screens
- A visual tool for exploring and annotating phenotypes within large datasets
- New methods to find similar looking cells and suggest useful annotations
- Faster and more complete discovery of phenotypes, more accurate classification

Piccinini et al., 2017, Cell Systems *4*, 651–655
June 28, 2017 © 2017 Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2017.05.012

Cell Systems
Tool

Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data

Filippo Piccinini,[1,8] Tamas Balassa,[2,8] Abel Szkalisity,[2] Csaba Molnar,[2] Lassi Paavolainen,[3] Kaisa Kujala,[3] Krisztina Buzas,[2,4] Marie Sarazova,[5] Vilja Pietiainen,[3] Ulrike Kutay,[5] Kevin Smith,[6,7] and Peter Horvath[2,3,9,*]

[1]Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) S.r.l., IRCCS, Via Piero Maroncelli 40, 47014 Meldola (FC), Italy
[2]Synthetic and Systems Biology Unit, Hungarian Academy of Sciences, Biological Research Center (BRC), Temesvári körút 62, 6726 Szeged, Hungary
[3]Institute for Molecular Medicine Finland, University of Helsinki, Tukholmankatu 8, 00014 Helsinki, Finland
[4]University of Szeged, Faculty of Dentistry, Tisza Lajos körút 64, 6720 Szeged, Hungary
[5]Insitute of Biochemistry, ETH Zurich, Otto-Stern-Weg 3, 8093 Zurich, Switzerland
[6]KTH Royal Institute of Technology, School of Computer Science and Communication, Lindstedtsvägen 3, 10044 Stockholm, Sweden
[7]Science for Life Laboratory, Tomtebodavägen 23A, 17165 Solna, Sweden
[8]These authors contributed equally
[9]Lead Contact
*Correspondence: horvath.peter@brc.mta.hu

http://dx.doi.org/10.1016/j.cels.2017.05.012

SUMMARY

High-content, imaging-based screens now routinely generate data on a scale that precludes manual verification and interrogation. Software applying machine learning has become an essential tool to automate analysis, but these methods require annotated examples to learn from. Efficiently exploring large datasets to find relevant examples remains a challenging bottleneck. Here, we present Advanced Cell Classifier (ACC), a graphical software package for phenotypic analysis that addresses these difficulties. ACC applies machine-learning and image-analysis methods to high-content data generated by large-scale, cell-based experiments. It features methods to mine microscopic image data, discover new phenotypes, and improve recognition performance. We demonstrate that these features substantially expedite the training process, successfully uncover rare phenotypes, and improve the accuracy of the analysis. ACC is extensively documented, designed to be user-friendly for researchers without machine-learning expertise, and distributed as a free open-source tool at www.cellclassifier.org.

INTRODUCTION

In the past, limits in automation, processing, and storage technology imposed practical restrictions on the number of images we could analyze. Typical research projects were limited to a few dozen to a few hundred. On that scale, it was possible to verify entire experiments manually. Today, high-content screening (HCS) experiments easily produce over 10,000 times more data (Neumann et al., 2010). It is no longer feasible to visually interpret and verify every data point. As we depend more and more on automated computational methods for complex and large-scale image analysis, we run the risk of only partially understanding the data. We must ask ourselves “Have I understood the data?” and “Is my analysis as accurate as possible?” Without the right analysis tools, our answers to these questions may be unsatisfactory.

In this paper, we introduce Advanced Cell Classifier (ACC), a machine-learning software designed to give a quicker and more complete understanding of large datasets and to train predictive models as accurately as possible. In 2011, we released ACC version 1.0 (ACC v1.0), a graphical image analysis software tool that offers access to a variety of machine-learning methods and provides accurate analysis (Horvath et al., 2011). Several large-scale cell-based phenotypic HCS studies have made use of ACC v1.0, including at least 15 human genome-wide RNAi screens and numerous extensive drug screens. These studies cover a wide variety of biological topics ranging from the studies of influenza A virus (Banerjee et al., 2014) to studies of acute lymphoblastic leukemia (Fischer et al., 2015).

ACC v1.0 shared a drawback with other similar machine-learning HCS analysis software (Orlov et al., 2008; Uhlmann et al., 2016; Held et al., 2010; Sommer et al., 2011; Laksameethanasan et al., 2013; Ogier and Dorval, 2012; Rämö et al., 2009; Misselwitz et al., 2010; Jones et al., 2008; Dao et al., 2016); it lacked discovery and data visualization tools to enable the user to fully explore and understand their data. We view this as a crucial shortcoming. Little attention has been given to the methods used to explore the data, understand it efficiently, or to ensure the quality of the annotations. The cost of collecting expert annotations is high and categorizing cells into strict classes is often ambiguous. Experts are often unsure if they have uncovered all the important phenotypes buried within the data because they lack the tools to fully explore it. There are also limitations in the annotation process. Existing software packages force the user to manually select cells to label or randomly select

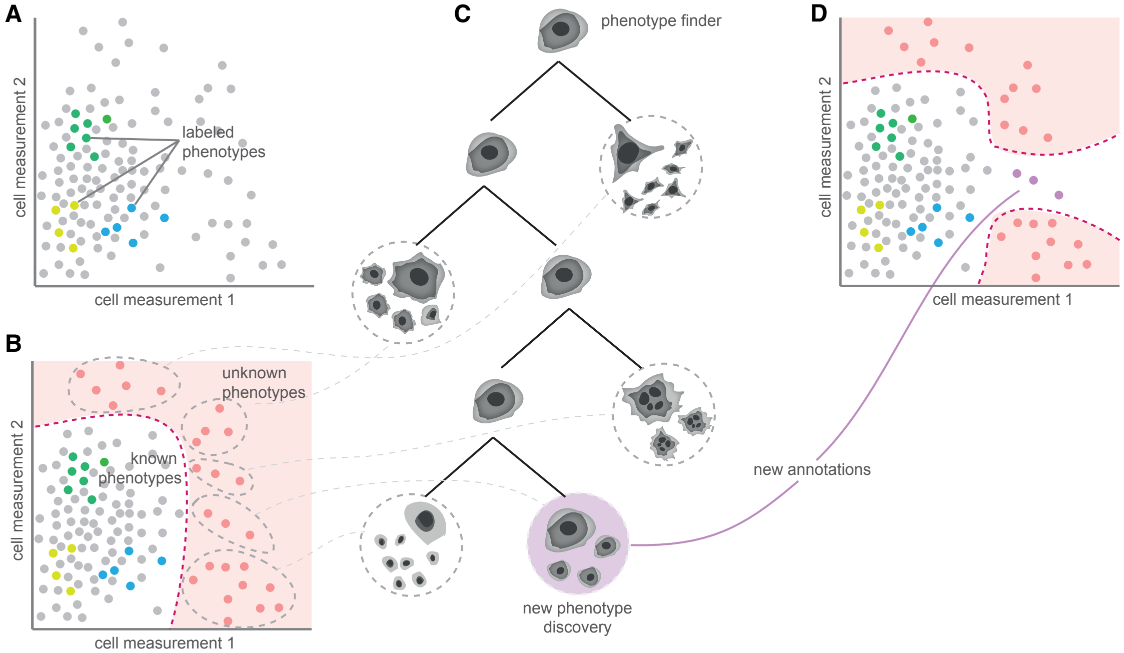


Figure 1. Phenotype Finder Tool

The phenotype finder tool organizes cells into a browsable hierarchy, which facilitates the discovery of new classes.

(A) Cells are represented by dots embedded in a two-dimensional synthetic feature space. Sets of cells annotated by the expert are shown in green, yellow, and blue. Unannotated cells are shown in gray.

(B) A one-class classifier is used to automatically determine which cells are least similar to the known cell types (pink region).

(C) Cells with the least similarity to known examples are sampled and clustered to construct a dendrogram. The expert can browse and analyze the representative cells shown in the tree, create a new phenotype class from these cells, or add cells to an existing one. In addition, irrelevant cells can be discarded to improve the classification results to be obtained in the next run.

(D) If a new phenotype is discovered, the region of non-annotated cells changes in the multidimensional feature space.

cells. These methods are inefficient and generate many wasteful annotations that ultimately do not prove useful for the classifier. Furthermore, these procedures make it difficult to discover new phenotypes or to intelligently refine decision boundaries (Smith and Horvath, 2014).

ACC v2.0 is a completely re-designed and user-friendly software tool with the goal of improving the collection and understanding of image data and the accuracy of the analysis (Figure S1). It allows researchers, even those without computer vision or machine-learning knowledge, to efficiently characterize and exploit their cell-based and image-based HCS experiments, leading to new discoveries. The main differences between ACC v2.0 and its previous versions include: (1) intelligent methods to explore and annotate large single-cell image data, including an active learning approach to improve the accuracy of the classifier, and similar cell search, an algorithm to find similar cells and increase the number of annotations for rare phenotypes; (2) an easy-to-use report generator to automatically obtain statistics on cell distribution and class incidence; (3) a new, re-designed and user-friendly interface; (4) detailed documentation, video tutorials, and online resources; and (5) improved data visualization methods. Finally, and most importantly, (6) we have implemented phenotype finder, a novel method to automatically discover new and biologically relevant cell phenotypes (Figure 1). The source code of ACC v2.0 is freely distributed as an open-source tool at www.cellclassifier.org.

RESULTS

To evaluate the effectiveness of the discovery and annotation tools included in ACC v2.0, we generated a synthetic dataset simulating images of cells from a high-content screen. This provided us with completely accurate knowledge concerning every cell and phenotype in every image, something that is impossible with images of actual cells. Using this information, we compared ACC v1.0 with ACC v2.0. A group of annotators were able to reach much higher recognition accuracy and found even extremely rare phenotypes in significantly less time using the tools we developed. A description of the synthetic dataset is given in the STAR Methods section along with the analysis, which is summarized in Figure S2.

We also applied ACC v2.0 to a drug screen and a small interfering (siRNA) screen to identify phenotypic classes using real data. In each case, the annotators were able to discover relevant phenotypic classes, including rare types, within a short time using our new methods. Details of the drug and the siRNA screens are provided in the STAR Methods section. Example images from the phenotypic classes identified using ACC v2.0 are shown in Figures 2 and S3.

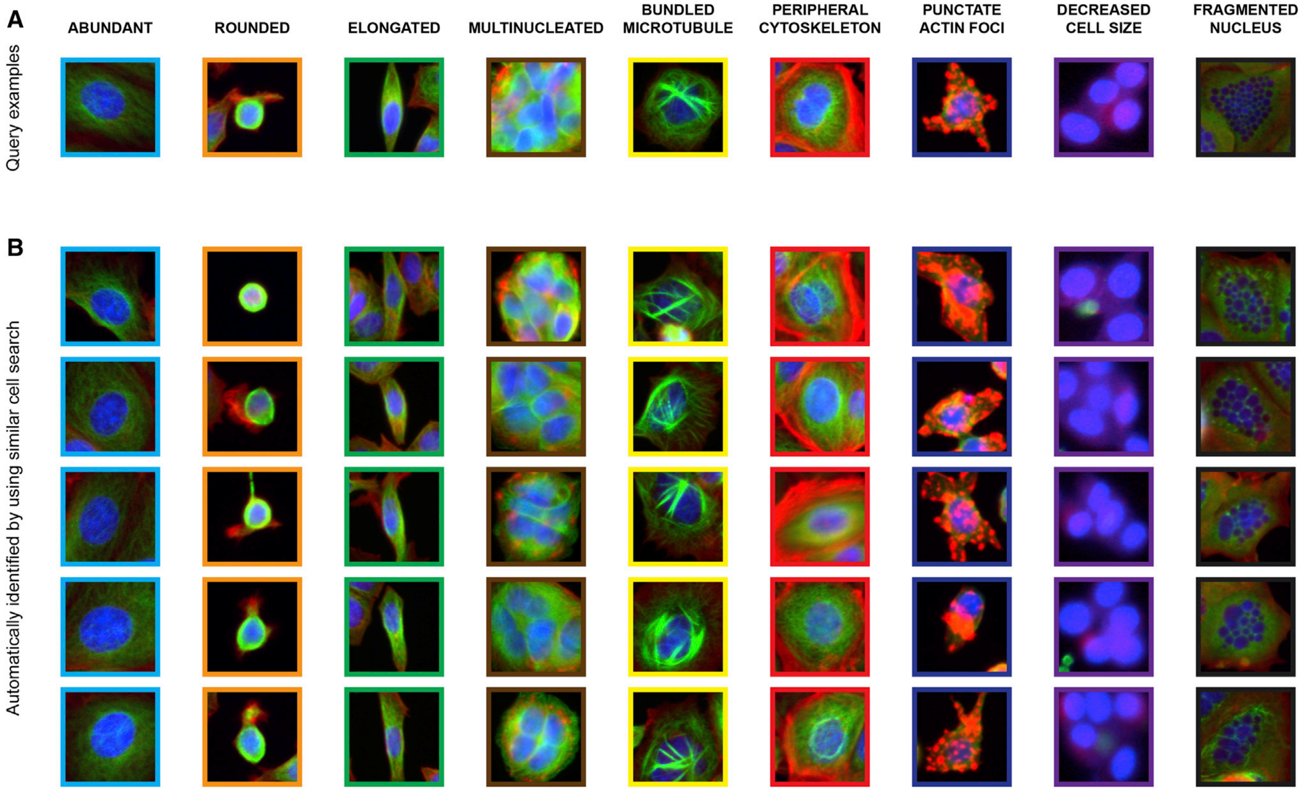


Figure 2. Efficient Example-Based Mining of Relevant Cell Phenotypes

(A) Cells from nine relevant cell phenotypes identified using the phenotype finder on data from the Broad Bioimage Benchmark Collection (Ljosa et al., 2012). Searching through thousands of images for more examples of a rare phenotype can be cumbersome, but the similar cell search function makes it simple.

(B) Results of the similar cell search given the query examples above. Similar cells were determined by computing the cosine similarity on image feature vectors between the query example and other cells in the dataset.

DISCUSSION

ACC v2.0 provides several new innovative tools designed to explore and collect the data necessary to train classifiers more efficiently and effectively. The ability of the classifier to correctly recognize cell types ultimately depend on the quality of data provided. While there is no universally accepted recipe for generating quality training data, many principles and techniques can be applied in practice to improve the efficiency and quality of the annotation process.

Perhaps the most fundamental principle is to ensure that the training data are complete in the sense that they includes examples of all the important phenotypes present in the screen. Although it may seem obvious, practically speaking, this can be tedious when the amount of data is very large. Another common issue is imbalance between classes. Often, interesting phenotypes are in the minority or occur very infrequently. If the data are imbalanced due to the presence of a rare class, the lack of representative data will make learning difficult (He and Garcia, 2009). Given a single example of a rare cell, ACC v2.0 can quickly identify additional, previously unidentified examples using the similar cell search feature, thereby helping balance the dataset and improve classification performance. Another way to improve data collection is to avoid redundant annotations and to prioritize annotations that are most useful for boosting classification performance. ACC v2.0 uses active learning to carefully select the most informative examples for labeling, which avoids irrelevant examples and refines regions where the classifier is uncertain.

Data quality can also be improved through iterative refinement of the classifier. During data collection, ACC v2.0 can train a classifier on existing annotated data and display predicted annotations on unlabeled data. By correcting erroneous predictions, the user adds valuable data points to the training set, which can help correct predictive errors.

In total, ACC v2.0 includes powerful new methods to mine microscopic image data, discover new phenotypes, and improve recognition performance. While no single method can be regarded as a silver bullet that solves all annotation problems, in our experience, the most effective strategy is to alternate between discovery tools as the biological task demands. ACC v2.0 gives the user access to a large variety of state-of-the-art machine-learning algorithms, has an intuitive user interface with advanced visualization, and allows for efficient navigation of image data. It is easy to use, well documented, and comes with helpful video tutorials. Using synthetic data and existing screens, we demonstrated that the discovery tools in ACC v2.0 improve the quality of training datasets and ultimately create classifiers with better phenotype recognition. Using our software, it is possible to discover interesting cell phenotypes hidden in large datasets.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Synthetic Dataset
  - ○ Drug HCS Dataset
  - ○ siRNA HCS Dataset
- METHOD DETAILS
  - ○ Overview of the Software
  - ○ Phenotype Finder
  - ○ Similar Cell Search
  - ○ Active Learning Methods to Prioritize Useful Annotations
  - ○ Manual Correction of Predictions
  - ○ Annotation of Manually Selected Cells
  - ○ Annotation of Randomly Selected Cells
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures, one table, four files, two data files, and five movies and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2017.05.012.

AUTHOR CONTRIBUTIONS

P.H. conceived and led the project. F.P. wrote the documentation, designed the GUI, prepared the figures, and co-supervised the project. U.K. and M.S. provided one of the HCS datasets and the knowledge behind it. P.H., K.S., T.B., V.P., and K.B. designed the experiments and analyzed the data. T.B. was the lead developer of the implementation and the phenotype finder module. A.S. implemented the active learning module. L.P and K.K. implemented the find similar cells module. K.S. implemented the suggest a classifier module. C.M. implemented the image visualization module and tested the code. F.P., K.S., T.B., K.B., V.P., and P.H. wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors wish to thank Beverley Isherwood, Neil Carragher, and Anne E. Carpenter for the information they provided about the HCS image dataset; ETH Zurich Microscopy Center (ScopeM) for HCS imaging; Arpad Balind for suggestions and feedback; Gabriella Tick and Máté Görbe for their help with the documentation; Dóra Bokor for proofreading the manuscript. F.P. was supported by a Short-Term Fellowship awarded by the Federation of European Biochemical Societies (FEBS). F.P., B.T., CS.M., A.SZ., and P.H. acknowledge support from the Hungarian National Brain Research Program (MTA-SE-NAP B-BIOMAG). U.K. received support from the SNF (313003A_166565). L.P., V.P., and P.H. acknowledge support from the Finnish TEKES FiDiPro Fellow Grant 40294/13 and European Union and the European Regional Development Funds (GINOP-2.3.2-15-2016-00026, GINOP-2.3.2-15-2016-00037).

Received: October 19, 2016
Revised: January 17, 2017
Accepted: May 19, 2017
Published: June 21, 2017

REFERENCES

Badertscher, L., Wild, T., Montellese, C., Alexander, L.T., Bammert, L., Sarazova, M., Stebler, M., Csucs, G., Mayer, T.U., Zamboni, N., et al. (2015). Genome-wide RNAi screening identifies protein modules required for 40S subunit synthesis in human cells. Cell Rep. *13*, 2879–2891.

Banerjee, I., Miyake, Y., Nobs, S.P., Schneider, C., Horvath, P., Kopf, M., Matthias, P., Helenius, A., and Yamauchi, Y. (2014). Influenza A virus uses the aggresome processing machinery for host cell entry. Science *346*, 473–477.

Caie, P.D., Walls, R.E., Ingleston-Orme, A., Daya, S., Houslay, T., Eagle, R., Roberts, M.E., and Carragher, N.O. (2010). High-content phenotypic profiling of drug response signatures across distinct cancer cells. Mol. Cancer Ther. *9*, 1913–1926.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100.

Cooper, J.A. (1987). Effects of cytochalasin and phalloidin on actin. J. Cell Biol. *105*, 1473–1478.

Dao, D., Fraser, A.N., Hung, J., Ljosa, V., Singh, S., and Carpenter, A.E. (2016). CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. Bioinformatics *32*, 3210–3212.

Fischer, U., Forster, M., Rinaldi, A., Risch, T., Sungalee, S., Warnatz, H.J., Bornhauser, B., Gombert, M., Kratsch, C., Stütz, A.M., et al. (2015). Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options. Nat. Genet. *47*, 1020–1029.

Fujikawa-Yamamoto, K., Teraoka, K., Zong, Z.P., Yamagishi, H., and Odashima, S. (1994). Apoptosis by demecolcine in V79 cells. Cell Struct. Funct. *19*, 391–396.

He, H., and Garcia, E.A. (2009). Learning from imbalanced data. IEEE Trans. Knowledge Data Eng. *21*, 1263–1284.

Held, M., Schmitz, M.H., Fischer, B., Walter, T., Neumann, B., Olma, M.H., Peter, M., Ellenberg, J., and Gerlich, D.W. (2010). CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. Nat. Methods *7*, 747–754.

Horvath, P., Wild, T., Kutay, U., and Csucs, G. (2011). Machine learning improves the precision and robustness of high-content screens using nonlinear multiparametric methods to analyze screening results. J. Biomol. Screen. *16*, 1059–1067.

Jones, T.R., Kang, I.H., Wheeler, D.B., Lindquist, R.A., Papallo, A., Sabatini, D.M., Golland, P., and Carpenter, A.E. (2008). CellProfiler Analyst: data exploration and analysis software for complex image-based screens. BMC Bioinformatics *9*, 482.

Jordan, M.A., and Wilson, L. (2004). Microtubules as a target for anticancer drugs. Nat. Rev. Cancer *4*, 253–265.

Kavallaris, M. (2010). Microtubules and resistance to tubulin-binding agents. Nat. Rev. Cancer *10*, 194–204.

Laksameethanasan, D., Tan, R.Z., Toh, G.W., and Loo, L.H. (2013). cellXpress: a fast and user-friendly software platform for profiling cellular phenotypes. BMC Bioinformatics *14* (*Suppl 16*), S4.

Laplante, M., and Sabatini, D.M. (2009). mTOR signaling at a glance. J. Cell Sci. *122*, 3589–3594.

Lehmussola, A., Ruusuvuori, P., Selinummi, J., Huttunen, H., and Yli-Harja, O. (2007). Computational framework for simulating fluorescence microscope images with cell populations. IEEE Trans. Med. Imaging *26*, 1010–1016.

Liu, L., Chen, L., Chung, J., and Huang, S. (2008). Rapamycin inhibits F-actin reorganization and phosphorylation of focal adhesion proteins. Oncogene *27*, 4998–5010.

Ljosa, V., Sokolnicki, K.L., and Carpenter, A.E. (2012). Annotated high-throughput microscopy image sets for validation. Nat. Methods *9*, 637.

Ljosa, V., Caie, P.D., Horst, R.T., Sokolnicki, K.L., Jenkins, E.L., Daya, S., Roberts, M.E., Jones, T.R., Singh, S., Genovesio, A., et al. (2013). Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. J. Biomol. Screen. *18*, 1321–1329.

MacLean-Fletcher, S., and Pollard, T.D. (1980). Mechanism of action of cytochalasin B on actin. Cell *20*, 329–341.

Misselwitz, B., Strittmatter, G., Periaswamy, B., Schlumberger, M.C., Rout, S., Horvath, P., Kozak, K., and Hardt, W.D. (2010). Enhanced CellClassifier: a multi-class classification tool for microscopy images. BMC Bioinformatics *11*, 30.

Mortensen, K., and Larsson, L. (2003). Effects of cytochalasin D on the actin cytoskeleton: association of neoformed actin aggregates with proteins involved in signaling and endocytosis. Cell. Mol. Life Sci. *60*, 1007–1012.

Moya, M.M., Koch, M.W., and Hostetler, L.D. (1993). One-class Classifier Networks for Target Recognition Applications (No. SAND-93-0084C) (Sandia National Labs), Technical Report.

Mühlradt, P.F., and Sasse, F. (1997). Epothilone B stabilizes microtubuli of macrophages like taxol without showing taxol-like endotoxin activity. Cancer Res. *57*, 3344–3346.

Neumann, B., Walter, T., Hériché, J.K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature *464*, 721–727.

Ogier, A., and Dorval, T. (2012). HCS-Analyzer: open source software for high-content screening data correction and analysis. Bioinformatics *28*, 1945–1946.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D.M., and Goldberg, I.G. (2008). WND-CHARM: multi-purpose image classification using compound image transforms. Pattern Recognit. Lett. *29*, 1684–1693.

Rämö, P., Sacher, R., Snijder, B., Begemann, B., and Pelkmans, L. (2009). CellClassifier: supervised learning of cellular phenotypes. Bioinformatics *25*, 3028–3030.

Schölkopf, B., Smola, A.J., Williamson, R.C., and Bartlett, P.L. (2000). New support vector algorithms. Neural Comput. *12*, 1207–1245.

Smith, K., and Horvath, P. (2014). Active learning strategies for phenotypic profiling of high-content screens. J. Biomol. Screen. *19*, 685–695.

Sommer, C., Straehle, C., Köthe, U., and Hamprecht, F.A. (2011). Ilastik: Interactive Learning and Segmentation Toolkit. Proceeding of the 2011 IEEE International Symposium on Biomedical Imaging: from Nano to Macro (ISBI 2011) (IEEE), pp. 230–233.

Uhlmann, V., Singh, S., and Carpenter, A.E. (2016). CP-CHARM: segmentation-free image classification made accessible. BMC Bioinformatics *17*, 51.

Witten, I., and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Advanced Cell Classifier version 2.0	Current manuscript	www.cellclassifier.org/download/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents or computational resources may be directed to, and will be fulfilled by, the Lead Contact Peter Horvath (horvath.peter@brc.mta.hu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Synthetic Dataset

To evaluate the effectiveness of the discovery and annotation tools included in ACC v2.0 (Figure S1), we designed and generated a synthetic dataset simulating images of cells from a high content screen. This provided us with completely accurate knowledge concerning every cell and phenotype in every image, something that is impossible with images of actual cells. Using this information, we compared ACC v1.0 to ACC v2.0 and evaluated how the tools we introduced improved the ability of annotators to discover phenotypes and train accurate classifiers.

The dataset was organized into 32 plates containing 384 images each. Every image contained human cells simulated using SIMCEP, a software tool for simulating fluorescence microscopy images of cell populations (Lehmussola et al., 2007). The amount of cells in each image was sampled uniformly between 5 and 40. In total, 12,288 images and 310,929 cells were generated. While this is a relatively small dataset by HCS standards, it provides sufficient statistical power to judge the efficacy of our tools. Eight distinct phenotypic classes were designed by varying the size and shape of the nucleus and cytoplasm, and by varying the number and size of subcellular vesicles (Figure S2A). Each cell is assigned a phenotype sampled from these classes, ranging in frequency from very common (appearing with 49.7% probability), to less common (24.8%), to extremely rare (0.01%). Furthermore, each image was generated with a dominant phenotype (approximately 80% of the cells in the image), and random cell classes made up the remainder of the phenotypes in the image. The synthetic images were processed with a CellProfiler pipeline to segment the cells and extract features. The synthetic data and pipeline are available in Data S1.

Three experts used ACC v1.0 and ACC v2.0 to annotate the dataset. They were given no prior information about the number of phenotypes or their locations. In each trial, the expert was given 30 minutes to annotate the synthetic data. The goals were (*1*) to discover all the phenotypes within the given time and (*2*) to create a high quality dataset which, when used to train a classifier, results in the highest accuracy on the whole dataset. The results of the experiment appear in Figure S2. All three experts were able to discover all eight phenotypes within the time limit using ACC v2.0 (Figure S2B). Using ACC v1.0, the three rarest phenotypes were only discovered by 2/3 of the experts. The experts were able to find phenotypes faster using ACC v2.0. Using the annotations collected by the experts, classifiers were trained and used to predict phenotypes for the entire dataset. The phenotypes predicted by trained classifiers were compared to the true phenotypes and normalized prediction accuracy was computed. The results show a substantial improvement (a 27% increase in recognition) with annotations collected using ACC v2.0 (Figure S2C). The last pane of the figure shows timelines of how many annotations were collected by each annotator using each version of the software (Figure S2D). The bold line shows the mean number of collected annotations.

Drug HCS Dataset

To demonstrate the capabilities of ACC v2.0 on real data, we analysed a dataset of MCF-7 breast cancer cells (BBBC021v1 (Caie et al., 2010), available from the Broad Bioimage Benchmark Collection (Ljosa et al., 2012): *https://www.broadinstitute.org/bbbc/BBBC021/*) treated with a collection of 113 small molecules in 8 different concentrations for 24 hours. A subset of this dataset is formatted for ACC and supplied as a test dataset with the software (*www.cellclassifier.org/download/*). The molecule set consists of a mechanistically distinct set of targeted and cancer-relevant cytotoxic compounds inducing a broad range of gross and subtle phenotypes. Cells were fixed, labelled for DNA, F-actin, and β-tubulin, and imaged by fluorescent microscopy acquiring multiple images of all data points and technical repeats (Caie et al., 2010). This resulted in 39,600 images containing approximately 2,000,000 cells.

The images were segmented and features were extracted with CellProfiler 2.2.0 (CellProfiler pipeline provided as File S1). To demonstrate the phenotype discovering capabilities of ACC v2.0, we started by manually annotating a few cells from a single class (standard abundant cells from the most common phenotype). We then used the phenotype finder tool to identify interesting cell phenotypes. Representative cells from each phenotypic class appear in Figure 2A. Names of the phenotypic classes were selected by using Cellular Microscopy Phenotype Ontology (CMPO; *http://www.ebi.ac.uk/cmpo/*), which provides a species-neutral controlled vocabulary for cellular phenotypes. To increase the number of annotations for rare classes such as multinucleated cells, we used the 'find similar cells' tool (Figure 2B). To improve the classifier's ability to distinguish between classes, we used the active learning module to refine decision boundaries. After approximately one hour we obtained around 1,500 annotated cells from 9 phenotype classes.

Finally, we assessed the predictive capability of the classifiers trained using the data we collected. We applied several different popular supervised classification methods and chose the one with best 10-fold cross validation performance. The Logistic Boost classifier achieved 88.4%, the highest recognition accuracy (Figure S4), and was applied to the entire screen. The final results of the analysis showing the number of cells in every phenotypic group for the different drug treatments can be found in Data S2.

We evaluated whether the identified phenotypes correlated to any known effects of the used drugs (Table S1). While most of the identified phenotypes were determined to be non-drug specific, we highlight two phenotypes showing strongly correlated characteristics to specific well-known drug effects: (*a*) multinucleated; and (*b*) bundled microtubule, cells with collapsed microtubule. The Cytochalasin B treatment leads to the classification of 20% of cells in the multinucleated group. Cytochalasin B is known to inhibit both the rate of actin polymerization and the interaction of the actin filaments in solution (MacLean-Fletcher and Pollard, 1980), thus preventing the formation of contractile microfilaments. This can result in a disturbed cell cycle, yielding an increased number of nuclei with variable shape and size within the cell (Cooper, 1987). The bundled microtubules group contains samples of a characteristic phenotype with microtubular bundles crossing through the centre of the cell. Most of the cells in this phenotypic class were treated with Taxol (Paclitaxel), Docetaxel, or Epothilone B. All of these agents bind the β-tubulin, and stabilize microtubules. They inhibit the microtubule function and alter their dynamics, as well as enhance the polymerization of tubulin, whereby they have antimitotic effect (Mühlradt and Sasse, 1997). In this case, the software identified a characteristic phenotype linked to different agents with the same mechanism of action.

We also identified phenotypes that correlate with well-known physiological cell status: (*1*) increased amount of punctate actin foci; and (*2*) fragmented nucleus, often the hallmark of apoptosis. The cells identified in the punctate actin class received Cytochalasin D as top hit. Cytochalasin D is an actin depolymerizing drug and can also induce the actin aggregation (Mortensen and Larsson, 2003). Interestingly, some of the cells treated with Rapamycin correlated strongly with this phenotype, while most of the Rapamycin–treated cells were classified to abundant or elongated cells classes. Rapamycin targets the mTOR (a mammalian target of Rapamycin-complex) with an essential role in the cell cycle and responses to changing nutrient levels. Inhibition of mTOR signalling by Rapamycin leads to defects in mitochondrial function, cell proliferation, cytoskeletal organization, protein synthesis, and can result in cell death in many ways: apoptosis, necrosis or autophagy (Laplante and Sabatini, 2009). Thanks to the new tools in ACC v2.0, we were also able to find the cells in a rarer actin-related phenotype of Rapamycin, reported to affect the F-actin reorganization by blocking the kinase activity of mTOR (Liu et al., 2008). Demecolcine (Colcemid) and Vincristine treatment resulted in an overrepresented number of cells of the phenotype fragmented nucleus. By inhibiting the polymerization of microtubules, they can arrest the cells in the metaphase, which can ultimately lead to apoptosis (Fujikawa-Yamamoto et al., 1994; Jordan and Wilson, 2004). These drugs are known to bind the tubulin and destabilize the microtubules, in contrast to the stabilizing agents, such as Docetaxel and Paclitaxel described above with a strong bundled microtubules phenotype (Kavallaris, 2010). This demonstrates the power of the ACC v2.0 to correctly categorize the drugs with opposite biological functions although they would have a potentially similar end-point outcome (cell death). As these results show, ACC v2.0 is an effective tool to find the novel, unknown biological effects of drugs or silencing/overexpression of certain genes, as well as for mining previously described phenotypes of interest.

siRNA HCS Dataset

We analysed images derived from a pilot siRNAs screen targeting selected hit factors as well as a suite of positive and negative controls for a genome-wide screen on 60S ribosomal subunit biogenesis in HeLa cells. The assay relies on an inducible RPL29-GFP reporter as a read-out, similar to our recent genome-wide analysis of 40S synthesis using RPS2-YFP (Badertscher et al., 2015). In brief, cells were transfected with the respective siRNAs by reverse transfection in 384 well plates. Reporter construct expression was induced by addition of tetracycline after 44 hours. 8 hours later, the culture medium was replaced by medium lacking tetracycline and cells were incubated for another 20 hours. Then, cells were fixed, DNA stained with Hoechst, and images taken by fluorescent microscopy acquiring 9 images per well.

Images were segmented based on Hoechst staining of cell nuclei, and 150 diverse features were extracted, including nuclear and cytoplasmic fluorescence intensities. We started by manually annotating a few cells from a single class of a mock-treated well. We then used the phenotype finder tool to identify interesting cell phenotypes. Representative cells from each phenotypic class appear in Figure S3A. To increase the number of annotations for rare classes such as large nuclei, we used the find similar cells tool (Figure S3B). We acquired ~500 annotated cells from 7 phenotype classes.

Finally, we assessed the predictive capability of the classifiers trained using the data we collected. We applied several different popular supervised classification methods and chose the one with best 10-fold cross validation performance. The Artificial Neural Network classifier achieved 95%, the highest recognition accuracy (Figure S5), and was applied to the entire screen.

METHOD DETAILS

Overview of the Software

ACC v2.0 is a software tool to apply sophisticated machine learning analysis to microscopic image data. Particular attention has been given to user-friendliness and tools to help non-experts explore their data, understand it, and train machine learning algorithms to be as accurate as possible. Figure S1A shows the main interface which consists of a menu bar, a toolbar, the main window, and cell view windows focusing on details of the currently highlighted cell. A separate image selector window allows the user to quickly switch between images and plates from the experiment (S1B). ACC v2.0 works in Windows, Linux, and Macintosh environments, and takes images as input with support for most common image formats (*e.g.* tif, bmp, png). In addition to the original image data, ACC v2.0 requires features, the image measurements extracted from the segmented objects (in txt, csv, or HDF5 format), and it supports contours of segmented sub-cellular/cellular objects (*e.g.* cells membranes and nuclei), such as those produced by CellProfiler (Carpenter et al., 2006).

Detailed documentation (File S2 and File S3) provides step-by-step instructions to use the software, and a series of video tutorials with how-to examples are also available (Movies S1–S5). A detailed flowchart diagram of the usage and services is presented on Figure S6. Typically, analysis begins with the user starting a project (‘‘Getting started’’, Movie S1) and importing data (‘‘ACC and CellProfiler’’, Movie S2, and ‘‘Input data structure’’, Movie S3). We provide a CellProfiler module designed to export necessary segmentation and feature data directly to ACC v2.0 (File S4). The next step is to begin exploring the data and assigning annotations to cells as members of a phenotypic class. Instructions on how to define classes, customize the visualization, and navigate the data are provided in Movie S1. Several innovative annotation strategies are available in ACC v2.0 which greatly improve the efficiency and quality of data annotation. These include the phenotype finder (Figure S1C, Figure 1), similar cell search, active learning, as well as manual and random selection. Examples on how to apply these strategies are shown in Movie S4. Once the annotated data is collected, up to sixteen different classifiers may be chosen to predict phenotypes for unannotated cells (Figure S1E) including well-known methods such as support vector machine (SVM), multilayer perceptron (MLP), and random forest using the Weka framework (Witten and Frank, 2005). Initial predictions may contain errors, but these can be corrected by replacing the incorrect prediction with the correct annotation and re-training the classifier. This process can be repeated until satisfactory results are achieved. Finally, the improved classifier is applied to the entire experiment and user selects the desired formats of the output including (*a*) cell-by-cell classification; (*b*) incidence and distribution of the different classes; (*c*) phenotype-based statistics of any selected features (Movie S5).

Phenotype Finder

Screening datasets often consist of tens of thousands of images, or orders of magnitude more. The amount of data is often so substantial that it exceeds the capabilities of a human expert to observe everything. Therefore, it is difficult for the annotator to know whether the training data he/she collected contains examples representing all the important phenotypes present in the screen. To address this, ACC v2.0 includes a phenotype finder tool which helps find and define new phenotypes efficiently, without requiring a priori knowledge about the underlying dataset (Figure 1). It does this by hierarchically grouping cells based on their appearance. Images of the cells are organized into a browsable tree-like structure (Figure 1C) which allows the user to quickly identify previously unseen phenotypes or subpopulations within a known phenotype (Movie S4). The phenotype finder operates on the assumption that true phenotypes will cluster together in a space of relevant image features. Using a bottom-up approach, a dendrogram can be constructed by first defining each observation (cell) as its own cluster, and by making pairwise similarity comparisons between each cluster. The dendrogram is constructed by greedily merging the most similar clusters first, and proceeding up the hierarchy until the entire set belongs to a single cluster (Figure S1D). In ACC v2.0, the similarity metric is defined as the Euclidean distance between the mean of the feature vectors associated with cells belonging to each cluster.

The computational complexity and memory consumption associated with hierarchical clustering means it cannot be directly applied to very large data sets (in practice, 25,000-50,000 cells can be clustered on a standard computer with 16 GB memory in 15-80 s). Because of this limitation, unless the dataset is small, the phenotype finder must be applied to subsets of the data - cells belonging to a known phenotype or cells that are hypothesized to belong to an undiscovered phenotype. In the first case, existing annotations are used to train a classifier that predicts which cells belong to a selected known phenotype. These are sampled and clustered using hierarchical clustering. In the second case, a one-class classifier is used to identify cells that do not fall within any currently known phenotype. One-class classification, also known as unary classification (Moya et al., 1993), is a method that can perform this task by estimating the support of a high-dimensional distribution given a set of positive samples (Figures 1A and 1B). A one-class SVM (Schölkopf et al., 2000) predicts the probability that a given cell belongs to any of the known phenotypes in this manner. Cells are sorted based on this probability, and the cells least likely to belong to an existing phenotype are clustered using hierarchical clustering (Figure 1C). The one-class SVM can be sensitive to the features it is provided. Oftentimes, HCS data contains some redundant features or features with little discriminative power. Feature selection methods can help by reducing the feature vector to the most informative features. We have tested three methods on synthetic and real HCS data: information gain, principal component analysis (PCA), and factor analysis. We found that information gain proves to be the most useful method, and the optimal number of selected features is in the range of 10-20 (Figure S7).

Once a subset of cells has been selected and the dendrogram has been constructed, it is displayed as a collapsible tree interface which allows the user to view the member cells for each cluster in the hierarchy and to quickly create new annotations from a cluster in the dendrogram (Figure 1C). If the user determines that the cluster shows signs of novelty it can be annotated as a new class, otherwise it can be inserted into an existing phenotype. If a new phenotype is discovered, the class boundaries will change (Figure 1D). Iteratively applying the phenotype finder in this manner will allow the user to quickly identify rare phenotypes or subpopulations within an existing phenotype.

Similar Cell Search

Often, important or interesting phenotypes occur very infrequently. While the phenotype finder can help initially discover the phenotype, without sufficient examples of the phenotype to train with, the classifier may struggle to identify it reliably. ACC v2.0 offers a solution to this issue. Given a single example of a rare cell, it can quickly identify new examples with similar appearance to the selected cell. This is accomplished by computing the cosine similarity on a reduced feature vector between the query cell and all other cells in the screen (Ljosa et al., 2013). The feature set is reduced in a pre-processing step to remove highly correlated features. The cells are sorted according to their similarity score, and the most similar cells are presented to the user (Figures 2B and S3B). The user may then select these cells and add them to the corresponding classes. In this manner, rare and important phenotypes can be quickly populated with reliable annotations.

Active Learning Methods to Prioritize Useful Annotations

Expert annotations are costly to acquire, but typically no guidance is provided as to which cells are useful to annotate. The user is often left to select cells manually, or is presented with a random sampling of cells. But there is no guarantee on the usefulness of the labels provided with these strategies. It can be advantageous to avoid redundant annotations and prioritize annotations that will be most useful for increasing classification performance. ACC v2.0 uses active learning to carefully select the cell which, if annotated, will be most useful in improving the classification accuracy. It avoids uninformative examples and refines regions where the classifier is uncertain. Using the currently annotated set (Figure S8A), a classifier is trained and makes predictions on unlabeled data and uses a query strategy to estimate which example will be most useful to improve classification performance (Figure S8B). We use two popular query strategies: uncertainty sampling and query by committee (Smith and Horvath, 2014). Uncertainty sampling selects the instance it is least certain how to label based on the classifier prediction probabilities. Query by committee selects the instance with the most disagreement between different classification algorithms. The expert annotates the requested cell, and the annotation is added to the training set. A new predictive model is trained and a new cell is selected for annotation according to the query strategy (Figure S8C). This process is repeated and iteratively improves classification performance. Potentially, new phenotypes can be discovered by exploring the boundary between classes (Figures S8D–S8E).

Manual Correction of Predictions

Another practical method of improving the quality of training data is to iteratively refine classifier performance by correcting errors in its predictions. ACC v2.0 can train a classifier on existing annotated data and visualize predicted annotations on unlabeled data. The user can visually inspect these predictions and add new annotations by confirming correct labels or by correcting erroneous predictions. By adding corrective annotations where the classifier made a mistake, the user adds valuable data points to the training set to improve performance.

Annotation of Manually Selected Cells

This annotation method is standard in most cell classification software. The expert is free to navigate the data and select which cells he/she wishes to annotate. Manual exploration has its merits, as the expert may often have a good intuition about which cells are useful to annotate early in process. However, there are several dangers to relying solely on this method. The user may be biased towards easy or redundant examples, may not balance the number of examples between classes, and may fail to discover rare phenotypes. When performing manual annotation, it is recommended that cells are selected from several different images. We also note that is it always possible to override other annotation modes and switch to a manual annotation if the user notices something interesting in the image.

Annotation of Randomly Selected Cells

To ensure a representative sampling of the data, it can be advantageous to randomly sample from the screen. This is a forced choice mode: the expert is required to annotate the phenotype class of a randomly selected cell from a randomly chosen image. This annotation mode thereby avoids selection bias by the user (Misselwitz et al., 2010). However, with this method and with manual annotation, luck may be required to discover rare phenotypes in large datasets.

DATA AND SOFTWARE AVAILABILITY

In this work we used the HCS image set BBBC021v1 (Caie et al., 2010), available from the Broad Bioimage Benchmark Collection (Ljosa et al., 2012). It is composed of 39,600 image files (13,200 fields of view imaged in three channels) in TIFF format. It can be downloaded at: *https://www.broadinstitute.org/bbbc/BBBC021/*. For any copyright issues regarding the dataset, follow the instructions provided at the referenced website.

ACC v2.0 is freely distributed at *www.cellclassifier.org* as an open-source tool. It is written in MATLAB R2015a and requires the Image Processing Toolbox 9.2. The source code, standalone versions, video tutorials, help documentation files, and additional modules are available at the official repository for all future software releases *www.cellclassifier.org*. Further developing the source code requires Matlab license. However, Windows, Linux, and Macintosh standalone compiled versions— that do not require license— are also available online. All the ACC materials are copyright protected and distributed under GNU General Public License version 3 (GPLv3).

nature COMMUNICATIONS

horvath.peter.2_10_22

ARTICLE

https://doi.org/10.1038/s41467-021-22866-x OPEN

Regression plane concept for analysing continuous cellular processes with machine learning

Abel Szkalisity[1,2], Filippo Piccinini[3], Attila Beleon[1], Tamas Balassa[1], Istvan Gergely Varga[4], Ede Migh[1], Csaba Molnar[1], Lassi Paavolainen[5], Sanna Timonen[5], Indranil Banerjee[6], Elina Ikonen[2], Yohei Yamauchi[7], Istvan Ando[4], Jaakko Peltonen[8,9], Vilja Pietiäinen[5], Viktor Honti[4] & Peter Horvath[1,5,10]✉

Biological processes are inherently continuous, and the chance of phenotypic discovery is significantly restricted by discretising them. Using multi-parametric active regression we introduce the Regression Plane (RP), a user-friendly discovery tool enabling class-free phenotypic supervised machine learning, to describe and explore biological data in a continuous manner. First, we compare traditional classification with regression in a simulated experimental setup. Second, we use our framework to identify genes involved in regulating triglyceride levels in human cells. Subsequently, we analyse a time-lapse dataset on mitosis to demonstrate that the proposed methodology is capable of modelling complex processes at infinite resolution. Finally, we show that hemocyte differentiation in Drosophila melanogaster has continuous characteristics.

[1] Synthetic and Systems Biology Unit, Biological Research Centre (BRC), Szeged, Hungary. [2] Department of Anatomy and Stem Cells and Metabolism Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [3] Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) IRCCS, Meldola, FC, Italy. [4] Institute of Genetics, Biological Research Center (BRC), Szeged, Hungary. [5] Institute for Molecular Medicine Finland-FIMM, Helsinki Institute of Life Science-HiLIFE, University of Helsinki, Helsinki, Finland. [6] Indian Institute of Science Education and Research (IISER), Mohali, India. [7] School of Cellular and Molecular Medicine, University of Bristol, BS8 1TD University Walk, Bristol, UK. [8] Faculty of Information Technology and Communication Sciences, Tampere University, FI-33014 Tampere University, Tampere, Finland. [9] Department of Computer Science, Aalto University, Aalto, Finland. [10] Single-Cell Technologies Ltd., Szeged, Hungary. ✉email: horvath.peter@brc.hu

Large-scale imaging technologies, such as high-content screening (HCS) and digital pathology imaging, have become the de facto tools for discovering drugs and genes and understanding tissue physiologies and pathologies, including cancer heterogeneity. This has induced a rapid growth in the amount of microscopy data, making it essential to elaborate appropriate bioinformatics tools to analyze them, and thus improve the current understanding of underlying biological processes[1–3].

Machine learning provides automation for analyzing big data, such as that acquired in large-scale, image-based experiments, and it has been successfully utilized for phenotypic analysis tasks[4]. Although a great variety of software tools are available for performing imaging assays in a supervised manner (e.g. CellProfiler Analyst, Ilastik, CellCognition, Advanced Cell Classifier[5]), all of them rely on the assumption that the underlying biological processes have stable steady states that can be dissected into discrete phenotypic classes (Fig. 1a). However, biological processes inherently contain continuous transitions between these phenotypes, consequently restricting the modelling to a set of discrete states reduces the potential to fully understand biological phenomena.

The application of traditional classification models for single-cell image analysis[6–8] is especially unreliable when the cells of interest change their morphological features gradually in the course of time (e.g. cell cycle). Annotation of such data is error-prone and laborious, and even field experts tend to make faulty decisions (e.g. in the case of samples with interclass properties), often leading to arbitrary labelling. Additionally, user-defined classes may obscure the real underlying distribution by inappropriate discretization.

Currently, none of the available and widely used software tools enable single-cell-based image analysis in a continuous, supervised manner. Instead, unsupervised models, such as Lineage Reconstruction Techniques (LRT)[9] and Dynamic Time Warping (DTW) prevail. Cycler[8] is an LRT and embeds 5 pre-selected image-based single-cell features to a one-dimensional (1D) continuous space called the cell-cycle trajectory. Similarly, Cai et al. used DTW to align mitotic cells into the mitotic standard time based on 6 selected features[10]. HipDynamics is a software for visualizing cell population dynamics in live-cell imaging data and it utilizes unsupervised linear regression to characterize the changes in user-selected features[11]. Indeed, these tools provide robust solutions for their targeted tasks, but the lack of expert interaction significantly reduces the potential to customize these methods for various purposes. Therefore, another set of tools known as Visual Analytics (VA) was developed, offering various techniques for experts to interactively change the machine learning model through a visualization interface, which is most often a continuous space (visualization map)[12,13]. CellCognition was a pioneer of supervised tools, designed with the intent to efficiently analyze biological processes, however still using classification[7].

Here, we propose a methodology called Regression Plane (RP), an interface for fully supervised, continuous machine learning

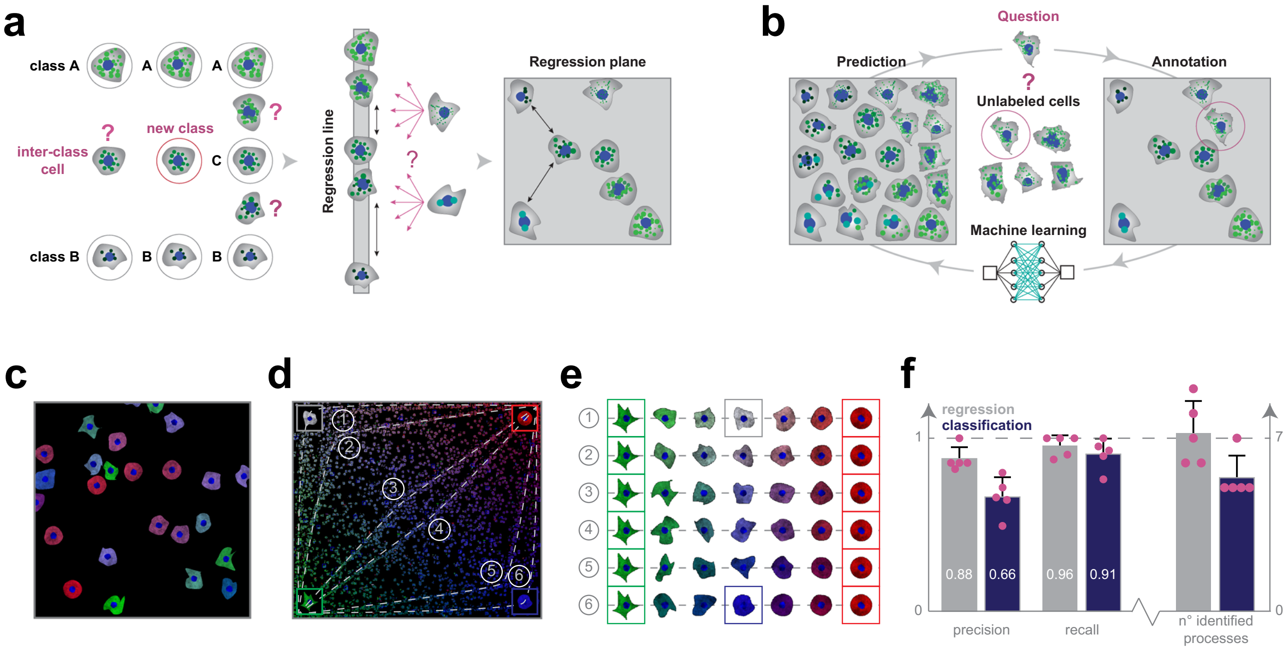


Fig. 1 Classification vs regression. a Regression plane concept. The classical way to model a biological process includes the phenotypical analysis of cells (i.e. subdividing cells into classes). However, in a high-content screening scenario, the multitude of different phenotypes makes it extremely challenging to create a set of representative classes. A possible solution builds on using a regression line, allowing to represent a single effect without the need of discretization. Nonetheless, biological processes are typically characterized by numerous ongoing effects. Thus, the regression plane represents a good trade-off between visualization capabilities and annotation complexity. Basically, it allows to represent a biological process with the limits of a planar graph. **b** Active regression. The aim of an active regression algorithm is to improve the training set (TS) to achieve better prediction performance. It is an iterative process where a cell that is difficult to annotate is proposed to the oracle who annotates it, and by doing so moves it to the TS used to train the regression model. **c** Synthetic dataset. Image from the synthetic dataset, generated using SIMCEP. **d** Experimental design. The designed processes overlayed on the space of perturbations. 6 processes are tracks in the space, and an extra process is formed of uniformly distributed cells (latent process 7). **e** Designed processes. The 6 continuous processes are modelled between two fixed endpoints: green cells of highly irregular shape and red, rounded cells. To assign a colour to the middle point of each process we interpolated between white (process 1) and blue (process 6). **f** Classification vs regression applied on synthetic data. Comparison of the performance of regression and classification. Statistics: precision, recall and the number of identified processes. Columns represent mean, error bars show the standard deviations from $n = 5$ independent users/experimental setup. Source data are provided as a Source Data file.

appropriate for image-based single-cell analysis. The idea originates from a study of an influenza A virus entry in which histone deacetylase-mediated reorganization of the microtubules led to various endosomal morphological and trafficking phenotypes that affected influenza infection[14]. The scatteredness of late endosomes and lysosomes (single output variable) was determined using regression instead of classification. Restricting the output to a single dimension prohibited the modelling of branching, circulating, parallel and crossing processes, therefore we extended the approach to utilize a 2D plane (Fig. 1a). Considering cellular steady states as graph nodes and gradual changes between the states as edges, the biological systems that correspond to planar graphs can be modelled with RP. Further extension of the modelling to 3D would increase the complexity of labelling and raise the chance of annotation errors. Additionally, to improve the quality of the annotated sets and decrease the time required from experts, we have incorporated active learning methods appropriate for regression-based phenotyping.

Results

Regression plane. Regression plane is implemented as an open-source module of Advanced Cell Classifier (ACC)[6], and it has been available since ACC v3.0. RP was incorporated into traditional phenotypic classification in a hierarchical manner: each class may be extended with a distinct regression plane, allowing multiple regression planes to be included in a single project. RP is easy to use, well documented and supported by video tutorials (Supplementary Software 1, Supplementary Movies 1, 2). Annotation is performed by assigning continuous labels to representative cells via placing them on a 2D plane. After training, RP predicts the position of every unlabelled cell and outputs versatile and easy-to-read visual representations at single-cell, population and treatment levels (for details see the Methods section).

Similarly to classification, a representative Training Set (TS) is also essential for RP. Active learning algorithms are routinely used in classification to find the most efficient TS[15] but are not widely used in regression[16]. In this work, we introduce various active regression algorithms by extending those used in classical active learning tasks (Fig. 1b, Supplementary Fig. 1a). These methods propose cells whose automatic prediction on the regression plane is uncertain or ambiguous. Details are reported in the Methods section.

Synthetic experiment: classification vs regression. To analyze data discovery capabilities of RP, we generated a synthetic HCS image dataset simulating perturbations of cell shape and protein expression (Fig. 1c, Supplementary Software 2). We designed gradual perturbations to enable smooth transition between cell states and hence facilitating the modelling of biological processes. We defined 6 processes as continuous changes from one cell state to another, plus an extra process (latent process 7) formed from uniformly distributed cells (Fig. 1d, e). Each well in the HCS plate was associated with an underlying process, and the corresponding images were generated by sampling cells uniformly from the process distribution. Based on the associated processes we defined a partitioning on the wells (those wells were in the same partition that had the same process associated to them), forming the ground-truth of our experiment. Details about the modelled biological processes are reported in the Methods section.

Subsequently, ten microscopy experts were divided into two groups and asked to identify the distinct underlying processes in the experiment (Supplementary Note 1), or equivalently to define a partitioning on the wells. The first group of five experts used ACC v2.1 (extended with Supplementary Software 3 to compensate for the advanced clustering features available in RP) to annotate cells with discrete labels, while the other group used RP only (ACC v3.0). Despite the great variety of the regression planes created by the microscopists (Supplementary Fig. 2), the results obtained using RP significantly outperformed the classification, both in terms of precision and recall (Fig. 1f). Specifically, the experts using RP performed better in estimating the number of ongoing processes, and achieved, on average, an improvement of approximately 20% in precision and 5% in recall, upon defining image sets containing cells with similar behaviour.

Lipid droplet study. Lipid droplets are storage units for neutral lipids, including triglycerides, and play a significant role in several disorders, including e.g. cardiovascular diseases. We evaluated whether siRNA perturbations of candidate genes, previously revealed to influence blood triglyceride (TG) levels in humans in a genome-wide association study[17], would affect the morphology of lipid droplets (LDs) in cultured hepatocytes (Huh7 cell line). Regarding their continuous changes in localization, number and size, LDs form a heterogeneous population reflecting different cellular metabolic states[18]. Thus, RP was used for the analysis of lipid droplets labelled with LipidToxGreen (Supplementary Fig. 3a–c), a probe used for quantitative analysis of neutral lipids. To train the model, 457 cells were placed on the regression plane by a cell biology expert (Fig. 2a).

We found that siRNA-mediated knockdown of TM6SF2 (a gene associated with decreased blood TGs) led to increased intracellular staining of neutral lipids, as it had been expected from the earlier evidence of TM6SF2 affecting hepatic lipid droplet content and TG secretion[19]. In contrast, the cells transfected with siRNAs targeting CD300LG (a gene associated with increased blood TGs[17]) showed a decreased amount of intracellular TGs, accompanied by the disappearance of (larger) LDs. Additional biochemical analysis measuring cellular TG levels confirmed these findings (Supplementary Fig. 3d). These data provide functional evidence for the role of CD300LG in regulating TG metabolism in hepatocytes.

Intriguingly, the knockdown of TM4SF5 (a gene associated with decreased blood TGs[17]) which codes for a protein functioning as an arginine sensor and mTORC1 regulator on lysosomal membranes[20], not shown earlier to affect TG levels in functional studies, promoted the increase of small LDs (Fig. 2b). Meta-visualization and clustering of the regression planes (Fig. 2c, Supplementary Fig. 3e–h) further supplemented the findings from an earlier study[17], and suggest that CD300LG and TM4SF5 may have biological effects on hepatic TG levels and LD composition, to be further addressed in future studies. Details are reported in Methods section.

Time-lapse microscopy: cell cycle analysis. We tested the capabilities of RP on 2 different time-resolved datasets. First, RP has been demonstrated to be capable of reproducing an unsupervised mitotic time model developed in the MitoCheck project (www.mitocheck.org).

Cai et al.[10] analyzed cell mitosis by performing time-lapse experiments to establish a canonical model for the morphological changes appearing during the mitotic progression of human cells. They reorganized the feature space according to the mitotic standard time instead of the imaging time (see Fig. 1f in ref. [10]), and by applying an unbiased peak-detection method in the warped feature space they identified up to 20 mitotic stages. The model was then used to integrate dynamic concentration data of several fluorescently labelled mitotic proteins, and to create a generic dynamic protein atlas of human cell division. Their public data include 3D images and segmented masks of 31 z-stacks. We intended to analyze this dataset without using prior feature

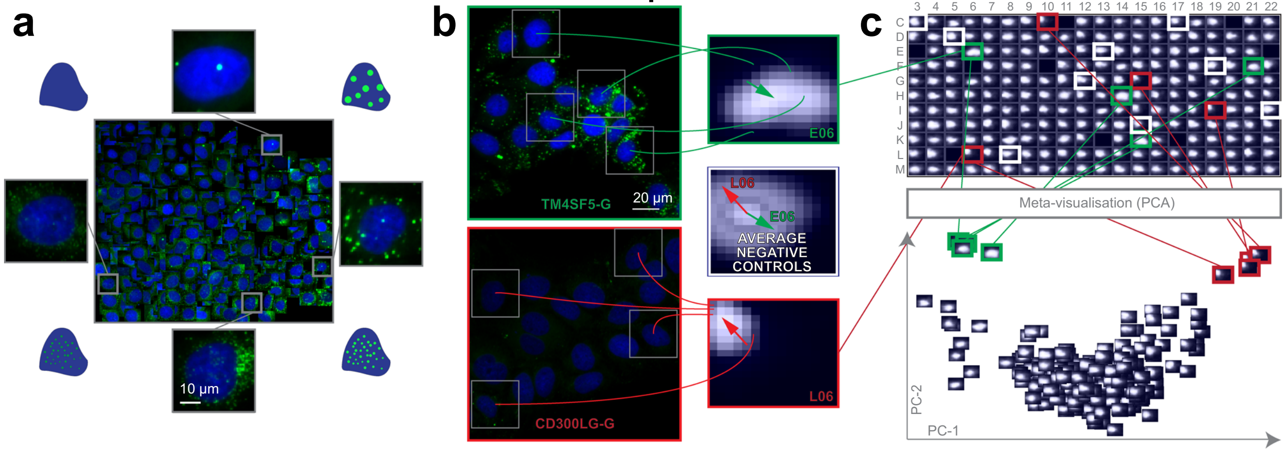


Fig. 2 Lipid droplet dataset. a Training set. Regression plane of 457 cells representing various lipid morphologies, created by an expert biologist. **b** RP output. Kernel Density Estimation (KDE)-maps of the predicted regression positions for cells treated with selected siRNAs. Arrows originate from the peak of the control KDE-map, and point to the peaks of the selected KDE-maps. **c** HCS analysis. Plate-based analysis performed by comparing well-based KDE-maps. Meta-visualization (in this case PCA–Principal Component Analysis) is obtained by extracting the principal components (PC1 and PC2) of the flattened KDE-maps.

information about the underlying process by exploiting regression techniques to characterize mitosis.

In our analysis, a field expert created a regression plane representing the process of mitosis, resulting in a training set of 585 cells (Fig. 3a). After prediction, the cells followed the designed circular path recalling canonical mitotic phases (Fig. 3b–d), while they also represented subtle phenotypic changes and single-cell differences in the regression plane. Additionally, we investigated whether the fluorescent tags have effect on the distribution of cells on the regression plane, and in most cases we did not observe undesired cellular behaviour due to the perturbations (Supplementary Fig. 4). Finally, we compared the results of the original methodology presented by Cai et al. (multi-dimensional dynamic time warping for creating the standard mitotic time, Fig. 3e) with the results obtained by RP (Fig. 3f), and we concluded that RP is capable of reproducing a mitotic time model equivalent to the original one. This indicates that RP can compete with complex analysis techniques, such as DTW. Moreover, RP provides the flexibility to customize the output space, enabling higher resolution analysis of user-defined sections of the biological process.

Time-lapse microscopy: blood cell differentiation. The fruit fly, *Drosophila melanogaster*, serves as a popular model system to study innate immune functions, such as phagocytosis, wound healing and capsule formation[21]. In the larva, these functions are executed by hemocytes, which are categorized into three main cell types: (1) phagocytic plasmatocytes, accounting for the majority of circulating hemocytes, (2) crystal cells, which play a role in melanization and wound healing, and (3) lamellocytes, which are large flat cells that appear only in certain tumorous genetic backgrounds or following immune induction[22]. Such an immune induction appears in nature as a result of egg-laying by a parasitoid wasp, *Leptopilina boulardi*. Following infestation, newly differentiating lamellocytes, together with plasmatocytes, eliminate the invader by forming a multilayer capsule around the wasp egg[23–25]. Lamellocytes are also produced when larvae are wounded with an insect pin[26] (Supplementary Fig. 5).

Cell lineage-tracing studies revealed that plasmatocytes, which had previously been considered as terminally differentiated phagocytic cells, show plasticity, and are capable of differentiating into encapsulating lamellocytes upon immune induction[22,27–29]. This transdifferentiation process has been underlined by recent single-cell RNA sequencing studies[30,31]. However, the cellular intermediates of the plasmatocyte-lamellocyte transition process have not been characterized morphologically in detail so far, and the routes of differentiation are still controversial[32]. A study by Anderl et al.[33] described two types of lamellocytes, and suggested that only the smaller type II lamellocytes (Supplementary Movie 3) differentiate from plasmatocytes, while the regular, flattened type I lamellocytes (Supplementary Movie 4) originate from dedicated precursors.

To clarify the potential routes of differentiation, we developed an ex vivo method for culturing Drosophila hemocytes, appropriate for monitoring their differentiation with time-lapse microscopy. Blood cell types can be characterized by their morphologies and in vivo transgenic reporter expression pattern[33]. The regression plane was manually trained using 109 cells based on their morphology and reporter gene expression (Fig. 4a).

The analysis revealed that 5.6% of the plasmatocytes transdifferentiated into lamellocytes upon immune induction (wounding) of the larvae (the threshold line is indicated in Fig. 4c), which is well reflected by the expression of cell type specific transgenes (Supplementary Movies 5, 6). After the formation of lamellocytes, no significant alterations in their cell size were observed, indicating that all types of lamellocytes are terminally differentiated cells. Most of the plasmatocytes (94.4%), however, did not differentiate into lamellocytes, but either spread out, increasing their cell size, or kept their size and morphology during the experiment, which is in line with the results of in vivo studies on blood cell differentiation in Drosophila.

However, in the case of lamellocytes instead of identifying 2 clearly separated subtypes I and II, we have observed that the differentiation processes are evenly distributed on the regression plane, as reflected by specific features (Fig. 4b, c, f). This finding suggests that type I and type II lamellocytes, both differentiating from plasmatocytes, are not definitely distinguishable cell types, but rather they are two extreme stages of a size continuum (Fig. 4e). Details are reported in the Methods section.

Discussion

Regression plane increases the resolution of classification to represent subtle phenotypic differences by exploiting regression techniques, extended by active learning. First, using artificial

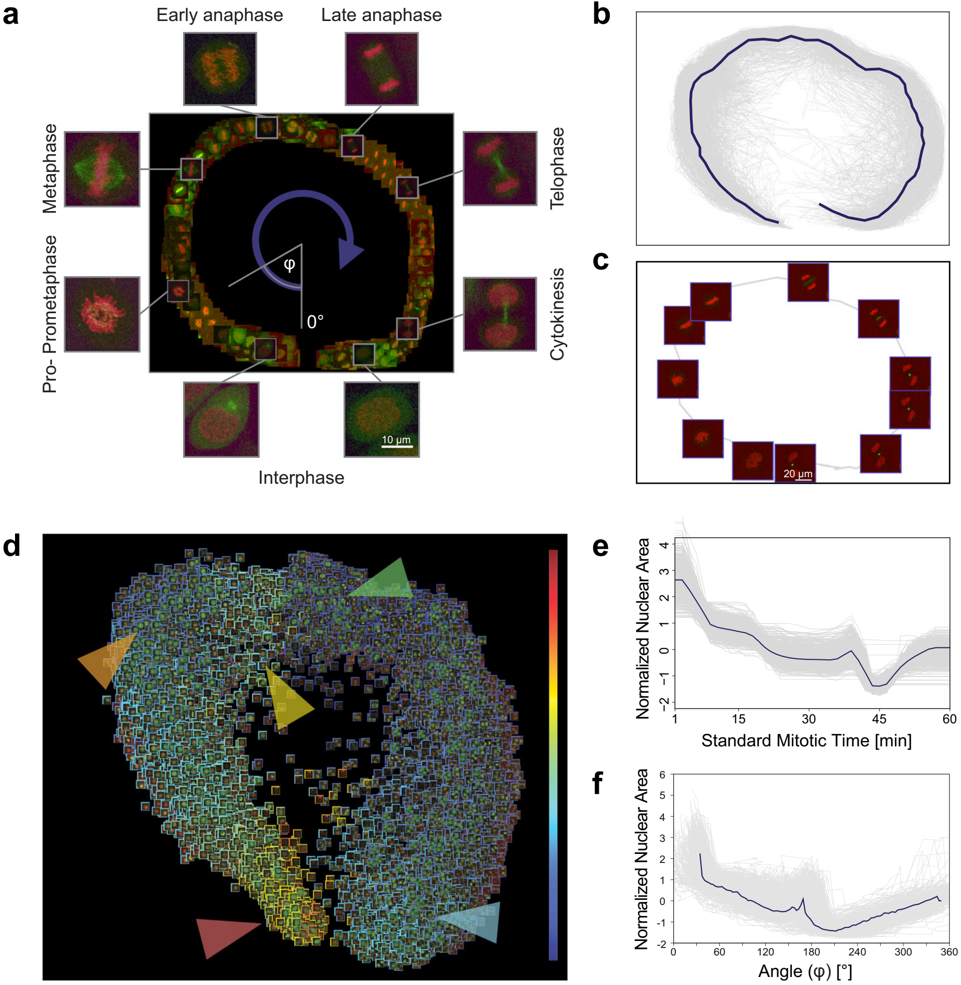


Fig. 3 Mitosis data analysis. a Regression plane of 585 cells annotated by a microscopy expert. **b** 498 trajectories for all the predicted cells. The median curve is shown in solid blue. **c** Example of a single-cell trajectory with representative cell icons visualized. **d** Regression plane with all ($n = 19{,}920$) predicted cells. The borders of the cell icons correspond to their nuclear area (Colour Frame module). Highlighted regions: early prophase region, large nuclear area (red). Metaphase region, nuclear area decreased (orange). Early-anaphase region, nuclear area is increasing as spindle fibres are pulling chromosomes apart (yellow). Anaphase, nuclear area dropped as the nucleus is considered as two separate objects with half the area (green). Late-telophase, nuclear area increasing up to half of the initial value (blue). **e** Trend for the normalized nuclear area according to standard mitotic time. Grey lines represent single-cell trajectories. **f** Trend for the normalized nuclear area according to the regression plane. Grey lines represent single-cell trajectories. The coordinates predicted by RP were converted to 1D by taking the angle argument of the polar coordinate representation as illustrated in **a**.

datasets we have demonstrated its capability to outperform the available classification tools in phenotypic discovery. Second, we have applied RP to analyze lipid droplets in hepatocytes during siRNA-mediated gene silencing, serving as a model of a heterogeneous population that reflects different cellular metabolic states. We have revealed genes playing a crucial role in regulating triglyceride levels in hepatocytes. Finally, we have identified the previously undiscovered continuous characteristics of hemocyte differentiation in *Drosophila melanogaster*. Our findings indicate that RP is a promising tool to explore biological data in a continuous manner, reflecting the non-discrete nature of biological processes.

Methods

Synthetic dataset. To generate the dataset we used a customized version of SIMCEP[34], provided as Supplementary Software 2. Synthetic microscopy images were organized into a 24-well plate format, and the dataset was composed of 9 images/well, for a total of 216 images and 8117 cells. The images of each well were generated by considering a predominant process mixed with other ones. To model the continuous processes we fixed two endpoints: green cells of highly irregular shape, and red, rounded cells (Fig. 1e). The degree of cell shape deformation decreases from the green to the red endpoint. Next, for each process we selected a middle point, and assigned a colour to that, ranging from white (process 1) to blue (process 6). The colour of the cells in each process was then defined by linear interpolation between the colour of the middle point and one of the two endpoints. The generated dataset was deposited to the Broad Bioimage Benchmark Collection (BBBC), and it is freely available at: https://data.broadinstitute.org/bbbc/image_sets.html (dataset ID: BBBC031).

Lipid droplet dataset. Huh7 hepatocellular carcinoma cell line (from Prof. Ilkka Julkunen, THL, Finland[35]) was authenticated using Promega StemElite^{TM} ID System at Genomics Unit of Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki. The cells were cultured in Minimum Essential Medium (MEM, Gibco® Life Technologies) supplemented with 10% FBS (fetal bovine serum, Gibco® Life Technologies), 100 IU/ml Penicillin and 100 μg/ml Streptomycin (Penicillin/Streptomycin combination, Gibco® Life Technologies) at 37 °C incubator with 5% CO_2. siRNAs (Supplementary Data 1) were transferred from source plates (Echo Qualified 384-Well Low Dead Volume Microplate, 384LDV, Labcyte) to the assay plates (384 -Well Flat Clear Bottom Black Polystyrene TC-Treated Microplates, Corning®, USA) in a final concentration of 10 nM with Echo® 550 Liquid Handler (Labcyte, UK) and Echo Cherry Pick software (version 1.4.4). 25 nl/well of transfection reagent Lipofectamine RNAiMAX (Invitrogen, Life Technologies, USA) in 5 μl of Opti-MEM (Gibco® Life Technologies) was added to the assay plate with Multidrop Combi nL Reagent Dispenser (Thermo Fisher Scientific Oy, Finland). The cells (750 cells in 20 μl of complete medium/well) were delivered to the wells with Multidrop Combi Reagent Dispenser (Thermo Fisher Scientific Oy, Finland) using a standard cassette

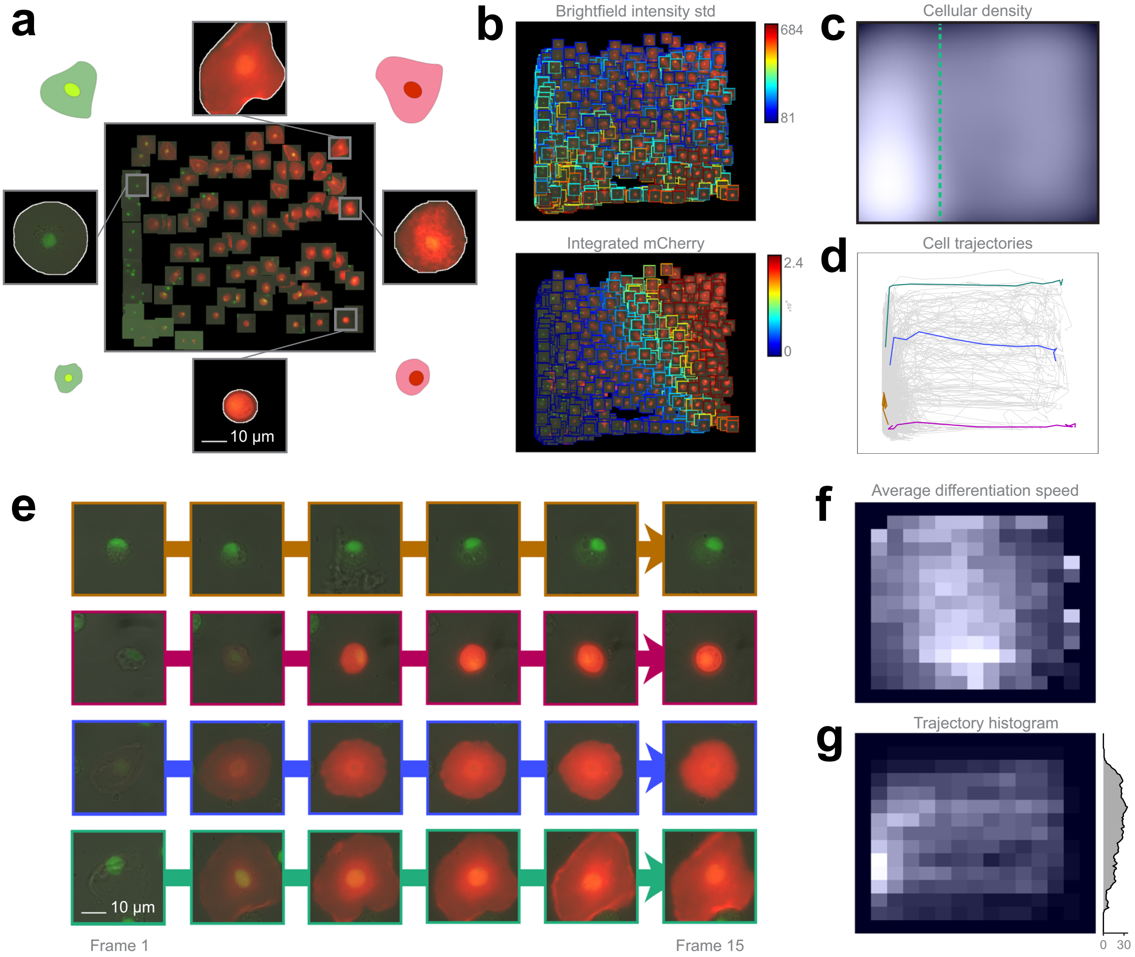


Fig. 4 Hemocyte dataset analysis. a Training set. 109 cells were placed on the regression plane by a microscopy expert. Cells were segmented by applying the NucleAIzer[40] deep learning method on brightfield microscopy images. **b** Single cell features. Colour-coded feature values overlay on the predicted cells. **c** Density plots. Kernel density estimation of single cells. **d** Single-cell trajectories. 2323 cell trajectories on the regression plane. **e** Selected cell trajectories. Representative phenotypes highlighted in **d**. **f** Differentiation speed histogram. Cell differentiation speed on the regression plane. **g** Trajectory histogram. 2D trajectory histogram on the regression plane and 1D projection with trajectory counts, including only those trajectories that reach beyond the green line in **c**.

(Thermo Fisher Scientific Oy). After 72 h of siRNA transfection the cells were fixed with 4% paraformaldehyde, quenched with 50 mM NH_4Cl and stained with Lipidtox Green (HCS LipidTox Green Neutral Lipid Stain, Invitrogen) and 300 nM DAPI (Sigma-Aldrich) for 30 min at RT. Nine images/well were acquired per channel for duplicate plates with an automated epifluorescence ScanR microscope (Olympus) equipped with a 150 W Mercury-Xenon mixed gas arc burner, a 20× long working distance objective (UIS2) and a digital monochrome CCD camera (Hamamatsu). The image resolution was 1344 × 1024 pixel and 16 bit per channel. The 2 identical plates contained a total of 3956 images of 232,084 cells (>2200 cells per siRNA). The generated dataset was deposited to FigShare[36].

To validate our findings, additional biochemical analysis was performed to siRNA-transfected Huh-7 cells. The cells were collected in 0.2 N NaOH, followed by lipid extraction. TGs and CEs were resolved on TLC plates using hexane/diethyl ether/acetic acid (80:20:1) as the mobile phase. Lipids were visualized by charring, the plates were scanned and the intensities were quantified by ImageJ.

Blood cell differentiation dataset. Early third instar Me larvae (eaterGFP as a marker of plasmatocytes, MSNF9MOmCherry as a marker of lamellocytes[33]) were immune induced by wounding the cuticle with an Austerlitz Insect Pin® of 0.2 mm in diameter. Wounded larvae were kept on standard Drosophila food at 25°C. Circulating blood cells were isolated 12 h after wounding. Blood samples of 10 larvae were collected, pooled in 300 µl Schneider's medium (Lonza, Cat: 04-351 Q) supplemented with 10% fetal bovine serum (FBS; Gibco®, Cat: 10270) plus 0.01 mg/ml gentamicin (Sigma, Cat: G3632), 0.065 mg/ml penicillin (Sigma, Cat: P7794) and 0.1 mg/ml streptomycin (Sigma, Cat: S6501). Next it was spread into a well chamber of an 8-well µ-slide (Ibidi, Cat: 80826). Both sample storage and microscopic analysis were carried out at 25 °C.

We acquired 15-frame image sequences/field (141 fields) on 3 channels: brightfield, mCherry, and EGFP, with 2-hour-gaps between the subsequent frames. Images were acquired with a high-content screening microscope (Operetta, Perkin Elmer) equipped with a 60× high-numeric-aperture objective and a digital high resolution 14-bit CCD camera, yielding a total of 4230 images (2 plates, 2115 images in each). The image size was 1360 × 1024 pixels and 8-bit per channel, in TIFF format. The generated dataset was deposited to FigShare[37].

Image segmentation and feature extraction. In order to classify the cells in an image, ACC requires the position and features of each cell to be analyzed. For this purpose, we first flattened illumination distortions of the acquired images by using CIDRE[38]. Then, we used CellProfiler[39] and the NucleAIzer deep learning framework[40] to segment the cells and extract the standard features describing morphology, intensity and texture characteristics. Details of the image analysis and the regression models used in each experiment are reported in Supplementary Note 2.

Regression models. Regression methods, a subgroup of supervised machine learning techniques, are aiming at approximating continuous target variables. Alike for classification, various models have been proposed for regression, ranging from linear regression to neural networks and random forests[41].

The diverse set of regression models raise the problem of model selection for RP. As the RP is completely user-defined, it is impossible to have any prior assumptions on the function to be learnt, hence model selection should be data-driven. RP provides cross-validation assessment of model performance by root mean squared error measure (RMSE) and relative RMSE[42]. Additionally, two important aspects are to be considered when selecting the model.

First, the two-dimensional output format of RP requires the use of multi-target regression, as we require a 2D position (expressed by 2 coordinates) to be predicted. Traditionally, regression models aim at predicting a single continuous variable, which may be naturally extended for multiple dimensions by considering the outputs as independent variables, also called the single-target (ST) method[43]. On the contrary, it has been reported several times that multi-target models that exploit the possible correlation between the output variables may yield significantly better results than the ST methods[44,45]. Consequently, when a strong relationship between the output variables is evident, choosing a multi-target regression model is more appropriate.

Secondly, models that are capable of providing a probabilistic output (i.e. those that provide not only the predictive mean, but also some sort of uncertainty) are less wide-spread for regression than for classification. However, uncertainties provide valuable information to assess the model's performance, and most of the active learning strategies essentially rely on them.

Gaussian processes (GPs) can be used as non-parametric regression models with a probabilistic output[46]. Instead of providing a single prediction for each cell, GP returns a normal distribution whose mean can be used as the predicted value. More importantly, its variance is an estimate for the uncertainty of the given cell. GP is originally considered as a single-target method, however, its multi-target extensions also exist and are known as co-kriging[44,47]. Although GP is a non-parametric method (hence training is not required in principle), it still has hyperparameters (mean, covariance, likelihood, inference functions and their parameters) that can be optimized for enhanced performance. The most frequently applied iterative optimization methods (gradient descents) require initial hyperparameter settings which significantly affect the quality of the ultimate hyperparameter set. Consequently, we have designed heuristic hyperparameter initialization methods for several mean and covariance functions as described in Supplementary Note 3. Due to the broad selection of implementable models, RP provides an interface (via Object Oriented Programming) to facilitate the extension of implemented regression methods. By default, the package contains bridges to several models from Weka[48], Mulan[49] and Matlab's Deep Learning Toolbox. The full list and instructions on how to include new models are provided in Supplementary Note 4.

Active regression. Usually, the most time-consuming part of statistical learning for biomedical applications (including shallow and deep learning) is the procedure of annotation, and – as transfer learning is rarely used – it is often repeated for new experiments. Active learning[50] aims at reducing the number of training samples needed to achieve the most representative training set by automatically proposing cells for annotation. It has previously been shown by Smith and Horvath[51] that active learning reduces the time cost of annotation in HCS compared to classical labelling. Most of the active classification methods are based solely on the predicted class labels, enabling the underlying model to be freely modified. However, these methods are not directly applicable for regression, as they assume that the predicted label is discrete. Active regression methods were developed by Cohn et al.[52], based on variance reduction for Neural Networks, Mixture of Gaussians and Locally Weighted Regression. Here we present active regression methods inspired by the general active classification approaches, and a specific method for Gaussian Processes utilizing its properties (Supplementary Fig. 1).

Committee members. The Committee Members approach is inspired by the *QueryByCommittee* active classification method. Similarly to cross-validation, a set of models (committee) is built up from the available training samples, and a measure of disagreement is defined for the committee. In case of regression, the classical measures cannot be applied directly for two reasons: (1) they rely on the fact that the output is discrete, and (2) they require a probabilistic model. Thus, we propose using the quadratic mean of the Euclidean distance between the committee consensus and the single committee predictions. Hence, the next cell to be labelled by the expert is defined by the following formula:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{argmax}\sqrt{\frac{\sum_{i=1}^{C} d(\hat{\boldsymbol{y}}_i,\bar{\boldsymbol{y}})^2}{C}} \tag{1}$$

where C is the size of the committee, $\hat{\mathbf{y}}_\mathbf{i}$ is the predicted position for $\mathbf{x}$ (a sample not taken from the TS) by the ith committee member, $\bar{\mathbf{y}}$ is the mean of $\hat{\mathbf{y}}$, and d is the Euclidean distance.

Empty regions. The Empty regions method targets the cells which were predicted to the least dense region of the regression plane in terms of training samples. This heuristic is supposed to explore those cell types that are not presented in the TS.

Out of bounds. By design, the regression plane is represented by a unit-square, and has limits in each direction. However, this limitation was not incorporated into the regression models, consequently it is possible that cells are predicted outside of the regression plane's boundaries. Therefore, we propose a strategy that selects these cells for annotation, ranked by their distance from the edges of the regression plane.

Uncertainty sampling. When a probabilistic regression model (such as GP) is available, then, instead of plain predictions, a posterior distribution is defined for each cell, enabling the application of active learning methods aiming at decreasing the variance of this posterior. Our proposed method targets the cell with the highest posterior variance, where the final value for the selection is determined by taking either the mean, the sum, the product, the minimum or the maximum of the 2 separate variances, calculated for each output dimension of the regression plane.

Overall uncertainty sampling. GP has an intriguing property, namely that the posterior distribution is independent of the actual TS positions; it only depends on the input features and the hyperparameters of the GP. In consequence, given fixed hyperparameters, it is possible to exactly calculate how the posterior variance changes, assuming that a new cell is included in the TS even without knowing its position on the regression plane. Executing this calculation for all possible candidates, the resulting cell proposed for annotation is the one that decreases overall variance the most. This approach is formulated by:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{argmin}\sum_{i=1}^{N} f_\sigma^{\boldsymbol{x}}(\boldsymbol{x}_i) \tag{2}$$

where N is the size of the full dataset (including the training dataset) and $f_\sigma^\mathbf{x}(\mathbf{x_i})$ is the variance for $\mathbf{x_i}$, supposing that the GP was trained on the available training set extended with $\mathbf{x}$. The predictive variances for individual samples are calculated from the diagonal elements of the predictive variance matrix according to ref. [46] by the following formula:

$$K\left(\boldsymbol{X}_*,\boldsymbol{X}_*\right) - K(\boldsymbol{X}_*,\boldsymbol{X})K(\boldsymbol{X},\boldsymbol{X})^{-1}K(\boldsymbol{X},\boldsymbol{X}_*) \tag{3}$$

where K is the kernel (covariance) function, $\mathbf{X}$ is the feature matrix of samples not yet predicted and $\mathbf{X}$ is the feature matrix of the training set's elements.

We assessed the performance of the proposed active learning methods with 4 regression models: Random Forest, Gaussian Process, Neural Network and Support Vector Machine; on 2 of our datasets: Lipid droplets and MitoCheck containing 457 and 585 annotated cells, respectively. In each scenario the experiment started with randomly isolating 1/3 of the available samples to a test set, leaving the remaining 2/3 in a pool. Then, 10 cells were randomly selected from the pool for initializing the training set, followed by iteratively extending it with 290 cells according to the active query strategy. In each iteration a regression model was trained, and the relative root mean square error (RRMSE) was calculated on the test set.

The results from 50 independent runs are displayed in Supplementary Fig. 1b, c. In all but one (Gaussian Process in the MitoCheck dataset) scenario there was at least one active learning technique that outperformed random sampling, despite the high variance of error values among different regression models. The Random Forest and Gaussian Process models achieved smaller RRMSE values than the other two methods inhibiting the active strategies' ability to significantly improve the performance in these cases. Still, the CommitteeMembers strategy resulted in the lowest average area under the curve value in 5 out of the 8 cases. We also note that although mean prediction error is the most widespread measure of active learning, other aspects of the model performance (e.g. model coverage) might be equally interesting for the users.

Regression plane output. RP provides output in various formats to satisfy the diverse needs of field experts (Supplementary Movie 2). The simplest output can be obtained by predicting an image in the main window of ACC, by clicking on a cell to see its raw regression plane position. Alternatively, in the regression plane one can select an arbitrary number of images, so that all cells in those images are going to be visualized on the regression plane with their icon at their predicted position. Importantly, these predictions can easily be added to the TS as well.

For well-based analysis, a multi-component report can be generated for each plate. The first component of the report is a pdf file containing a heatmap (simple cell count in a discretized regression plane) and a kernel density estimation (KDE) visualizing the distribution of cells on the regression plane in the particular well (Fig. 2b, c). Besides, the difference and the most dense position shift between single wells, and the average of user-defined control wells are also included.

Secondly, RP provides standard visualization tools (PCA, t-SNE[53] and NeRV[54]) for assessing the relationships among the wells. Each of these methods can generate the figure of Plot of plots (PoP*;* Fig. 2c, Supplementary Fig. 4a–c). In PoP each well is represented by its KDE/heatmap, and the distance between these representations corresponds to the difference between the wells' regression plane distributions (i.e. similar wells are close in PoP, whilst differing ones are farther from each other). In case of plates with higher well-numbers (e.g. 96 or 384) this may result in an overwhelmingly dense diagram, so the PoPs can be re-loaded to RP where they can be examined interactively. Importantly, in the RP-PoP, wells of similar perturbations (replicates) can be highlighted with colours. In addition to these tools for visualization, a clustergram can also be generated, providing a way to compare the perturbations by performing hierarchical clustering (Supplementary Fig. 3e–h, Supplementary Fig. 4d). The matrix in the middle of the clustergram visualizes pairwise Kullback–Leibler divergence (KLD) between the cell-number weighted average of the replicate wells. Clustering is performed on the pairwise KLD matrix with correlation as distance, and average linkage.

Additionally, RP enables the analysis of underlying image features by the Colour Frame (CF) module. CF works by visualizing the feature distribution of cells from the regression plane, using an artificial colour scale. In particular, the user selects a specific feature and adjusts the visualization settings to define a colour for each cell icon's frame in the regression plane. (Figs. 3d and 4b). Notably, CF can be used either for fine tuning of the TS, or for assessing features of interest after prediction.

Finally, the Trajectory Plot (TP) facilitates the assessment of live-cell data composed of time-resolved image sequences of the same fields. Organizing the corresponding single-cells into trajectories using the predicted coordinates of the regression plane enables the visualization of the dynamics of underlying processes (Fig. 4c–f). TP is a multifunctional visualization tool that facilitates a better understanding of the continuous aspect of biological processes and offers several possibilities to investigate cell fates or to compare the development of particular cells as a function of time. Filtering functions help to find subgroups of phenotypes with different behaviours. Interestingly, the dynamics of the process can be perceived by animating the evolution of trajectories (Supplementary Movies 7, 8). We note that the observed distances and the derived speed of motion in trajectories are completely user-defined, hence they should be interpreted relative to the designed training set. This is a general property of supervised methods and represents a trade-off between customizability and fully unbiased approaches.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Synthetic dataset: https://data.broadinstitute.org/bbbc/image_sets.html (dataset ID: BBBC031). Lipid droplet dataset: https://doi.org/10.6084/m9.figshare.c.5067638.v1[36]. Mitocheck dataset: http://www.mitocheck.org/mitotic_cell_atlas/downloads/v1.0.1/mitotic_cell_atlas_v1.0.1_fulldata.zip. The training set for the Mitocheck data generated in this study is available as Supplementary Data 2. Drosophila dataset: https://doi.org/10.6084/m9.figshare.c.5075093.v1[37]. Source data are provided with this paper.

Code availability

RP is a new module of ACC (current version 3.1). ACC is written in MATLAB (The MathWorks, Inc., USA). ACC supports the most common image formats (e.g. tif, bmp, png) and it works under Windows 64-bit, Linux, and OS X environments. Source code and standalone versions (which do not require a MATLAB license), video tutorials, and help documentation files are publicly available at: www.cellclassifier.org. All the ACC materials are copyright protected and distributed under GNU General Public License version 3 (GPLv3). Further software involved in this study: CellProfiler v1 is available freely at: https://cellprofiler.org/previous-releases. The CIDRE framework is freely available at: https://github.com/smithk/cidre. The nucleAIzer pipeline source code is available at: https://github.com/spreka/biomagdsb. The experiments involving Matlab were conducted with Matlab v9.5.0.1298439 (R2018b). The data analysis involving ImageJ was conducted with version 1.49b.

Received: 11 September 2020; Accepted: 30 March 2021;
Published online: 05 May 2021

References

1. Carragher, N. et al. Concerns, challenges and promises of high-content analysis of 3D cellular models. *Nat. Rev. Drug Discov.* **17**, 606–606 (2018).
2. Caicedo, J. C. et al. Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).
3. E. Moen, et al. Deep learning for cellular image analysis. *Nat. Methods* **16**, 1–14 (2019).
4. Sommer, C. & Gerlich, D. W. Machine learning in cell biology–teaching computers to recognize phenotypes. *J. ell Sci.* **126**, 5529–5539 (2013).
5. Smith, K. et al. Phenotypic image analysis software tools for exploring and understanding big image data from cell-based assays. *Cell Syst.* **6**, 636–653 (2018).
6. Piccinini, F. et al. Advanced cell classifier: user-friendly machine-learning-based software for discovering phenotypes in high-content imaging data. *Cell Syst.* **4**, 651–655 (2017).
7. Held, M. et al. CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nat. methods* **7**, 747–754 (2010).
8. Gut, G., Tadmor, M. D., Pe'er, D., Pelkmans, L. & Liberali, P. Trajectories of cell-cycle progression from fixed cell populations. *Nat. Methods* **12**, 951–954 (2015).
9. Kester, L. & van Oudenaarden, A. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* **23**, 166–179 (2018).
10. Cai, Y. et al. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature* **561**, 411–415 (2018).
11. Kerz, M. et al. A novel automated high-content analysis workflow capturing cell population dynamics from induced pluripotent stem cell live imaging data. *J. Biomol. Screen.* **21**, 887–896 (2016).
12. Sacha, D. et al. What you see is what you can change: human-centered machine learning by interactive visualization. *Neurocomputing* **268**, 164–175 (2017).
13. Buja, A. et al. Data visualization with multidimensional scaling. *J. Comput. Graph. Stat.* **17**, 444–472 (2008).
14. Yamauchi, Y. et al. Histone deacetylase 8 is required for centrosome cohesion and influenza A virus entry. *PLoS Pathog.* **7**, e1002316 (2011).
15. Sverchkov, Y. & Craven, M. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS Comput. Biol.* **13**, e1005466 (2017).
16. Kumar, P. & Gupta, A. Active learning query strategies for classification, regression, and clustering: a survey. *J. Comput. Sci. Technol.* **35**, 913–945 (2020).
17. Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
18. Olzmann, J. A. & Carvalho, P. Dynamics and functions of lipid droplets. *Nat. Rev. Mol. Cell Biol.* **20**, 137–155 (2019).
19. Mahdessian, H. et al. TM6SF2 is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid droplet content. *Proc. Natl Acad. Sci. USA* **111**, 8913–8918 (2014).
20. Jung, J. W. et al. Transmembrane 4 L six family member 5 senses arginine for mTORC1 signaling. *Cell Metab.* **29**, 1306–1319 (2019).
21. Evans, C. J., Hartenstein, V. & Banerjee, U. Thicker than blood: conserved mechanisms in Drosophila and vertebrate hematopoiesis. *Dev. Cell* **5**, 673–690 (2003).
22. Honti, V., Csordás, G., Kurucz, É., Márkus, R. & Andó, I. The cell-mediated immunity of *Drosophila melanogaster*: hemocyte lineages, immune compartments, microanatomy and regulation. *Dev. Comp. Immunol.* **42**, 47–56 (2014).
23. Nappi, A. J., Vass, E., Frey, F. & Carton, Y. Superoxide anion generation in Drosophila during melanotic encapsulation of parasites. *Eur. J. Cell Biol.* **68**, 450–456 (1995).
24. Russo, J., Dupas, S., Frey, F., Carton, Y. & Brehelin, M. Insect immunity: early events in the encapsulation process of parasitoid (*Leptopilina boulardi*) eggs in resistant and susceptible strains of Drosophila. *Parasitology* **112**, 135–142 (1996).
25. Lanot, R., Zachary, D., Holder, F. & Meister, M. Postembryonic hematopoiesis in Drosophila. *Dev. Biol.* **230**, 243–257 (2001).
26. Márkus, R., Kurucz, É., Rus, F. & Andó, I. Sterile wounding is a minimal and sufficient trigger for a cellular immune response in *Drosophila melanogaster*. *Immunol. Lett.* **101**, 108–111 (2005).
27. Stofanko, M., Kwon, S. Y. & Badenhorst, P. Lineage tracing of lamellocytes demonstrates Drosophila macrophage plasticity. *PloS One* **5**, e14051 (2010).
28. Kroeger, P. T. Jr, Tokusumi, T. & Schulz, R. A. Transcriptional regulation of eater gene expression in Drosophila blood cells. *Genesis* **50**, 41–49 (2012).
29. Honti, V. et al. Cell lineage tracing reveals the plasticity of the hemocyte lineages and of the hematopoietic compartments in *Drosophila melanogaster*. *Mol. Immunol.* **47**, 1997–2004 (2010).
30. Cattenoz, P. B. et al. Temporal specificity and heterogeneity of Drosophila immune cells. *EMBO J.* **39**, e104486 (2020).
31. Tattikota, S. G. et al. A single-cell survey of Drosophila blood. *Elife* **9**, e54818 (2020).
32. Csordás, G., Gábor, E. & Honti, V. There and back again: the mechanisms of differentiation and transdifferentiation in Drosophila blood cells. *Dev. Biol.* **469**, 135–143 (2021).
33. Anderl, I. et al. Transdifferentiation and proliferation in two distinct hemocyte lineages in *Drosophila melanogaster* larvae after wasp infection. *PLoS Pathog.* **12**, e1005746 (2016).
34. Lehmussola, A., Ruusuvuori, P., Selinummi, J., Huttunen, H. & Yli-Harja, O. Computational framework for simulating fluorescence microscope images with cell populations. *IEEE Trans. Med. Imaging* **26**, 1010–1016 (2007).
35. Nakabayashi, H., Taketa, K., Miyano, K., Yamane, T. & Sato, J. Growth of human hepatoma cell lines with differentiated functions in chemically defined medium. *Cancer Res.* **42**, 3858–3863 (1982).
36. "FigShare - 2020_ACC_RP_LipidDroplets_siRNA,". https://doi.org/10.6084/m9.figshare.c.5067638.v1 (2020).
37. "FigShare - 2020_ACC_RP_DrosophilaBloodCells,". https://doi.org/10.6084/m9.figshare.c.5075093.v1 (2020).
38. Smith, K. et al. CIDRE: an illumination-correction method for optical microscopy. *Nat. Methods* **12**, 404–406 (2015).
39. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
40. Hollandi, R. et al. nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.* **10**, 453–458 (2020).

41. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2009).
42. Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W. & Vlahavas, I. Multi-target regression via input space expansion: treating targets as inputs. *Mach. Learn.* **104**, 55–98 (2016).
43. Borchani, H., Varando, G., Bielza, C. & Larrañaga, P. A survey on multi-output regression. *Wiley Interdiscip. Rev.* **5**, 216–233 (2015).
44. Boyle, P. & Frean, M. Dependent gaussian processes. *Adv. Neural Inf. Process. Syst.* **17**, 217–224 (2004).
45. Han, Z., Liu, Y., Zhao, J. & Wang, W. Real time prediction for converter gas tank levels based on multi-output least square support vector regressor. *Control Eng. Pract.* **20**, 1400–1409 (2012).
46. Rasmussen, C. E. & Williams, C. K. *Gaussian Processes for Machine Learning* Vol. 1 (MIT Press, 2006).
47. Cressie, N. A. C. *Statistics for Spatial Data* (John Wiley & Sons, 1993).
48. Hall, M. et al. The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* **11**, 10–18 (2009).
49. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. & Vlahavas, I. Mulan: a java library for multi-label learning. *J. Mach. Learn. Res.* **12**, 2411–2414 (2011).
50. Settles, B. Active learning literature survey (2009).
51. Smith, K. & Horvath, P. Active learning strategies for phenotypic profiling of high-content screens. *J. Biomol. Screen.* **19**, 685–695 (2014).
52. Cohn, D. A., Ghahramani, Z. & Jordan, M. I. Active learning with statistical models. *J. Artif. Intell. Res.* **4**, 129–145 (1996).
53. Maaten, Lvd & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
54. Venna, J., Peltonen, J., Nybo, K., Aidos, H. & Kaski, S. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J. Mach. Learn. Res.* **11**, 451–490 (2010).

Acknowledgements

The authors thank Antti Lehmussola and Pekka Ruusuvuori (Tampere University of Technology, Finland) for the information provided about the SIMCEP software; Samuli Ripatti and Ida Surakka (FIMM, University of Helsinki, Finland) for their valuable comments on our experiments related to the genes associated with dyslipidemia; Anna Uro (Faculty of Medicine, University of Helsinki) for providing expertise in the biochemical quantification of lipid levels; Mariliina Arjama (FIMM, University of Helsinki, Finland) for technical expertise in cell culture; the FIMM High Throughput Biomedicine Unit for providing access to high throughput robotics and siRNA library and the FIMM High Content Imaging and Analysis unit for HC-imaging (HiLIFE, University of Helsinki and Biocenter Finland); Olli Kallioniemi (FIMM, University of Helsinki, Finland) for support in HC-imaging capabilities and funding; Gabriella Tick and Máté Görbe (BRC, Szeged, Hungary) for their help with the software documentation; the Finnish Grid and Cloud Infrastructure (urn:nbn:fi:research-infras-2016072533) for computational resources; Dóra Bokor (BRC, Szeged, Hungary) for proofreading the manuscript. A.Sz., B.T., A.B., E.M., Cs.M. and P.H. acknowledge support from the Hungarian National Brain Research Program (MTA-SE-NAP B-BIOMAG), from the LENDULET-BIOMAG Grant (2018-342), from the European Regional Development Funds (GINOP-2.3.2-15-2016-00006, GINOP-2.3.2-15-2016-00026, GINOP-2.3.2-15-2016-00037), from the H2020 (ERAPERMED-COMPASS, DiscovAIR) and from the Chan Zuckerberg Initiative (Deep Visual Proteomics). A.Sz. and E.I. acknowledges support from University of Helsinki (Centre of Excellence matching funds) and Academy of Finland (project 324929). V.P., L.P. and P.H. acknowledge support from the Finnish TEKES FiDiPro Fellow Grant 40294/13 and FIMM High Content Imaging and Analysis Unit (FIMM-HCA; HiLIFE-HELMI) and Biocenter Finland, Finnish Cancer Society, Juselius Foundation, Academy of Finland Centre of Excellence in Translational Cancer Biology, Kymenlaakso and Finnish Cultural Foundation. V.P. acknowledges University of Helsinki post-doctoral research project grant. F.P. acknowledges support from the Union for International Cancer Control (UICC) for a UICC Yamagiwa-Yoshida (YY) Memorial International Cancer Study Grant (ref: UICC-YY/678329). V.H. and I.A. acknowledge the Hungarian National Research Fund (OTKA NKFI-2 NN118207). V.H. acknowledges support from the National Research, Development and Innovation Office (OTKA K-131484). L.P. and J.P. acknowledge support from the Academy of Finland, decision numbers 295694, 313748, 327352 and 310552.

Author contributions

P.H. conceived and led the project. A.Sz. developed the Regression Plane tool. A.Sz. and A.B. developed the trajectory tool. T.B. debugged and released the software. A.Sz., F.P., T.B., I.G.V., E.M., Cs.M., L.P., S.T., V.P. and V.H. designed and performed the experiments. A.Sz., F.P., A.B. and V.P. tested the software tool. E.I., Y.Y., I.A., J.P. and P.H. supervised the project. F.P. prepared the documentation and website. A.Sz., F.P., V.P., V.H. and P.H. wrote the manuscript. A.Sz., F.P. and I.B. prepared the figures included in the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing isnterests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-22866-x.

Correspondence and requests for materials should be addressed to P.H.

Peer review information *Nature Communications* thanks Emmanuel Gustin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

Cite as: J. L. Daly *et al*., *Science*
10.1126/science.abd3072 (2020).

Neuropilin-1 is a host factor for SARS-CoV-2 infection

James L. Daly[1]*, Boris Simonetti[1]*†, Katja Klein[2]*, Kai-En Chen[3]‡, Maia Kavanagh Williamson[2]‡, Carlos Antón-Plágaro[1]‡, Deborah K. Shoemark[4], Lorena Simón-Gracia[5], Michael Bauer[6], Reka Hollandi[7], Urs F. Greber[6], Peter Horvath[7,8], Richard B. Sessions[1], Ari Helenius[9], Julian A. Hiscox[10,11], Tambet Teesalu[5], David A. Matthews[2], Andrew D. Davidson[2], Brett M. Collins[3], Peter J. Cullen[1]†, Yohei Yamauchi[2]†

[1]School of Biochemistry, Faculty of Life Sciences, Biomedical Sciences Building, University of Bristol, BS8 1TD, UK. [2]School of Cellular and Molecular Medicine, Faculty of Life Sciences, Biomedical Sciences Building, University of Bristol, BS8 1TD, UK. [3]Institute for Molecular Bioscience, the University of Queensland, St. Lucia, QLD 4072, Australia. [4]School of Biochemistry and BrisSynBio Centre, Faculty of Life Sciences, Biomedical Sciences Building, University of Bristol, BS8 1TD, UK. [5]Laboratory of Cancer Biology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia. [6]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland. [7]Synthetic and Systems Biology Unit, Biological Research Centre (BRC), Szeged, Hungary. [8]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [9]Institute of Biochemistry, ETH Zurich, Zurich, Switzerland. [10]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK. [11]Singapore Immunology Network, Agency for Science, Technology, and Research, 138648, Singapore.

*These authors contributed equally to this work.

†Corresponding author. Email: bs13866@bristol.ac.uk (B.S.); pete.cullen@bristol.ac.uk (P.J.C.); yohei.yamauchi@bristol.ac.uk (Y.Y.)

‡These authors contributed equally to this work.

SARS-CoV-2, the causative agent of COVID-19, uses the viral Spike (S) protein for host cell attachment and entry. The host protease furin cleaves the full-length precursor S glycoprotein into two associated polypeptides: S1 and S2. Cleavage of S generates a polybasic Arg-Arg-Ala-Arg C-terminal sequence on S1, which conforms to a C-end rule (CendR) motif that binds to cell surface Neuropilin-1 (NRP1) and Neuropilin-2 (NRP2) receptors. Here, we used X-ray crystallography and biochemical approaches to show that the S1 CendR motif directly bound NRP1. Blocking this interaction using RNAi or selective inhibitors reduced SARS-CoV-2 entry and infectivity in cell culture. NRP1 thus serves as a host factor for SARS-CoV-2 infection and may potentially provide a therapeutic target for COVID-19.

SARS-CoV-2 is the coronavirus responsible for the current COVID-19 pandemic (*1*, *2*). A striking difference between the S protein of SARS-CoV-2 and SARS-CoV is the presence, in the former, of a polybasic sequence motif, RRAR, at the S1/S2 boundary. It provides a cleavage site for a host proprotein convertase, furin (*3*–*5*) (fig. S1A). The resulting two proteins, S1 and S2, remain non-covalently associated, with the serine protease TMPRSS2 further priming S2 (*6*). Furin-mediated processing increases infectivity and affects the tropism of SARS-CoV-2, while furin inhibition diminishes SARS-CoV-2 entry, and deletion of the polybasic site in the S protein reduces syncytia formation in cell culture (*3*–*5*, *7*).

The C terminus of the S1 protein generated by furin cleavage has an amino acid sequence (^{682}RRAR^{685}), that conforms to a [R/K]XX[R/K] motif, termed the ‘C-end rule’ (CendR) (fig. S1B) (*8*). CendR peptides bind to Neuropilin-1 (NRP1) and NRP2, transmembrane receptors that regulate pleiotropic biological processes, including axon guidance, angiogenesis, and vascular permeability (*8*–*10*). To explore the possibility that the SARS-CoV-2 S1 protein may associate with neuropilins we generated a GFP-tagged S1 construct (GFP-S1) (fig. S1C). When expressed in HEK293T cells engineered to express the SARS-CoV-2 receptor ACE2, GFP-S1 immunoprecipitated endogenous NRP1 and ACE2 (Fig. 1A). We transiently co-expressed NRP1-mCherry and either GFP-S1 or GFP-S1 ΔRRAR (a deletion of the terminal ^{682}RRAR^{685} residues) in HEK293T cells. NRP1 immunoprecipitated the S1 protein, and deletion of the CendR motif reduced this association (Fig. 1B). Comparable binding was also observed with mCherry-NRP2, a receptor with high homology to NRP1 (fig. S1, D and E). In both cases, residual binding was observed with the ΔRRAR mutant indicating an additional CendR-independent association between neuropilins and the S1 protein.

To probe the functional relevance of this interaction, we generated HeLa wild type and NRP1 knock out (KO) cell lines stably expressing ACE2, designated as HeLa^{wt}+ACE2 and HeLa^{NRP1KO}+ACE2 respectively (the level of ACE2 expression was comparable between these lines) (fig. S1F). Using a clinical isolate SARS-CoV-2 (SARS-CoV-2/human/Liverpool/REMRQ001/2020), we performed viral infection assays and fixed the cells at 6 and 16 hours post infection (hpi). SARS-CoV-2 infection was reduced in HeLa^{NRP1KO}+ACE2 relative to HeLa^{wt}+ACE2 (Fig. 1C). HeLa cells lacking ACE2 expression were not infected (fig. S1G). In Caco-2 cells, a human colon adenocarcinoma cell line endogenously expressing ACE2 and widely used in COVID-19 studies, the suppression of NRP1 expression by shRNA greatly reduced SARS-CoV-2 infection at both 7 and 16 hpi respectively, whereas that of vesicular stomatitis virus (VSV) pseudotyped with VSV-G was unaffected (Fig. 1D and figs. S1H and S2A). To determine if NRP1 was required for early virus infection, we established a sequential staining procedure using antibodies against SARS-CoV-2 S and N proteins to distinguish extracellular and intracellular viral particles (fig. S2B). While NRP1 depletion did not affect SARS-CoV-2 binding to the Caco-2 cell surface (Fig. 1E), virus uptake was halved in NRP1-depleted cells compared to control cells after 30 min of internalization (Fig. 1F). Thus, NRP1 enhances SARS-CoV-2 entry and infection.

We also observed that SARS-CoV-2-infected HeLa^{wt}+ACE2 cells displayed a multi-nucleated syncytia cell pattern, as reported by others (Fig. 1C) (*5*). Using an image analysis algorithm and supervised machine learning (fig. S2, C to F) (*11*), we quantified syncytia of infected HeLa^{wt}+ACE2 and HeLa^{NRP1KO}+ACE2 cells. At 16 hpi, the majority of HeLa^{wt}+ACE2 cells formed syncytia, while in HeLa^{NRP1KO}+ACE2 cells this phenotype was reduced (fig. S2G). When infected with a SARS-CoV-2 isolate lacking the furin cleavage site (SARS-CoV-2 ΔS1/S2) (fig. S1A) the differences in infection and syncytia formation were less pronounced (fig. S2, H and I). However, a significant decrease in infection of HeLa^{NRP1KO}+ACE2 was still observed at 16 hpi, indicating that NRP1 may additionally influence infection through a CendR-independent mechanism (fig. S2H).

The extracellular regions of NRP1 and NRP2 are composed of two CUB domains (a1 and a2), two coagulation factor domains (b1 and b2), and a MAM domain (*9*). Of these, the b1 domain contains the specific binding site for CendR peptides (fig. S3A) (*12*). Accordingly, the mCherry-b1 domain of NRP1 immunoprecipitated GFP-S1, and a shortened GFP-S1 construct spanning residues 493-685 (figs. S1C and S3B). Isothermal titration calorimetry (ITC) established that the b1 domain of NRP1 directly bound a synthetic S1 CendR peptide (^{679}NSPRRAR^{685}) with an affinity of 20.3 μM at pH 7.5, which was enhanced to 13.0 μM at pH 5.5 (Fig. 2A). Binding was not observed to a S1 CendR peptide in which the C-terminal arginine was mutated to alanine (^{679}NSPRRAA^{685}) (Fig. 2A). We co-crystallized the NRP1 b1 domain in complex with the S1 CendR peptide (Fig. 2B). The resolved 2.35 Å structure revealed 4 molecules of b1 with electron density of the S1 CendR peptide clearly visible in the asymmetric unit (fig. S3C). S1 CendR peptide binding displayed remarkable similarity to the previously solved structure of NRP1 b1 domain in complex with its endogenous ligand VEGF-A$_{164}$ (Fig. 2B and fig. S3D) (*12*). The key residues responsible for contacting the C-terminal R685 of the CendR peptide - Y297, W301, T316, D320, S346, T349 and Y353 - are almost identical between the two structures (Fig. 2B and fig. S3D). The R682 and R685 sidechains together engage NRP1 via stacked cation-π interactions with NRP1 side chains of Y297 and Y353. By projecting these findings onto the structure of the NRP1 ectodomain, the b1 CendR binding pocket appears to be freely accessible to the S1 CendR peptide (fig. S3E) (*13*).

Site-directed mutagenesis of the S1 R685 residue to aspartic acid drastically reduced GFP-S1$^{493-685}$ immunoprecipitation by mCherry-b1, confirming the critical role of the C-terminal arginine (Fig. 2C). Mutagenesis of the T316 residue within the mCherry-b1 domain of NRP1 to arginine also reduced association with GFP-S1$^{493-685}$, consistent with its inhibitory impact on VEGF-A$_{164}$ binding (*12*) (Fig. 2D). Accordingly, incubation of mCherry-b1 with VSV particles pseudotyped with trimeric S resulted in immunoprecipitation of processed forms of S1, which was dependent on the T316 residue (fig. S3F). Next, we transiently expressed either GFP, full length NRP1 wt-GFP or full length NRP1-GFP harboring the T316R mutation in HeLa^{NRP1KO}+ACE2 cells. GFP expression and ACE2 expression levels were comparable and both constructs retained similar cell surface localization (fig. S3, G and H). SARS-CoV-2 infection was significantly enhanced in cells expressing NRP1 wt-GFP compared to GFP control, whereas it was not enhanced in cells expressing the T316R mutant (Fig. 2E). Thus, the SARS-CoV-2 S1 CendR and NRP1 interaction promotes infection.

To establish the functional relevance of the S1 CendR-NRP1 interaction, we screened monoclonal antibodies (mAb#1, mAb#2, mAb#3) raised against the NRP1 b1b2 ectodomain. All three bound to the NRP1 b1b2 domain, displayed staining by immunofluorescence in NRP1-expressing PPC-1 (human primary prostate cancer) cells but not in M21 (human melanoma) cells that do not express NRP1 (fig. S4A) (*8*), and stained the extracellular domain of NRP1-GFP expressed in cells (fig. S4B). Of these antibodies, mAb#3, and to a lesser extent mAb#1, bound to the CendR-binding pocket with high specificity, as defined by reduced ability to bind to a b1b2 mutant that targets residues (S346, E348, T349) at the opening of the binding pocket (Fig. 3A) (*12*). Incubation of Caco-2 cells with mAbs#1 and 3, reduced SARS-CoV-2 infection compared to a control mAb targeting avian influenza A virus (H11N3) hemagglutinin (Fig. 3B). Consistent with this, mAb#3 inhibited binding of GFP-S1$^{493-685}$ and mCherry-b1 (Fig. 3C). As a comparison, Caco-2 and Calu-3 cells were incubated with soluble ACE2, which inhibited SARS-CoV-2 infection in both cases (fig. S4C).

Next, we turned to the small molecule EG00229, a selective NRP1 antagonist that binds the b1 CendR binding pocket and inhibits VEGF-A binding (Fig. 3D) (*14*). ITC established that EG00229 bound to the NRP1 b1 domain with a K$_d$ of 5.1 and 11.0 μM at pH 7.5 and 5.5 respectively (Fig. 3E). EG00229 inhibited the direct binding between b1 and the S1 CendR peptide, and the immunoprecipitation of GFP-S1$^{493-685}$ by mCherry-b1 (Fig. 3E and fig. S4D). Finally, incubation of Caco-2 cells with EG00229 reduced the efficiency of SARS-CoV-2 infection at 7 and 16 hpi (Fig. 3F). Thus, the SARS-CoV-2 interaction with NRP1 can be targeted to reduce viral infectivity in relevant human cell lines (fig. S5).

Cell entry of SARS-CoV-2 depends on priming by host cell proteases (*5*, *6*, *15*). Our data indicate that a component of SARS-CoV-2 S protein binding to cell surface neuropilins occurs via the S1 CendR motif generated by the furin cleavage of S1/S2. While not affecting cell surface attachment, this interaction promotes entry and infection by SARS-CoV-2 in physiologically relevant cell lines widely used in the study of COVID-19. The molecular basis for the effect is unclear, but neuropilins are known to mediate the internalization of CendR ligands through an endocytic process resembling macropinocytosis, (*8*, *16*, *17*). Interestingly, gene expression analysis has revealed an up-regulation of NRP1 and NRP2 in lung tissue from COVID-19 patients (*18*). A SARS-CoV-2 virus with a natural deletion of the S1/S2 furin cleavage site demonstrated attenuated pathogenicity in hamster models (*19*). NRP1 binding to the CendR peptide in S1 is thus likely to play a role in the increased infectivity of SARS-CoV-2 compared with SARS-CoV. The ability to target this specific interaction may provide a route for COVID-19 therapies.

REFERENCES AND NOTES

1. WHO Coronavirus disease, 2019 (COVID-19) Weekly Epidemiological Update – 31 August 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200831-weekly-epi-update-3.pdf?sfvrsn=d7032a2a_4
2. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020). doi:10.1016/S1473-3099(20)30120-1 Medline
3. D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, J. S. McLellan, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020). doi:10.1126/science.abb2507 Medline
4. A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, D. Veesler, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281–292.e6 (2020). doi:10.1016/j.cell.2020.02.058 Medline
5. M. Hoffmann, H. Kleine-Weber, S. Pöhlmann, A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell* **78**, 779–784.e5 (2020). doi:10.1016/j.molcel.2020.04.022 Medline
6. M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M. A. Müller, C. Drosten, S. Pöhlmann, SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8 (2020). doi:10.1016/j.cell.2020.02.052 Medline
7. J. Shang, Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach, F. Li, Cell entry mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11727–11734 (2020). doi:10.1073/pnas.2003138117 Medline
8. T. Teesalu, K. N. Sugahara, V. R. Kotamraju, E. Ruoslahti, C-end rule peptides mediate neuropilin-1-dependent cell, vascular, and tissue penetration. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16157–16162 (2009). doi:10.1073/pnas.0908201106 Medline
9. H. F. Guo, C. W. Vander Kooi, Neuropilin functions as an essential cell surface receptor. *J. Biol. Chem.* **290**, 29120–29126 (2015). doi:10.1074/jbc.R115.687327 Medline
10. A. Plein, A. Fantin, C. Ruhrberg, Neuropilin regulation of angiogenesis, arteriogenesis, and vascular permeability. *Microcirculation* **21**, 315–323 (2014). doi:10.1111/micc.12124 Medline
11. R. Hollandi, A. Szkalisity, T. Toth, E. Tasnadi, C. Molnar, B. Mathe, I. Grexa, J. Molnar, A. Balind, M. Gorbe, M. Kovacs, E. Migh, A. Goodman, T. Balassa, K. Koos, W. Wang, J. C. Caicedo, N. Bara, F. Kovacs, L. Paavolainen, T. Danka, A. Kriston, A. E. Carpenter, K. Smith, P. Horvath, nucleAIzer: A parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.* **10**, 453–458.e6 (2020). doi:10.1016/j.cels.2020.04.003
12. M. W. Parker, P. Xu, X. Li, C. W. Vander Kooi, Structural basis for selective vascular endothelial growth factor-A (VEGF-A) binding to neuropilin-1. *J. Biol. Chem.* **287**, 11082–11089 (2012). doi:10.1074/jbc.M111.331140 Medline
13. B. J. Janssen, T. Malinauskas, G. A. Weir, M. Z. Cader, C. Siebold, E. Y. Jones, Neuropilins lock secreted semaphorins onto plexins in a ternary signaling complex. *Nat. Struct. Mol. Biol.* **19**, 1293–1299 (2012). doi:10.1038/nsmb.2416 Medline
14. A. Jarvis, C. K. Allerston, H. Jia, B. Herzog, A. Garza-Garcia, N. Winfield, K. Ellard, R. Aqil, R. Lynch, C. Chapman, B. Hartzoulakis, J. Nally, M. Stewart, L. Cheng, M. Menon, M. Tickner, S. Djordjevic, P. C. Driscoll, I. Zachary, D. L. Selwood, Small molecule inhibitors of the neuropilin-1 vascular endothelial growth factor A (VEGF-A) interaction. *J. Med. Chem.* **53**, 2215–2226 (2010). doi:10.1021/jm901755g Medline
15. J. K. Millet, G. R. Whittaker, Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. *Virology* **517**, 3–8 (2018). doi:10.1016/j.virol.2017.12.015 Medline
16. M. Simons, E. Gordon, L. Claesson-Welsh, Mechanisms and regulation of endothelial VEGF receptor signalling. *Nat. Rev. Mol. Cell Biol.* **17**, 611–625 (2016). doi:10.1038/nrm.2016.87 Medline
17. H. B. Pang, G. B. Braun, T. Friman, P. Aza-Blanc, M. E. Ruidiaz, K. N. Sugahara, T. Teesalu, E. Ruoslahti, An endocytosis pathway initiated through neuropilin-1 and regulated by nutrient availability. *Nat. Commun.* **5**, 4904 (2014). doi:10.1038/ncomms5904 Medline
18. M. Ackermann, S. E. Verleden, M. Kuehnel, A. Haverich, T. Welte, F. Laenger, A. Vanstapel, C. Werlein, H. Stark, A. Tzankov, W. W. Li, V. W. Li, S. J. Mentzer, D. Jonigk, Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in Covid-19. *N. Engl. J. Med.* **383**, 120–128 (2020). doi:10.1056/NEJMoa2015432 Medline
19. S.-Y. Lau, P. Wang, B. W.-Y. Mok, A. J. Zhang, H. Chu, A. C.-Y. Lee, S. Deng, P. Chen, K.-H. Chan, W. Song, Z. Chen, K. K.-W. To, J. F.-W. Chan, K.-Y. Yuen, H. Chen, Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. *Emerg. Microbes Infect.* **9**, 837–842 (2020). doi:10.1080/22221751.2020.1756700 Medline
20. A. D. Davidson, M. K. Williamson, S. Lewis, D. Shoemark, M. W. Carroll, K. J. Heesom, M. Zambon, J. Ellis, P. A. Lewis, J. A. Hiscox, D. A. Matthews, Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Med.* **12**, 68 (2020). doi:10.1186/s13073-020-00763-0 Medline
21. M. Berger Rentsch, G. Zimmer, A vesicular stomatitis virus replicon-based bioassay for the rapid and sensitive determination of multi-species type I interferon. *PLOS ONE* **6**, e25858 (2011). doi:10.1371/journal.pone.0025858 Medline
22. K. Smith, Y. Li, F. Piccinini, G. Csucs, C. Balazs, A. Bevilacqua, P. Horvath, CIDRE: An illumination-correction method for optical microscopy. *Nat. Methods* **12**, 404–406 (2015). doi:10.1038/nmeth.3323 Medline
23. Ronneberger, O., Fischer, P., and Brox. T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-assisted intervention.* 234-241, Springer International Publishing (2015).
24. F. Piccinini, T. Balassa, A. Szkalisity, C. Molnar, L. Paavolainen, K. Kujala, K. Buzas, M. Sarazova, V. Pietiainen, U. Kutay, K. Smith, P. Horvath, Advanced cell classifier: User-friendly machine-learning-based software for discovering phenotypes in high-content imaging data. *Cell Syst.* **4**, 651–655.e5 (2017). doi:10.1016/j.cels.2017.05.012 Medline
25. B. A. Appleton, P. Wu, J. Maloney, J. Yin, W.-C. Liang, S. Stawicki, K. Mortara, K. K. Bowman, J. M. Elliott, W. Desmarais, J. F. Bazan, A. Bagri, M. Tessier-Lavigne, A. W. Koch, Y. Wu, R. J. Watts, C. Wiesmann, Structural studies of neuropilin/antibody complexes provide insights into semaphorin and VEGF binding. *EMBO J.* **26**, 4902–4912 (2007). doi:10.1038/sj.emboj.7601906 Medline
26. C. W. Vander Kooi, M. A. Jusino, B. Perman, D. B. Neau, H. D. Bellamy, D. J. Leahy, Structural basis for ligand and heparin binding to neuropilin B domains. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6152–6157 (2007). doi:10.1073/pnas.0700043104 Medline
27. B. G. Dorner, S. Steinbach, M. B. Hüser, R. A. Kroczek, A. Scheffold, Single-cell analysis of the murine chemokines MIP-1α, MIP-1β, RANTES and ATAC/lymphotactin by flow cytometry. *J. Immunol. Methods* **274**, 83–91 (2003).

doi:10.1016/S0022-1759(02)00498-2 Medline

28. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010). doi:10.1107/S0907444909047337 Medline
29. P. R. Evans, G. N. Murshudov, How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204–1214 (2013). doi:10.1107/S0907444913000061 Medline
30. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007). doi:10.1107/S0021889807021206 Medline
31. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010). doi:10.1107/S0907444909052925 Medline
32. V. B. Chen, W. B. Arendall 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010). doi:10.1107/S0907444909042073 Medline
33. L. Holm, P. Rosenström, Dali server: Conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545-9 (2010). doi:10.1093/nar/gkq366 Medline

ACKNOWLEDGMENTS

We thank the Bristol Synthetic Biology Centre and the Advanced Computing Research Centre for provision of HPC (Bluegem), and the University of Bristol Wolfson Bioimaging Facility. We thank the University of Queensland Remote Operation Crystallisation and X-ray facility (UQ-ROCX) and the staff for their support with the crystallization experiments, and the staff of the Australian Synchrotron for assistance with X-ray diffraction data collection. **Funding:** JLD was supported by a Wellcome Trust studentship from the Dynamic Molecular Cell Biology Ph.D. program (203959/Z/16/Z), CAP was supported by Beca Fundación Ramón Areces Estudios Postdoctorales en el Extranjero and MKW was supported by a MRC grant (MR/R020566/1) awarded to ADD. This project has received funding from the MRC (MR/P018807/1), Wellcome Trust (104568/Z/14/2), Lister Institute of Preventive Medicine, and Elizabeth Blackwell Institute for Health Research Rapid Response Call (COVID-19) awarded to PJC, the European Research Council under the European Union's Horizon 2020 research and innovation program (No 856581 - CHUbVi), and from MRC-AMED (MR/T028769/1) awarded to YY, the Swiss National Science Foundation and Kanton Zurich awarded to UFG. BMC is supported by an Australian National Health and Medical Research Council (NHMRC) Senior Research Fellowship (APP1136021) and Project Grant (APP1156493), and the United States Food and Drug Administration grant number HHSF223201510104C 'Ebola Virus Disease: correlates of protection, determinants of outcome and clinical management' amended to incorporate urgent COVID-19 studies awarded to JAH, ADD and DAM. RH and PH acknowledge support from the LENDULET-BIOMAG Grant (2018-342), from H2020-discovAIR (874656), and from Chan Zuckerberg Initiative, Seed Networks for the HCA-DVP. TT was supported by the European Regional Development Fund (Project No. 2014-2020.4.01.15-0012), by European Research Council grant GLIOGUIDE and Estonian Research Council (grants PRG230 and EAG79, to T.T.). **Author contributions:** JLD, BS, AH, PJC and YY conceived the study. JLD, BS, KK and YY performed most of the experiments. KK, MKW, DAM and ADD performed all work with infectious SARS-CoV-2 supervised by ADD. MKW and ADD isolated SARS-CoV-2 strains used for the work. KC, CAP, MB, LSG, UFG, KK, RBS, DKS, JAH and TT did experimental work and/or provided essential reagents. RH and PH performed image analysis. BS, ADD, BMC, PJC and YY supervised the research. JLD, BS, ADD, PJC and YY wrote the manuscript and made the figures. All authors read and approved the final manuscript. **Competing onterests:** T. Teesalu is an inventor of patents on CendR peptides and shareholder of Cend Therapeutics Inc., a company that holds a license for the CendR peptides and is developing the peptides for cancer therapy. J. Hiscox is a member of the Department of Health, New and Emerging Respiratory Virus Threats Advisory Group (NERVTAG) and the Department of Health, Testing Advisory Group. U. Greber is a consultant to F. Hoffmann-La Roche Ltd, Switzerland. All other authors declare no competing interests. **Data and materials availability:** Coordinates and structure factors for the NRP1 b1 - S1 CendR peptide complex have been deposited at the Protein Data Bank (PDB) with accession code 7JJC. All other data are available in the manuscript or the supplementary material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/cgi/content/full/science.abd3072/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S3
References (*20*–*33*)
MDAR Reproducibility Checklist

14 June 2020; accepted 12 October 2020
Published online 20 October 2020
10.1126/science.abd3072

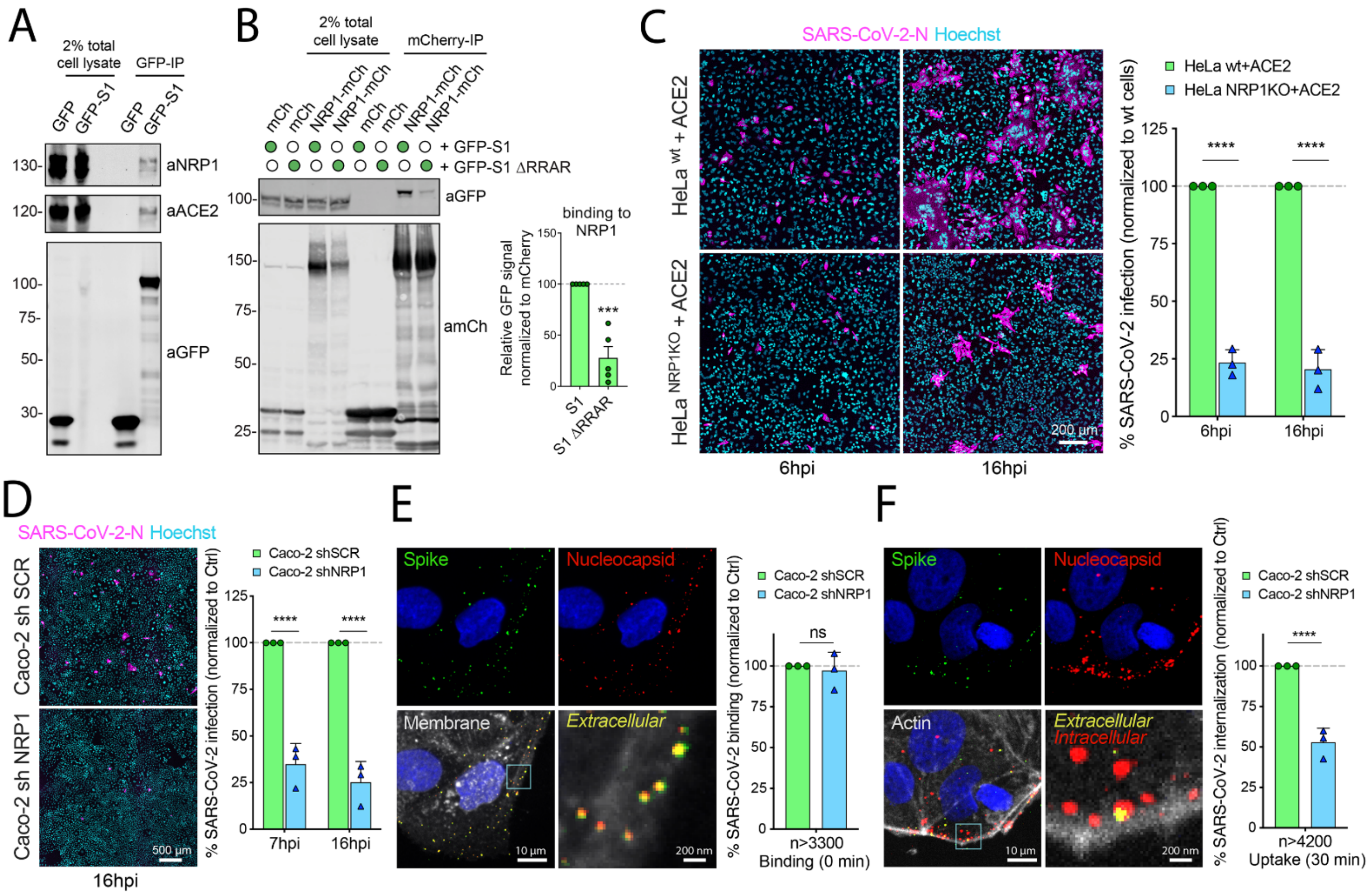


Fig. 1. NRP1 Interacts with S1 and enhances SARS-CoV-2 infection. (**A**) HEK293T cells transduced to express ACE2 were transfected to express GFP or GFP-tagged S1 and lysed after 24h. The lysates were subjected to GFP-nanotrap and the immune-isolates were blotted for ACE2 and NRP1 (N=3). (**B**) HEK293T cells were co-transfected to express GFP-tagged S1 or GFP-S1 ΔRRAR and mCherry or mCherry-tagged NRP1 and subjected to GFP-nanotrap (N=5). Two-tailed unpaired t-test; P= 0.0002. (**C**) HeLa^{wt}+ACE2 and HeLa$^{NRP1\ KO}$+ACE2 cells were infected with SARS-CoV-2. Cells were fixed at 6 or 16 hpi and stained for N protein (magenta) and Hoechst (cyan), and virus infectivity was quantified (N=3). Two-tailed unpaired t-test; P=0.00002 and 0.00088. Scale bar=200 μm. (**D**) Caco-2 cells expressing shRNA against NRP1 or a non-targeting control (SCR) were infected with SARS-CoV-2 and fixed at 7 or 16 hpi. The cells were stained for N protein (magenta) and Hoechst (cyan), and infectivity was quantified (N=3). Two-tailed unpaired t-test; P=0.0005 and 0.00032. Scale bar=500 μm. (**E**) Caco-2 shSCR or shNRP1 cells were inoculated with MOI=50 of SARS-CoV-2 and incubated in the cold for 60 min, and fixed. A two-step antibody staining procedure was performed using anti-S and -N Abs to distinguish external (green) and total (red) virus particles, and the binding of particles per cell was quantified for over 3300 particles per condition (N=3). Two-tailed unpaired t-test; P=0.6859. (**F**) Caco-2 shSCR or shNRP1 cells were bound with SARS-CoV-2 as in (**E**), followed by incubation at 37 °C for 30 min. The cells were fixed and stained as in (**E**). Viral uptake was quantified for over 4200 particles per condition (N=3). Two-tailed unpaired t-test; P=0.00079. Scale bars for (**E**) and (**F**) = 10 μm and 200 nm (zoom panels). The square regions were zoomed in. The bars, error bars, circles and triangles represent the mean, SEM (**B**) and SD (**C**-**F**), individual data points, respectively. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$.

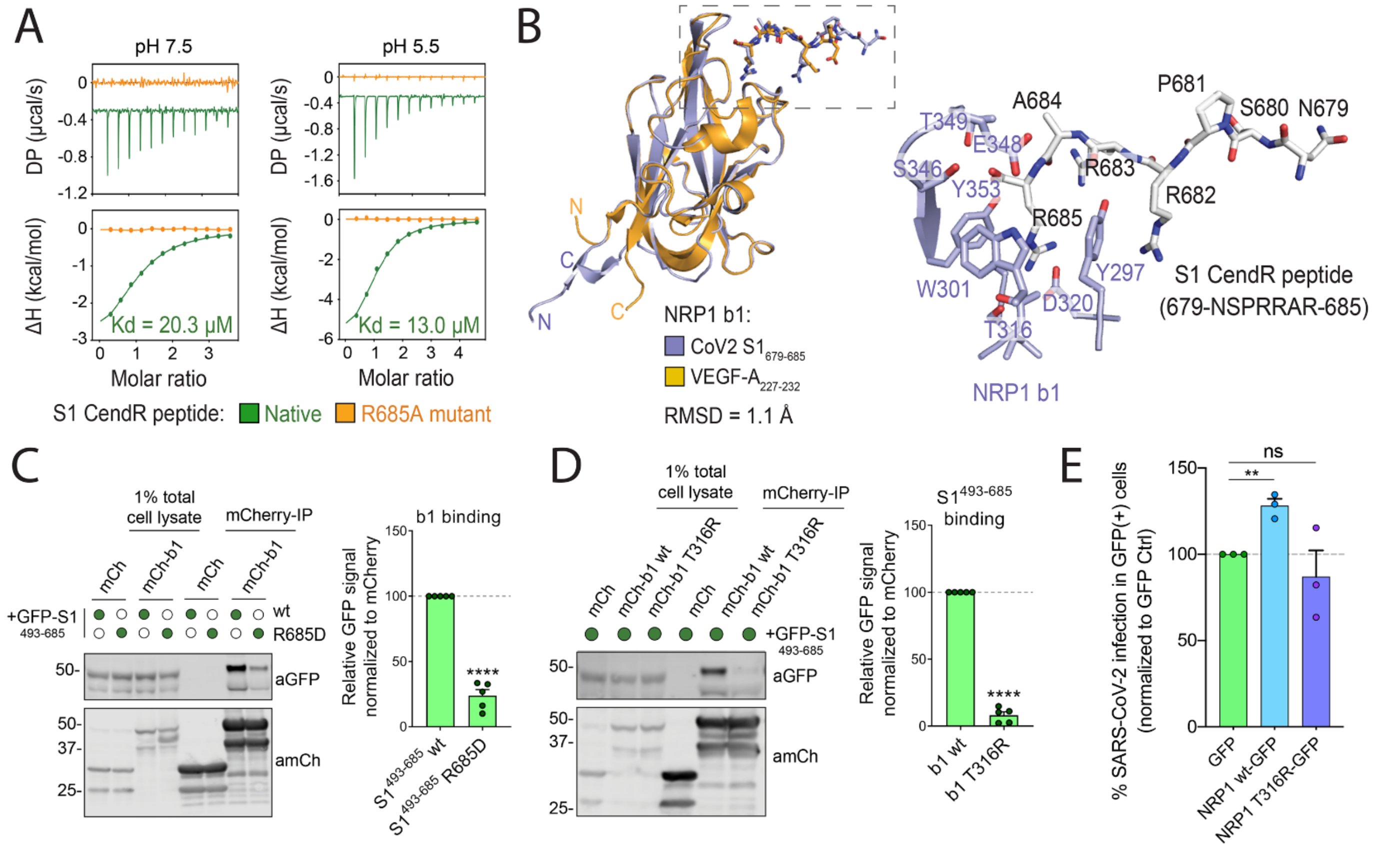


Fig. 2. Molecular basis for CendR binding of SARS-CoV-2 S1 with NRP1. (**A**) Binding of NRP1 b1 with native (green line) and mutant (orange line) form of S1 CendR peptide (corresponding to residues 679-685) by ITC at two different pH conditions (N=3). All ITC graphs represents the integrated and normalized data fit with 1-to-1 ratio binding. (**B**) Left: NRP1 b1 – S1 CendR peptide complex superposed with NRP1 b1 – VEGF-A fusion complex (PDB ID: 4DEQ). Bound peptides are shown in stick representation. RMSD = root mean square deviation. Right: Enlarged view highlighting the binding of S1 CendR peptide b1. Key binding residues on b1 are shown in stick representation. (**C**). HEK293T cells were co-transfected with combinations of GFP-tagged S1$^{493-685}$ and S1$^{493-685}$ R685D, and mCherry or mCherry-NRP1 b1, and subjected to mCherry-nanotrap (N=5). Two-tailed unpaired t-test; P <0.0001. (**D**). HEK293T cells were co-transfected with combinations of GFP-tagged S1$^{493-685}$ and mCherry, mCherry-NRP1 b1 or mCherry-NRP1 b1 T316R mutant and subjected to mCherry-nanotrap (N=5). Two-tailed unpaired t-test; P <0.0001. (**E**) HeLa^{NRP1KO} + ACE2 cells transfected with GFP, NRP1 wt-GFP or NRP1 T316R-GFP constructs were infected 24 h later with SARS-CoV-2. At 16 hpi the cells were fixed and stained for SARS-CoV-2-N, and viral infection quantified in the GFP-positive subpopulation of cells (N=3). The percentage of infection was normalized to that of GFP-transfected cells. Two-tailed unpaired t-test; p = 0.002. The bars, error bars and circles represent the mean, SEM (**C-D**) and SD (**E**), individual data points, respectively. *P< 0.05, **P< 0.01, ***P< 0.001, ****P< 0.0001.

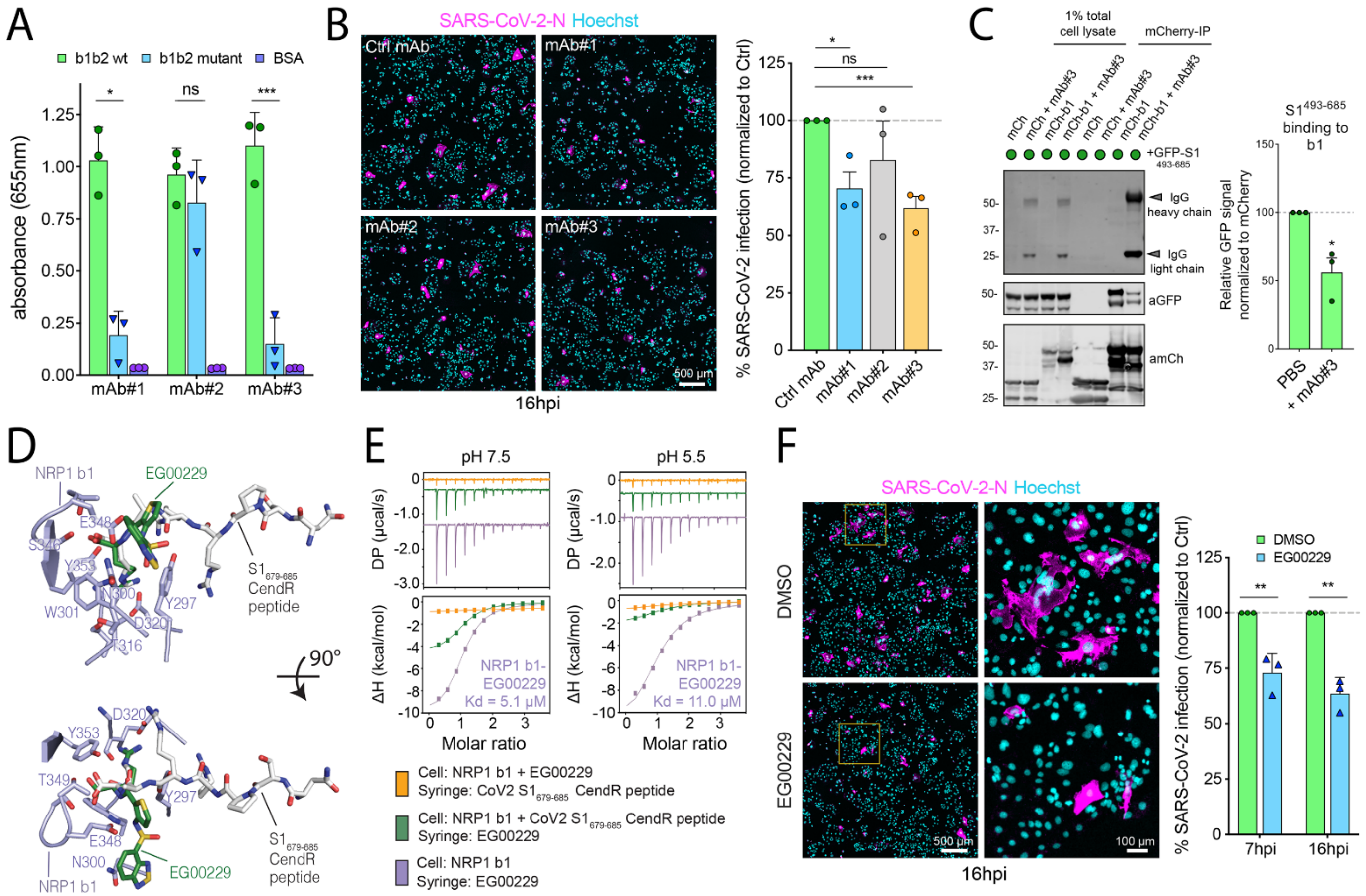


Fig. 3. Selective inhibition of the S1-NRP1 interaction reduces SARS-CoV-2 infection. (**A**) ELISA of anti-NRP1 monoclonal antibodies (mAb#1, mAb#2, mAb#3) at 3 μg/mL using plates coated with NRP1 b1b2 wild type, b1b2 mutant (S346A, E348A, T349A) or BSA, used as control (N=3). Binding is represented as arbitrary units of absorbance at 655 nm. Two-tailed unpaired t-test; P = 0.0207, 0.2430, 0.0007. (**B**) Cells were pre-treated with 100 μg/mL of anti-H11N3 (Ctrl) mAb, mAb#1, 2 or 3 for 1 h prior to infection with SARS-CoV-2. Cells were fixed at 16 hpi and stained for N protein (magenta) and Hoechst (cyan) (N=3). Two-tailed unpaired t-test; P=0.015, 0.36, 0.0003. Scale bar=500 μm. (**C**) HEK293T cells were co-transfected with combinations of mCherry or mCherry-b1 and GFP-tagged S1$^{493-685}$ and subjected to mCherry-nanotrap with or without co-incubation with mAb#3 (N=3). Two-tailed unpaired t-test; P = 0.0143. (**D**) NRP1 b1 – S1 CendR peptide complex superimposed with NRP1 b1 – EG00229 inhibitor complex (PDB ID:3I97). Key binding residues on b1, bound peptides and EG00229 are shown in stick representation. (**E**) ITC analysis of EG00229 binding to b1 domain of NRP1 at two different pH conditions. Pre-incubation with EG00229 blocks S1 CendR peptide binding (orange line), and the CendR peptide can reduce binding of EG00229 (green line). (N=3). All ITC graphs represents the integrated and normalized data fit with 1-to-1 ratio binding. (**F**). Cells were pre-treated with 100 μM of EG00229 or DMSO prior to infection with SARS-CoV-2. Cells were fixed at 7 and 16 hpi and stained for N protein (magenta) and Hoechst (cyan) (N=3). The square region was zoomed in. Scale bars=500 μm and 100 μm (zoom panel). Two-tailed unpaired t-test; P = 0.0059 and 0.0013. The bars, error bars, circles and triangles represent the mean, SEM (**C**) and SD (**A**, **B**, **F**) and individual data points, respectively. $*P< 0.05$, $**P< 0.01$, $***P< 0.001$, $****P< 0.0001$.

Science

Neuropilin-1 is a host factor for SARS-CoV-2 infection

James L. Daly, Boris Simonetti, Katja Klein, Kai-En Chen, Maia Kavanagh Williamson, Carlos Antón-Plágaro, Deborah K. Shoemark, Lorena Simón-Gracia, Michael Bauer, Reka Hollandi, Urs F. Greber, Peter Horvath, Richard B. Sessions, Ari Helenius, Julian A. Hiscox, Tambet Teesalu, David A. Matthews, Andrew D. Davidson, Brett M. Collins, Peter J. Cullen and Yohei Yamauchi

published online October 20, 2020

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02628-4 OPEN

Intelligent image-based in situ single-cell isolation

Csilla Brasko[1], Kevin Smith[2,3], Csaba Molnar[4], Nora Farago[1,4,5], Lili Hegedus[4], Arpad Balind[4], Tamas Balassa[4], Abel Szkalisity[4], Farkas Sukosd[1], Katalin Kocsis[1], Balazs Balint [6], Lassi Paavolainen[7], Marton Z. Enyedi[4], Istvan Nagy [4,6], Laszlo G. Puskas[4,5], Lajos Haracska[4], Gabor Tamas[1] & Peter Horvath[4,7]

Quantifying heterogeneities within cell populations is important for many fields including cancer research and neurobiology; however, techniques to isolate individual cells are limited. Here, we describe a high-throughput, non-disruptive, and cost-effective isolation method that is capable of capturing individually targeted cells using widely available techniques. Using high-resolution microscopy, laser microcapture microscopy, image analysis, and machine learning, our technology enables scalable molecular genetic analysis of single cells, targetable by morphology or location within the sample.

[1] University of Szeged, Szeged, Hungary Közép fasor 52, 6726 Szeged Hungary. [2] School of Computer Science and Communication, KTH Royal Institute of Technology, Lindstedtsvägen 3-5, 10044 Stockholm Sweden. [3] Science for Life Laboratory, Tomtebodavägen 23A, 17121 Solna Sweden. [4] Biological Research Centre of the Hungarian Academy of Sciences, Temesvári krt. 62., 6726 Szeged Hungary. [5] Avidin Biotechnology Ltd, Alsó Kikötő sor 11, 6726 Szeged Hungary. [6] SeqOmics Biotechnology Ltd, Vállalkozók útja 7, 6782 Mórahalom Hungary. [7] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Tukholmankatu 8, Helsinki 00014, Finland. Csilla Brasko, Kevin Smith, Csaba Molnar, Gabor Tamas and Peter Horvath contributed equally to this work. Correspondence and requests for materials should be addressed to P.H. (email: horvath.peter@brc.mta.hu)

Much of our current understanding of biology is built upon population-averaged measurements, including many models for cellular networks and signaling[1]. However, measurements averaging the behavior of large populations of cells can lead to false conclusions if they mask the presence of rare but critical subpopulations[2]. It is now well recognized that heterogeneities within a small subpopulation can carry important consequences for the entire population. For example, genetic heterogeneity plays a crucial role in drug resistance and the survival of tumors[3]. Even genetically homogeneous cell populations possess large degrees of phenotypic cell-to-cell variability due to individual gene expression patterns[4]. To better understand biological systems with cellular heterogeneity, we increasingly rely on single-cell molecular analysis methods[5]. However, single-cell isolation, the process by which we target and collect individual cells for further study, is still technically challenging and lacks a perfect solution.

A number of isolation methods are capable of collecting cells based on certain single-cell properties in a high-throughput manner, including fluorescence-activated cell sorting (FACS), immunomagnetic cell sorting, microfluidics, and limiting dilution[6,7]. However, these harvesting techniques disrupt and dissociate the cells from the microenvironment, and they are incapable of targeting the cell based on location within the sample or by phenotypic profile. In contrast, micromanipulation and laser capture microdissection[8] (LCM) are microscopy-based alternatives that directly capture single cells from suspensions or solid tissue samples. They can target cells by location or phenotype, and this contextual information can provide important insights when interpreting data from genetic analysis. LCM and micromanipulation methods can isolate specific subpopulations without substantial disruption of the tissue while limiting contamination (e.g., from chemical treatments needed for FACS). This is an important advantage for assaying single-cell gene expression and molecular processes. Recently, other single-cell isolation techniques have been introduced to perform mass spectrometry on single cells[9]. However, all these methods have a crucial limitation—they require manual operation to choose cells for isolation and to precisely target and extract them. These human-operated steps are error-prone and laborious, which greatly limits capacity.

We developed a technique to increase the accuracy and throughput of microscopy-based single-cell isolation by automating the target selection and isolation process. Computer-assisted microscopy isolation (CAMI) combines image analysis algorithms, machine-learning, and high-throughput microscopy to recognize individual cells in suspensions or tissue and automatically guide extraction through LCM or micromanipulation. To demonstrate the capabilities of our approach, we conducted three sets of experiments that require targeted single-cell isolation to collect individual cells without disturbing their microenvironment. We show that CAMI-selected cells can be successfully used for digital PCR (dPCR) and next-generation sequencing through these experiments.

Results

The CAMI system. A diagram summarizing CAMI technology is provided in Fig. 1. During preparation, samples are collected in variable formats etched with registration landmarks (Supplementary Note 1), and potentially treated with compounds according to the assay (Fig. 1a). Samples may come from tissue or cell cultures, and they are imaged with an automated high-throughput microscope (Fig. 1b). Images from the microscope are sent to our image analysis software that uses state-of-the-art algorithms to correct illumination, identify and segment cells (even in cases of overlap, Supplementary Note 2)[10], and extract multiparametric cellular measurements[11] (Fig. 1c). Advanced Cell Classifier software[12] trains machine-learning algorithms to automatically recognize the cellular phenotype of every cell in the sample based on their extracted properties (Fig. 1d), and these data along with the location and contour of each cell are sent to our interactive online database computer-aided microscopic isolation online (CAMIO; Fig. 1e). CAMIO provides an interface to approve the cells chosen to be extracted. If the user wishes, he/she may add or remove cells, or correct mistakes in the contour and classified phenotype. Selected cells are then extracted by micromanipulation or laser microdissection combined with a catapulting system (Fig. 1f) and collected in a microtube or high-throughput format for molecular characterization such as sequencing or dPCR (Fig. 1g). The software components we developed to support this technology are freely available (Supplementary Software).

As a proof of principle, we conducted three sets of experiments to demonstrate the capabilities of the technology to target, isolate, and analyze individual cells without disturbing their microenvironment (Fig. 2). These experiments were chosen because they could not have been analyzed using conventional automated isolation techniques (e.g., FACS), and alternative solutions would have required laborious manual operation.

Cell selection by phenotype validated by dPCR. First, we check whether immunofluorescent-labeled cells selected using machine-learning in CAMI corresponded to mRNA quantification in individual neurons extracted from 10 µm thick sections of the rat cerebral cortex. To accomplish this, we applied fluorescent labels to tissue fixed in 4% paraformaldehyde using immunohistochemistry with nNOS antibodies (Fig. 2a). Then, we automatically targeted and extracted individual cells that were predicted to belong to two phenotypic categories using CAMI technology (Fig. 2b). Cells that were most confidently predicted to be nNOS-expressing interneurons and non-labeled pyramidal cells were selected and isolated using laser microdissection (Fig. 2c). These individual cells were then catapulted and collected in PCR tubes containing SingleCellProtect stabilization and lysis buffer and directly used for cDNA conversion (Fig. 2d). The cDNA mixture was divided into two parts and used for single-cell dPCR to measure nNOS and RS18 gene expression for each extracted cell[13]. The dPCR results confirm that single nNOS-expressing neurons were reliably separated from nearby cells of other types within the same tissue (Fig. 2e). We also checked that the RNA did not significantly degrade by confirming that the number of transcripts of RS18, a housekeeping gene used as a control and for normalization between cells, matched values previously observed using live-cell aspirates with good fidelity[13,14] (Supplementary Data 1).

Whole-transcriptome sequencing of pyramidal cells. Next, we applied CAMI to target pyramidal cells for isolation from the same cortex sections. Using an online cell isolation tool we developed CAMIO, we automatically identified pyramidal cells based on morphological image features and selected cells specifically from layers L2 and L3 of the somatosensory cortex. The cells were extracted using LCM, pooled in SingleCellProtect buffer, and amplified using a REPLI-g WTA Single Cell Kit that contains an optimized Phi 29 polymerase and uses multiple displacement amplification technology. After quality control, we prepared sequencing libraries from the purified cDNA, sequenced the fragments on an Ion Torrent PGM, and recorded a list of gene expression. The experiment was repeated for three biological replicates (50, 50, and 300 cells), and the whole-transcriptome

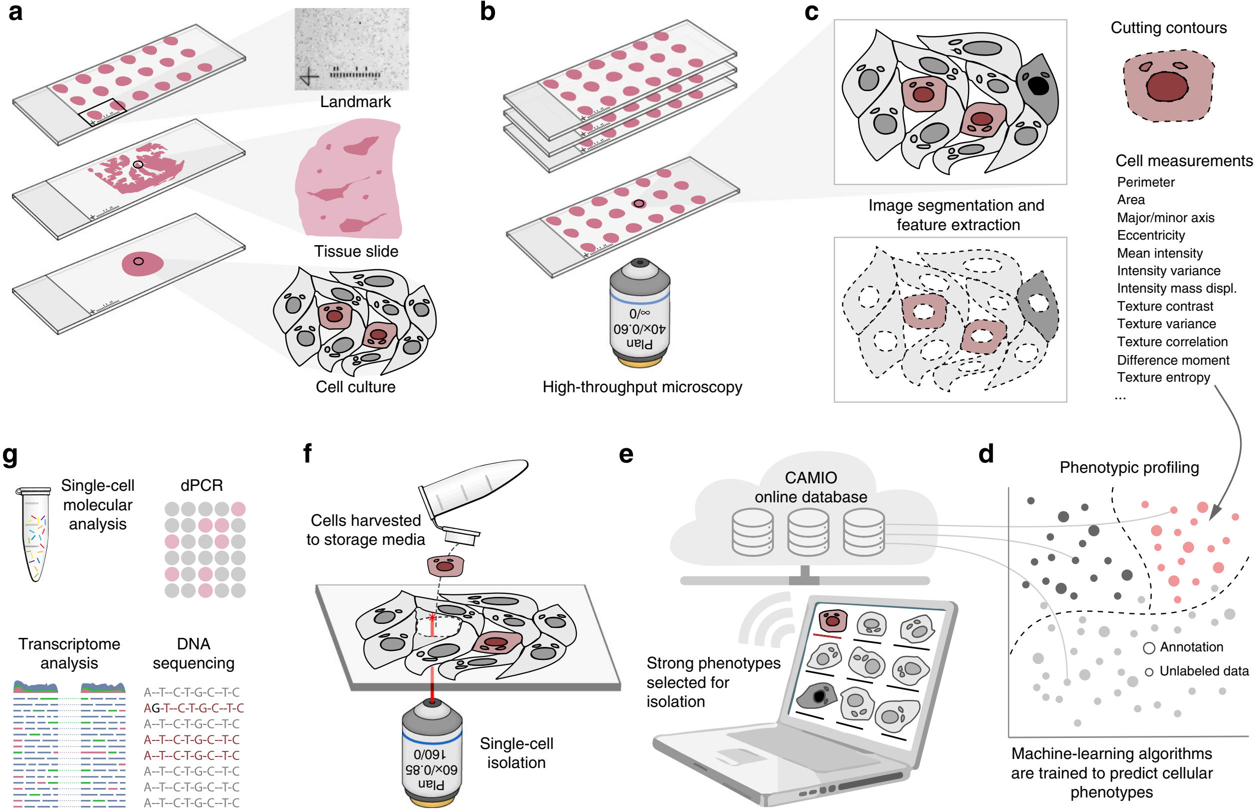


Fig. 1 Summary of computer-assisted microscopy isolation technology. **a** Tissue or cultured samples are prepared in a variety of formats, etched with registration landmarks, and treated according to the assay. **b** Samples are imaged with an automated high-throughput microscope. **c** Image analysis software applies algorithms to correct illumination, identify and segment cells (even in cases of overlap)[10], and extract multiparametric cellular measurements. Our software automatically defines a cutting contour using these data. **d** Advanced Cell Classifier software trains and optimizes machine-learning algorithms to automatically recognize cellular phenotypes based on extracted properties. **e** The raw images and analysis data are sent to our interactive online database, which provides an interface to review and select imaged cells. Cells exhibiting strong phenotypes are recommended for extraction. The user can add or remove cells or correct mistakes on the contour and classified phenotype prior to extraction. **f** Selected cells are extracted by micromanipulation or laser microdissection combined with a catapulting system and collected in a microtube or high-throughput format. **g** Outside the CAMI workflow, the collected cells can be molecularly characterized (e.g., digital PCR or next-generation sequencing)

profiles were compared (Fig. 2f). A comparison of the profiles revealed high correlations (Pearson's R) and high overlap in the top-100 expressed genes between the replicates (Supplementary Data 2, Supplementary Data 3). In a similar procedure, 50 astrocytes were also collected and sequenced, revealing negligible correlation with the pyramidal cells (Fig. 2f, Supplementary Data 2). This experiment shows that it is possible to automatically collect populations of a distinct type of cell from a specific region of fixed tissue in a high-throughput manner, and to perform reproducible whole-transcriptome sequencing using CAMI extraction.

Identification of upstream regulators by phenotyping. Last, we demonstrate that CAMI technology can provide a highly sensitive and cost-effective alternative to RNA interference (RNAi) library screens to uncover novel gene functions. While RNAi knockdowns test one gene at a time—measuring population responses (~20,000 experiments for a genome-wide library)—CAMI technology can be used to select individual cells from a mixture of stably silenced cell lines. Pooled cells exhibiting interesting phenotypes can be collected for further analysis, and the cell's silenced gene can be identified. The DNA of extracted cells is sequenced using universal primers flanking the specific silencing short hairpin RNA (shRNA)-coding region present in each cell of the library. As a proof of concept, we followed this approach to identify both known and novel genes involved in the response to DNA damage. We prepared a mixture of single shRNA-expressing stable human embryonic kidney cell lines (limited to 10 cell lines in our study). DNA damage was induced in the cells through UV exposure. In normal cells, this results in the recruitment of DNA repair proteins to the damage site and the formation of nuclear foci[15]. A fluorescent marker indicating polymerase η expression allowed us to visualize the formation of foci as spots within the nucleus (Fig. 2g). In the absence of upstream regulators, recruitment of repair proteins to the damage sites is prevented, resulting in a homogeneous expression of polymerase η (Fig. 2h). Using CAMI, we automatically identified 150 foci-forming and 150 homogeneous cells, captured them, and sequenced their shRNA-coding DNA region using next-generation sequencing (NGS). Our results confirm the identification of previously published upstream regulators of polymerase η (SPARTAN, BRCA2, and RAD18)[16–18], and identified RAD52 and FANCA as promising new potential regulators (Fig. 2i).

Discussion

LCM has been around for nearly 20 years[19], yet only now have the technologies matured sufficiently and computational techniques become sophisticated enough to support targeted automatic, environment-preserving, high-throughput single-cell isolation as we propose. Computer-driven automation increases throughput over manual techniques by orders of magnitude (from several hundred to over a thousand cells per day with CAMI, see

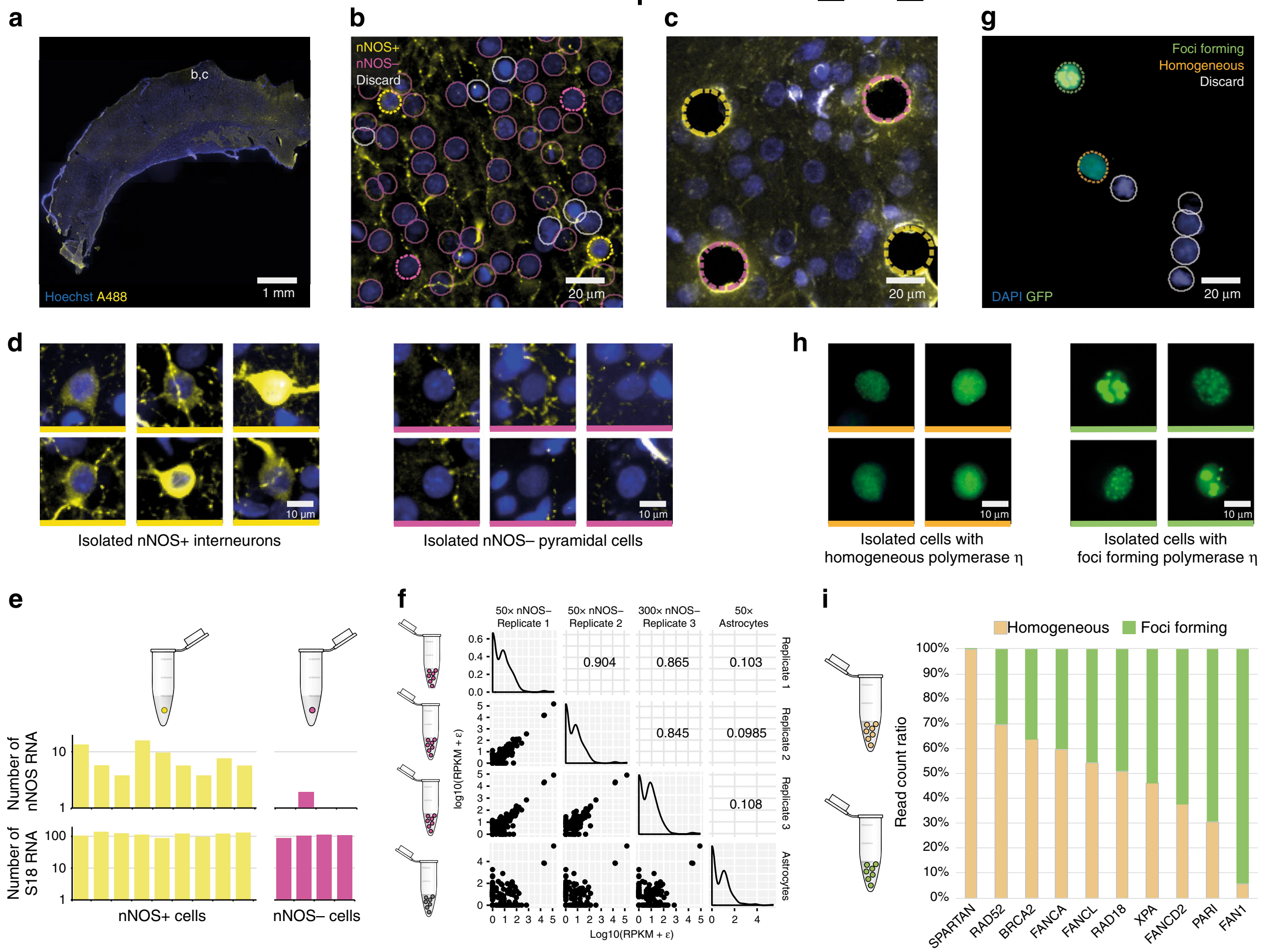


Fig. 2 Computer-assisted microscopy isolation (CAMI) opens the door to new types of high-throughput single-cell molecular analysis through non-disruptive collection of individual cells from fixed tissue and selection of cells by phenotypic morphology or location. **a** Coronal sections of rat brain labeled with mouse-anti-NeuN antibody (blue) and rabbit anti-nNOS antibody (yellow) were imaged with a high-throughput microscope. **b** High-resolution detail of a region of the somatosensory cortex indicated in **a**. Outlines show nuclear segmentations and phenotype classifications predicted by our software. Cells outlined in yellow are predicted to be nNOS+, cells outlined in magenta are nNOS−, and gray indicates cells that should be discarded (e.g., due to artifacts). Dotted lines indicate cells that were targeted for extraction. **c** The same region after extracting two nNOS+ and two nNOS− cells. **d** Individual cells automatically selected and extracted using CAMI, nNOS-expressing interneurons on the left and nonexpressing cells on the right. **e** Expression levels measured by dPCR show that CAMI reliably separates cells. Cells identified as nNOS+ show significantly higher expression (7.96 ± 0.48) than those identified as nNOS− (0.48 ± 0.95), two-sampled *t*-test $p = 0.0061$. Expression levels of housekeeping gene S18 did not vary significantly between cells identified as nNOS+ (116.37 ± 16.54) and nNOS− (103.98 ± 10.29), two-sampled *t*-test $p = 0.1992$. **f** Whole-transcriptome gene expression profiles of nNOS− cells (two 50-cell replicates and one 300-cell) and astrocytes (50 cells) extracted by CAMI and sequenced by Ion Torrent PGM. Analysis reveals strong correlations (Pearson's R) between the nNOS− replicates, and weak correlations between the astrocytes and nNOS− cells. **g** CAMI also enables a novel, cost-effective alternative to RNAi screening. Cells with interesting phenotypes are identified and extracted from mixed populations of stable shRNA-expressing silenced cell lines. After UV exposure, cells normally recruit polymerase η to repair DNA damage, which is visualized as foci by our green fluorescent marker. Absence of an upstream regulator can disrupt the foci formation and lead to homogeneous polymerase η expression. **h** CAMI identified 150 foci-forming and 150 homogeneous cells and extracted them. **i** Extracted cells were sequenced using next-generation sequencing (NGS). The ratio between the two populations revealed known upstream regulators of polymerase η (BRCA2, RAD18, and SPARTAN) and identified promising new regulators, Rad52 and FANCA

Supplementary Note 3, compared to 10 with patch-clamp harvesting), and microscopy-based isolation boasts several advantages over conventional high-throughput isolation techniques. These include non-disruptive collection of individual cells from fixed tissue or cell culture and selection of cells based on phenotypic morphology or location within the tissue. The throughput, precision, and versatility of CAMI enable new modes of highly reproducible molecular analysis and make it an attractive technique to drive new discoveries, for example, through alternative RNA and CRISPR/Cas9-screening approaches or through clinical applications using fresh or archived tissue samples.

Methods

Set-up. As a first step for every experiment, we etched 50 × 50 µm landmarks into poly-L-lysine-coated slides (one landmark per slide) using a microdissection microscope (Zeiss Axio Observer microscope with PALM MicroBeam manipulator). The landmarks are easily recognized by software and serve as an absolute zero position to register image data between microscopes. The landmarks were designed to indicate the orientation in order to avoid any errors due to rotation of the coordinates (Supplementary Note 1). An image of the landmark is also acquired and stored. Optionally, unique barcodes may also be etched into the slide to identify samples.

Tissue preparation. Male Wistar rats (between 200 and 350 g) were anesthetized by inhalation of 2-Bromo-2-chloro-1,1,1-trifluoroethane followed by intraperitoneal administration of 1 ml of 4% chloral hydrate per 100 g of body weight. Animals were then transcardially perfused with ice-cold saline for 2–4 min (10 ml/1 min) followed by 4% paraformaldehyde (PFA) made up in 0.1 M phosphate buffer (PB, pH = 7.4) for 10 min. Coronal sections of 10 µm thickness were cut with a Leica vibratome (Leica, VT 1000 S). All procedures were performed with the approval of the University of Szeged and in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Immunohistochemistry. Rat coronal sections were washed twice with 0.1 M PB for 10 min, followed by two washes with Tris-buffered saline (TBS, pH = 7.4) for 10 min. Sections were blocked for 2 h in a solution containing 20% normal horse serum and 0.2% Triton-X-100 made up in TBS. The sections were then incubated with primary antibodies to detect markers for neuronal cell populations including a mouse-anti-NeuN antibody (1:2000, MAB377, Chemicon, Temecula, CA) and a rabbit anti-nNOS antibody (1:1200, 160870 Cayman Chemical Company, Ann Arbor, MI). The antibodies were diluted in TBS and incubated for 2 days. After incubation, sections were washed four times with TBS for 10 min, and secondary antibodies including donkey anti-rabbit Alexa Fluor 488 (711-545-152, Jackson ImmunoResearch Laboratories, West Grove, PA), donkey anti-rabbit Cy3 (711-165-152, Jackson ImmunoResearch Laboratories), and donkey anti-mouse Alexa Fluor 568 (A10037, Thermo Fisher) were applied in 1:400 dilution and incubated for 2.5 h at room temperature. During the last 30 min of the incubation, Hoechst blue (Sigma, B2261) was added in 1 µg/ml concentration. Finally, sections were washed 3× with TBS for 10 min and then washed 2× with 0.1 M PB for 5 min before mounting in vectrashield (H-1000 Vector Laboratories, Burlingame, CA).

Imaging set-up and acquisition. Prior to extraction, a high-throughput screening campaign was performed using an automated imaging system (Operetta, PerkinElmer, Germany) allowing us to automatically analyze thousands of cells and to select the best examples for isolation. A 20× long working distance objective with 0% overlap was used to collect 1200 images with two fluorescent channels: Hoescht 333 and Alexa 568. The system we propose is compatible with any open format microscope where image position and pixel size can be measured, and has been successfully tested with a confocal slide scanner (Pannoramic Confocal, 3DHistech, Hungary) and a laser-scanning confocal microscope (FV 1000, Olympus, Japan) using 20× water and 40× oil emerging objectives. Alternatively, manual cell selection can be performed directly using the dissection microscope. However, throughput is significantly reduced.

Image analysis and pattern recognition. We developed a software pipeline to precisely outline every cell from the screen and to predict its phenotypic class. This software allows us to quickly visualize and select the best cell candidates from relevant subpopulations for isolation. The pipeline is composed of three steps: preprocessing, segmentation and feature extraction, and classification. The preprocessing step corrects artifacts due to uneven illumination present in the images using a quasi-newtonian optimization technique[20]. In the segmentation step, cells were outlined and properties were extracted using CellProfiler software[11] with custom pipelines. If individual cells were well separated, the default nuclei segmentation method was used: a seed-based adaptive Otsu thresholding. For cells in close proximity to one another, this method often fails. To overcome this, we used a two-step approach that first identifies nucleus centers using an *à trous* wavelet transform[21] and then expands the seeds to fit the boundaries using either CellProfiler secondary objects or high-order active contours[10] in the case of overlapping cells (Supplementary Note 2). This method allowed us to reliably identify cells with overlapping nuclei, which are typically discarded from molecular analysis. Custom CellProfiler modules implementing these methods and the pipelines used are provided (Supplementary Software). Nucleus segmentations were used to construct a polygon approximation of a 3 µm ring around the nucleus. This defines the cutting regions for isolation. It ensures that the laser does not destroy molecular information from the nucleus and also minimizes contamination from extracellular sources. After the segmentation step is complete, 92 single-cell properties describing the intensity, texture, and shape of the nuclei were extracted using CellProfiler and stored in the Advanced Cell Classifier (ACC) format[12].

We used supervised machine-learning algorithms to predict the phenotypic class for every cell in the screens based on the extracted features. Using Advanced Cell Classifier software[11], segmented objects were labeled according to their phenotypes. Using these data as a training set, ACC was used to train several machine-learning models using multiple methods to predict phenotypic class of all cells; 10-fold cross-validation was used to select the best-performing model. A random forest classifier achieved 91% cross-validation accuracy and was trained using every annotation (Supplementary Figure 1). It was then used to predict the phenotypic class for every cell. ACC software with modules to upload single-cell information for selected subpopulations to an online repository is included as Supplementary Software.

Single-cell online repository and selection tool. Cell phenotype predictions are ranked by confidence, and the 200 cells with highest confidence for the interneuron and pyramidal phenotype classes were automatically uploaded to CAMIO, an online single-cell data repository and selection tool we developed (Supplementary Software). The purpose of this tool is to visualize individual cells and facilitate the selection of appropriate candidates for isolation. Individual cells are displayed, organized by experiment and phenotypic class. Cells can be selected for isolation automatically or through manual verification. Selected cells are sent instantly to the single-cell isolation device. The CAMIO interface allows the user to verify and correct the proposed cutting regions for each cell. It also records the location of the etched landmark relative to each object. The CAMIO interface is shown in Supplementary Figure 2, and an online read-only version of the system can be tested at https://camio-webapp.herokuapp.com/.

Image coordinate registration between microscopes. To register data between microscopes, the landmark etched in each sample slide is automatically detected by our software using two-dimensional cross-correlation. The landmark location is used as the zero position and orientation reference to transform data from one microscope to another. The offset between the orientation landmark and the microscope coordinate system is recorded in the source microscope (high-throughput microscope) and recorded. It is also measured in the target microscope (laser microdissection microscope). With this information, coordinates defining the cutting region for a cell can be transformed from the source image coordinates to the target microscope coordinates using the following relation

$$(x^2, y^2)^T = (y^1, x^1)^T - (y^1_{\mathrm{off}}, x^1_{\mathrm{off}})^T + (x^2_{\mathrm{off}}, y^2_{\mathrm{off}})^T$$

where $x^1$and y^1 are the coordinates in the source microscope, x^1_{off} and y^1_{off} are the origin offsets in the source microscope, and x^2_{off} and y^2_{off} are the origin offsets in the target microscope. By applying this transform, contours of cells from the high-throughput microscope and CAMIO can be registered in the laser microdissection microscope.

Single-cell isolation. To prevent contamination, a custom-designed closed hood was mounted on the isolation microscope and a UV sterilizer was built in (UVR-M Biosan) that was run before every experiment for 30 min. Temperature in the hood and laboratory was 20 °C, and humidity was kept at 50–60% to prevent sample drying.

After cells were selected for isolation using the CAMIO online tool, samples were hydrated with 0.1 M PB. The tissues were initially overhydrated. Immediately prior to cell isolation, the liquid was entirely removed from the surface. This practice allowed a more flexible schedule when cutting. The cutting path for each cell was provided by CAMIO. As a last step before each cell was extracted, we acquired an image of the specimen in situ to document the cell before isolation. This allowed us to perform quality control and refer to the source image when examining results from further analysis. A Zeiss PALM laser microdissection microscope was used for isolation with a 63× LCM-compatible magnification objective (LD Plan-Neofluar, 63×). The cutting was performed using the ultraviolet (337 nm) N2 laser microbeam system of Zeiss PALM, emitting 3 ns pulses. The laser-cutting speed was 1% (~ 4.7 µm/s), and cutting time ranged between 10 and 20 s per cell, depending on the contour of the cell and stage velocity. The cutting energy varied between 36 and 48 µJ depending on the glass thickness. By keeping the laser pulses short and low-power, we promoted a "cold cutting" that is less harmful to the samples.

Isolated cells were pressure-catapulted into PCR tube caps containing 4 µl SingleCellProtect^{TM} (Avidin Ltd., Szeged, Hungary) buffer media facing downward for storage. To avoid dripping and evaporation of the media, microtubes were kept at −20 °C and catapulting was performed when the buffer transitions from frozen to liquid state (between 10 and 20 s after removal from the fridge). After collection, the tubes were closed and immediately stored at −80 °C.

Single-cell reverse transcription and dPCR of rat cortical neurons. Reverse transcription of individual microdissected cells was carried out in two steps. The first step was performed for 5 min at 65 °C in a total reaction volume of 7.5 µl containing the cell captured in 4 µl SingleCellProtect^{TM} (Avidin Ltd., Cat.No.: SCP-250), 0.45 µl TaqMan Assays (Thermo Fisher), 0.45 µl 10 mM dNTPs (Thermo Fisher, Cat.No.: 10297018, 1.5 µl 5× first-strand buffer, 0.45 µl 0.1 mol/l DTT, 0.45 µl RNase inhibitor (Thermo Fisher, Cat.No.:N8080119), and 100 U of reverse transcriptase (Superscript III, Thermo Fisher, Cat.No.: 18080055). The second step of the reaction was carried out at 55 °C for 1 h, and then the reaction was stopped by heating at 75 °C for 15 min. The reverse transcription reaction mix was stored at −20 °C until PCR amplification.

For dPCR analysis, the reverse transcription reaction mixture (7.5 µl) was divided into two parts: 6 µl was used for amplification of the gene of interest and 1.5 µl cDNA was used for amplifying the housekeeping gene, RS18. Template cDNA was supplemented with nuclease-free water to a final volume of 8 µl. TaqMan Assays (2 µl; Thermo Fisher), 10 µl OpenArray Digital PCR Master Mix (Thermo Fisher, Cat.No.: 4458095), and nuclease-free water (3 µl) were mixed to obtain a total volume of 20 µl, and the mixture was evenly distributed on four subarrays (256 nanocapillary holes) of an OpenArray plate by using the OpenArray autoloader. Processing of the OpenArray slide, cycling in the OpenArray NT cycler, and data analysis were done as previously described[22]. For our dPCR protocol amplification, reactions having CT values less than 23 or greater than 33 were considered primer dimers or background signals, respectively, and excluded from the data set.

The following Taqman Assays were used: RS18 (Thermo Fisher, Cat.No.: 4331182, Rn01428913_gH), NOS1 (Thermo Fisher, Cat.No.: 4331182, Rn00583793_m1), NPY (Thermo Fisher, Cat.No.: 4331182, Rn01410145_m1).

Whole-transcriptome sequencing of rat cortical neurons. For RNA and subsequent cDNA amplification, REPLI-g WTA Single Cell Kit (Qiagen, Cat.No.: 150063) was used with the Amplification of Total RNA from Single Cells' protocol according to the manufacturer's guidelines with the exception that 3 µl Lysis buffer was added to 8 µl 1× SingleCellProtect solution containing either 50 or 300 collected cells (three replicates were collected—two with 50 cells and one with 300 cells, denoted 50× nNOS—Replicate 1, 50× nNOS—Replicate 2, and 300× nNOS —Replicate 3). In addition, 50 astrocytes were collected, denoted 50× Astrocytes). All subsequent steps were performed as described in the protocol manual. The quality and quantity control of cDNA pools were performed on TapeStation using genomic DNA ScreenTape and Reagents (Agilent Technologies, Cat.No.: 5067-5365 and 5067-5365) and Qubit using dsDNA High-Sense assay (Thermo Fisher, Cat.No.: Q32854), and were purified using Agencourt AMPure XP magnetic beads (Beckman Coulter, Cat.No.: A63881). Fragment libraries were constructed from purified cDNA using NEBNext Fast DNA Fragmentation & Library Prep Set for Ion Torrent (New England Biolabs, Cat.No.: E6285) according to the manufacturer's instructions. Briefly, cDNA was enzymatically digested, and the fragments were end-repaired; the fragmentation time was adjusted to cDNA quality and quantity (generally 5–8 min of fragmentation). Fragmented cDNA pools were purified with Agencourt AMPure XP magnetic beads. Purified fragments were end-repaired, Ion Xpress Barcode Adaptors (Thermo Fisher, Cat.No.: 4474521) were then ligated and the template fragments size-selected using AMPure beads. Adaptor-ligated fragments were PCR-amplified, cleaned-up using AMPure beads, quality-checked on D1000 ScreenTape and Reagents using TapeStation instrument (Agilent Technologies, Cat.No.: 5067-5582 and 5067-5583), and finally quantified using Ion Library TaqMan Quantitation Kit (Thermo Fisher, Cat.No.: 4468802). The library templates were processed for sequencing using the Life Technologies Ion OneTouch protocols and reagents. Library fragments were clonally amplified onto Ion Sphere Particles (ISPs) through emulsion PCR and then enriched for template-positive ISPs. More specifically, Ion PGM emulsion PCR reactions utilized the Ion OneTouch 200 Template Kit (Thermo Fisher, Cat.No.: 4480974), and emulsions and amplification were generated using the Ion OneTouch System (Thermo Fisher). Enrichment was completed by selectively binding the ISPs containing amplified library fragments to streptavidin-coated magnetic beads, removing empty ISPs through washing steps, and denaturing the library strands to allow for collection of the template-positive ISPs. For all reactions, these steps were accomplished using the ES module of the Ion OneTouch System. Template-positive beads were deposited onto the Ion 318 chips (Thermo Fisher, Cat.No.: 4484354); finally, sequencing was performed with the Ion PGM Hi-Q Sequencing Kit (Thermo Fisher, Cat.No.: A25592) on Ion Torrent PGM instrument generating between 2.9 and 5.3 million reads per sample.

Ion Torrent PGM sequencing data processing and expression analysis. The PGM sequencing data were processed using Genomics Workbench ver 9.0.1 (CLC Bio). Raw sequencing data were trimmed by removal of low-quality (quality limit: 0.05) and short (length limit: 40 bases) sequences so that only high-quality sequences were used in further analysis. Sequences were mapped on the Rattus norvegicus 6.0 genome (Rnor_6.0) using the CLC RNA-Seq algorithm, allowing mapping to intergenic regions, using default parameters except for the following: minimum alignment length 80%, minimum similarity 80% with the maximum number of hits for a read set to 30. Total read counts were used as a measure of gene expression in all samples.

The level of correlation between the biological replicates was determined by using the Pearson's product–moment correlation coefficient (PCC) which infers the linear relationship between two data sets based on the covariance and SD from the expression values. These values computed between each nNOS− replicate and the astrocytes are provided in Fig. 2f.

Sample preparation to detect subnuclear foci formation in human cells. HEK293 cells stably expressing different shRNAs were harvested in Dulbecco's modified Eagle's medium (Sigma, Cat. No. D6429) supplemented with 10% fetal bovine serum (Gibco, Cat. No. 10270) and 300 µg/ml G418 (Biochrom, Cat. No. A291-25) at 37 °C. The cells were transfected with GFP-polymerase η-expressing plasmid using the Lipofectamine 2000 transfection reagent (Invitrogen, Cat. No. 11668). Cells were plated in six-well plates 24 h before transfection. Then, the growth media was removed and changed to 1.5 ml OptiMEM per well. An amount of 3 µg plasmid DNA and 5 µl Lipofectamine 2000 reagent were used for each well. Both the DNA and the transfection reagent were diluted in 250–250 µl OptiMEM, mixed by vortexing, and incubated for 5 min. After mixing the two tubes, the solution was further incubated for 20 min, added to the cells dropwise and incubated for 4 h, and then the transfection media were removed and changed to 3 ml fresh growth media.

After 48 h had elapsed since transfection, cells were exposed to 20 J/m^2 UVC light to induce DNA damage and polymerase η foci formation. After 3 h of incubation, cells were counted in a Burker chamber and mixed in the same amounts to avoid the over-representation of any type of cell line. The fixation step was carried out using 3% PFA solution for 10 min. The cells were suspended and dropped to poly-L-lysine-covered slides. After the fixation, the sample was washed with PBS, followed by the staining of the nuclei with 0.5 µg/ml DAPI solution in PBS and then washing with MQ. The samples were kept in a humidity chamber until microscopic analysis to prevent drying.

Direct amplification and sequencing of shDNA fragments from human cells. Cells were captured in 5 µl catapult buffer, 150 cells for each phenotype (0.1 mM EDTA, 1 mM Tris pH 8, 0.5% Igepal). After capture, we added 0.5 µl Proteinase K (1 mg/ml) to the samples and incubated them at 60 ˚C for 20 min, followed by 3 min at 98 ˚C. Next, a two-step amplification reaction was carried out in 20 µl volume. In the first PCR, we used 10 µM shDNA-specific primer pair with a universal tag sequence, 1 × PCR buffer, 2.0 mM MgCl$_2$, 2.5 mM dNTPs, and 1 unit of AmpliTaq Gold DNA Polymerase (Thermo Fisher, Cat. No. 8080241). Thermal cycler conditions were: 95 °C for 2 min, 25 cycles of 95 °C for 15 s, 60 °C for 15 s, 72 °C for 30 s, and finally 2 min at 72 °C. In the second PCR, 1 µl from the first amplification reaction was used as a template with primers complementary to the universal tag sequence. The 5' end of the primers consisted of Illumina specific adaptor sequences. Other PCR conditions were the same as the first ones performing 30 PCR cycles this time. PCR reactions were run on a 2% agarose gel for amplification quality control. Successfully amplified samples were quantified using the qPCR-based quantification method (Kapa Biosystems, Cat. No. KK4854) on LightCycler480 qPCR (Roche, Indianapolis, IN). Finally, Illumina sequencing was carried out on the Illumina MiSeq system with Standard Flow Cell v2 (Illumina, Cat. No. MS-102-2002), following the manufacturer's instructions. Sequencing data were analysed using proprietary NGSeXplorer bioinformatics software. Sequencing reads were mapped to a reference sequence that contained all the 10 shDNA-specific sequences. Read counts were measured at each shDNA sequence. These data were then used to calculate the shDNA patterns for the two phenotypic groups (foci-forming and homogeneous cells).

Limiting factors and sources of error. Our approach is capable of isolating from several hundred to over a thousand cells per day. However, the throughput of CAMI is still limited by several bottlenecks. In the imaging and set-up stage, the main bottleneck is the microscope set-up (locating the landmark, configuring the microscope settings, finding focus, and selecting the region of interest). This comes at a fixed time cost per sample slide (see Supplementary Note 3 for timings). In the software analysis stage, the main bottleneck is processing enough images to find the desired number of cells for isolation. Rare phenotypes require searching through more images to find interesting cells to isolate. In the isolation stage, the main bottleneck is the laser microdissection, which takes ~10 s per cell (catapulting and stage movement are substantially quicker than cutting). Depending on the experimental parameters (rarity of the phenotype, number of desired cells, etc), the most costly bottleneck changes (Supplementary Note 3). When collecting relatively few cells (less than 100) of a common phenotype, the imaging and its set-up is the limiting factor. For very rare cell types that require the software to process thousands of images to find isolation candidates, the analysis software is the limiting factor (although this can be mitigated using distributed computing). When a large number of cells are desired (more than 1000), laser-cutting is the limiting factor.

Because CAMI relies on a diverse set of complex technologies, there exist several potential sources of error. It is difficult to mount a slide in perfect alignment with the stage for different microscopes, so there is potential for angular misalignment between microscope coordinates (multiple landmarks can mitigate this). When setting up the microscope, the user must select the appropriate areas of the sample to image or they may have difficulty finding isolation candidates. When imaging, problems with focus and artifacts in the image can cause errors in cell segmentation. The segmentation software itself is prone to errors, and the machine-learning predictions are imperfect. Errors in cell isolation contours may be caused by imperfect registration between microscope coordinates, which can result in poorly cut cells. In the laser microdissection step, losing focus or choosing the wrong cutting speed may result in failure to properly cut the cell (slow cutting can burn the cell, fast cutting can cause problems ejecting). Finally, the catapulting laser must be correctly calibrated or cells may be lost in the collection step.

Data availability. The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files.

Received: 1 January 2017 Accepted: 14 December 2017
Published online: 15 January 2018

References

1. Altschuler, S. & Wu, L. Cellular heterogeneity: when do differences make a difference? *Cell* **141**, 559–563 (2010).
2. Pelkmans, L. Using cell-to-cell variability—a new era in molecular biology. *Science* **336**, 425–426 (2012).
3. Heppner, G. H. Tumor heterogeneity. *Cancer Res.* **44**, 2259–2265 (1984).
4. Tay, S. et al. Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267–271 (2010).
5. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
6. Anselmetti, D. *Single Cell Analysis: Technologies and Applications*. (Wiley, Hoboken, 2009).
7. Gross, A. et al. Technologies for single-cell isolation. *Int. J. Mol. Sci.* **16**, 16897–16919 (2015).
8. Espina, V. et al. Laser-capture microdissection. *Nat. Protoc.* **1**, 586–603 (2006).
9. Fujii, T. et al. Direct metabolomics for plant cells by live single-cell mass spectrometry. *Nat. Protoc.* **10**, 1445–1456 (2015).
10. Molnar, C. et al. Accurate morphology preserving segmentation of overlapping cells based on active contours. *Sci. Rep.* **6**, 32412 (2016).
11. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
12. Piccinini, F. et al. Advanced cell classifier: user-friendly machine-learning-based software for discovering phenotypes in high-content imaging data. *Cell Syst.* **4**, 651–655 (2017).
13. Farago, N. et al. Digital PCR to determine the number of transcripts from single neurons after patch-clamp recording. *Biotechniques* **54**, 327–336 (2013).
14. Molnar, G. et al. GABAergic neurogliaform cells represent local sources of insulin in the cerebral cortex. *J. Neurosci.* **34**, 1133–1137 (2014).
15. Friedberg, E. C. et al. *DNA Repair and Mutagenesis*. (ASM Press, Washington, DC, 2006).
16. Watanabe, K. et al. Rad18 guides polη to replication stalling sites through physical interaction and PCNA monoubiquitination. *EMBO J.* **23**, 3886–3896 (2004).
17. Juhasz, S. et al. Characterization of human Spartan/C1orf124, an ubiquitin-PCNA interacting regulator of DNA damage tolerance. *Nucleic Acids Res.* **40**, 10795–10808 (2012).
18. Buisson, R. et al. Breast cancer proteins PALB2 and BRCA2 stimulate polymerase η in recombination-associated DNA synthesis at blocked replication forks. *Cell Rep.* **6**, 553–564 (2014).
19. Emmert-Buck, M. R. et al. Laser capture microdissection. *Science* **274**, 998–1001 (1996).
20. Smith, K. et al. CIDRE: an illumination-correction method for optical microscopy. *Nat. Methods* **12**, 404–406 (2015).
21. Olivo-Marin, J. C. Extraction of spots in biological images using multiscale products. *Pattern Recognit.* **35**, 1989–1996 (2002).
22. Farago, N. et al. Human neuronal changes in brain edema and increased intracranial pressure. *Acta Neuropathol. Commun.* **4**, 78 (2016).

Acknowledgements

A.B., C.B., T.B. and P.H. acknowledge the Hungarian National Brain Research Program (MTA-SE-NAP B-BIOMAG). P.H. and L.P. acknowledge support from the Finnish TEKES FiDiPro Fellow Grant 40294/13. N.F. and G.T. were supported by the National Research, Development and Innovation Office of Hungary (VKSZ-14-1-2015-0155), the Hungarian Academy of Sciences and the ERC INTERIMPACT project. P.H. and I.N. were supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. L.H. and P.H. acknowledge the European Union and the European Regional Development Funds (GINOP-2.3.2-15-2016-00006, GINOP-2.3.2-15-2016-00020).

Author contributions

P.H. and K.S. conceived the project. P.H. led the project. G.T. and L.H. co-supervised the RNA and DNA studies, respectively. C.B., A.B., F.S. and L.H. executed the pipeline and performed single-cell isolation. C.M., A.B., L.P., T.B., A.S. and P.H. wrote the software components. C.B., N.F. and L.G.P. performed dPCR analysis. C.B., B.B. and I.N. performed transcriptome analyses. L.H., M.E. and L.H. performed DNA analysis. K.S., C.B., L.H., C.M., A.B., I.N., L.P., L.H., G.T. and P.H. wrote the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02628-4.

Competing interests: P.H. is the founder and shareholder of Single-cell technologies Ltd.; L.G.P. is the founder and shareholder of Avidin Ltd. that provided SingleCellProtect reagent; B.B. and I.N. had consulting positions at SeqOmics Biotechnology Ltd. at the time the study was conceived. Single-cell technologies Ltd., Avidin Ltd., and Seqomics Biotechnology Ltd. were not directly involved in the design and execution of the experiments or in the writing of the manuscript. This does not alter the author's adherence to all the *Nature* policies on sharing data and materials. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21291-4 OPEN

Automatic deep learning-driven label-free image-guided patch clamp system

Krisztian Koos[1,7], Gáspár Oláh[2,7], Tamas Balassa[1], Norbert Mihut[2], Márton Rózsa[2], Attila Ozsvár[2], Ervin Tasnadi[1], Pál Barzó[3], Nóra Faragó[2,4,5], László Puskás[4,5], Gábor Molnár[2], József Molnár[1], Gábor Tamás[2] & Peter Horvath[1,6]✉

Patch clamp recording of neurons is a labor-intensive and time-consuming procedure. Here, we demonstrate a tool that fully automatically performs electrophysiological recordings in label-free tissue slices. The automation covers the detection of cells in label-free images, calibration of the micropipette movement, approach to the cell with the pipette, formation of the whole-cell configuration, and recording. The cell detection is based on deep learning. The model is trained on a new image database of neurons in unlabeled brain tissue slices. The pipette tip detection and approaching phase use image analysis techniques for precise movements. High-quality measurements are performed on hundreds of human and rodent neurons. We also demonstrate that further molecular and anatomical analysis can be performed on the recorded cells. The software has a diary module that automatically logs patch clamp events. Our tool can multiply the number of daily measurements to help brain research.

[1] Synthetic and Systems Biology Unit, Biological Research Centre, Eötvös Loránd Research Network, Szeged, Hungary. [2] MTA-SZTE Research Group for Cortical Microcircuits of the Hungarian Academy of Sciences, Department of Physiology, Anatomy and Neuroscience, University of Szeged, Szeged, Hungary. [3] Department of Neurosurgery, University of Szeged, Szeged, Hungary. [4] Laboratory of Functional Genomics, Institute of Genetics, Biological Research Centre, Szeged, Hungary. [5] Avidin Ltd, Szeged, Hungary. [6] Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [7] These authors contributed equally: Krisztian Koos, Gáspár Oláh. ✉email: horvath.peter@brc.hu

Research of the past decade uncovered the unprecedented cellular heterogeneity of the mammalian brain. It is well accepted now, that the complexity of the rodent and human cortex can be best resolved by classifying individual neurons into subsets by their cellular phenotypes[1–3]. By characterizing molecular, morphological, connectional, physiological, and functional properties several neuronal subtypes have been defined[4,5]. Revealing cell-type heterogeneity is still incomplete and challenging since classification based on quantitative features requires large amounts of individual cell samples, often thousands or more, encompassing a highly heterogeneous cell population. Recording morphological, electrophysiological, and transcriptional properties of neurons requires different techniques combined on the same sample such as patch clamp electrophysiology, posthoc morphological reconstruction, or single-cell transcriptomics. The fundamental technique to achieve such trimodal characterization of neurons is the patch clamp recording, which is highly laborious and expertise intense. Therefore, there is a high demand to efficiently automate this labor intense and challenging process.

Recently, the patch clamp technique has been automated and improved to a more advanced level[6,7]. Blind patch clamping was first done in vitro and only later performed in vivo[8–10]. In this case, the pipette is gradually moved forward and the brain cells are detected automatically by a resistance increase at the pipette tip. Automated systems soon incorporated image-guidance by using multiphoton microscopy on genetically modified rodents[11–13]. Further improvements include the integration of tools for monitoring animal behavior[14], the design of an obstacle avoidance algorithm before reaching the target cell[15] or the development of a pipette cleaning method which allows the immediate reuse of the pipettes up to ten times[16,17]. Automated multi-pipette systems were developed to study the synaptic connections[18,19]. It is also shown that cell morphology can be examined using automated systems[20]. One crucial step for image-guided automation is pipette tip localization. Different label-free pipette detection algorithms were compared previously[21]. Some automated patch clamp systems already contain pipette detection algorithms, e.g., intensity clustering[11] or thresholding-based[22] for fluorescence imaging, or Hough transform-based[23] for DIC optics. The other crucial step is the automatic detection of the cells which has only been performed in two-photon images so far. It is currently not possible to efficiently fluorescently stain human brain tissues. Alternatively, detection of cells in label-free images would open up new application possibilities in vitro[23], e.g., experiments on surgically removed human tissues. Most recently, deep learning[24] has been emerging to a level that in the case of well-defined tasks, outperforms humans, and often reaches human performance on ill-defined problems like detecting astrocyte cells[25].

In this paper, we describe a system we developed in order to overcome time-consuming and expertise-intense neuron characterization and collection. This fully automated differential interference contrast microscopy (DIC, or label-free in general) image-guided patch clamping system (DIGAP) combines 3D infrared video microscopy, cell detection using deep convolutional neural networks and a glass microelectrode guiding system to approach, attach, break-in, and record biophysical properties of the target cell.

The steps of the visual patch clamp recording process are illustrated in Fig. 1. Before the first use of the system, the pipette has to be calibrated, so that it can be moved relative to the field of view of the camera (1). Thereafter, a position update is made after every pipette replacement (2) using the built-in pipette detection algorithms (3) to overcome the problem caused by pipette length differences. At this point, the system is ready to perform patch clamp recordings. We have acquired and annotated a single cell image database on label-free neocortical brain tissues, to our knowledge the largest 3D set of this kind. A deep convolutional neural network has been trained for cell detection. The system can automatically select a detected cell for recording (4). When a cell is selected, multiple subsystems are started simultaneously

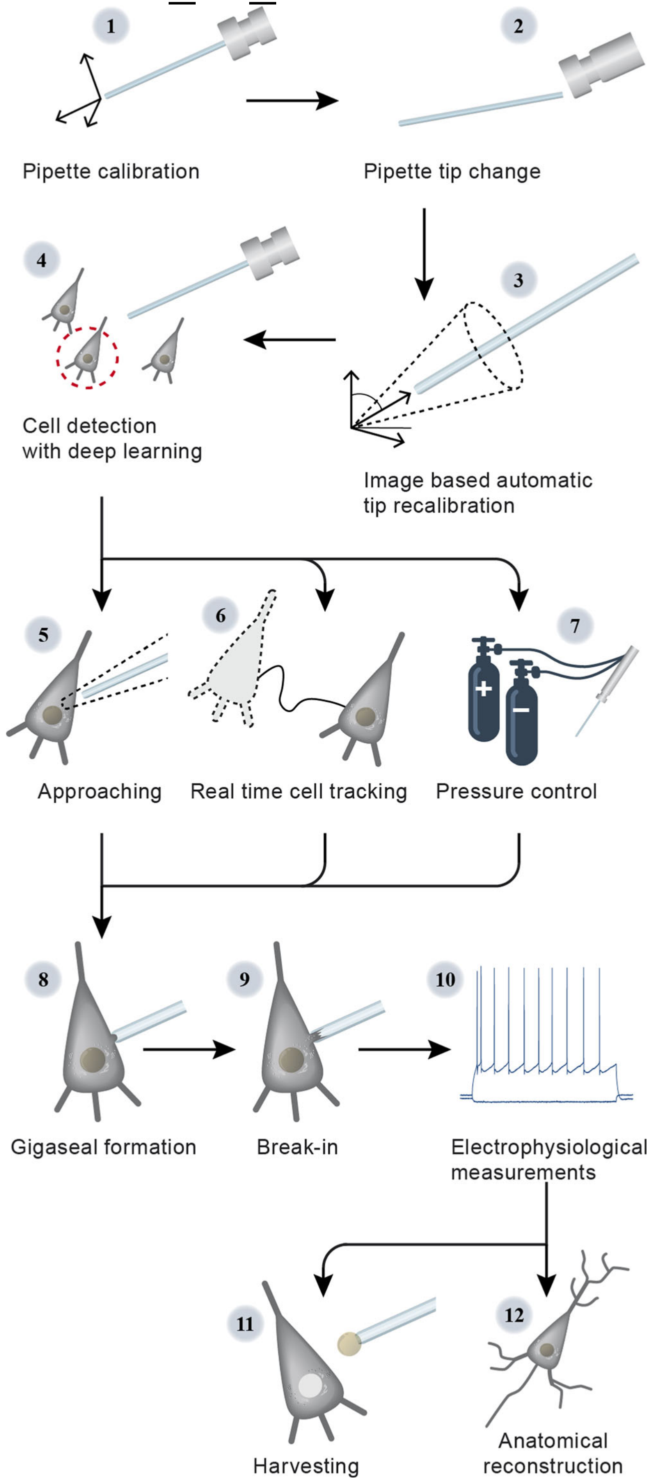


Fig. 1 Steps of DIGAP procedures. 1: Pipette calibration by the user, 2: pipette replacement after recording, 3: image-based automatic pipette tip detection, 4: automatic cell detection, 5: pipette navigation to the target cell, 6: 3D cell tracking, 7: pressure regulation, 8: gigaseal formation, 9: break-in, 10: electrophysiological recording, 11: nucleus and cytoplasm harvesting, 12: anatomical reconstruction of the recorded cell.

that perform the patch clamping: (i) A subsystem controls the movement of the micropipette next to the cell. If any obstacle is found in the way, an avoidance algorithm tries to bypass it (5). (ii) A cell tracking system follows the possible shift of the cell in 3D (6). (iii) During the whole process, a pressure regulator system assures that the requested pressure on the pipette tip is available (7).

Once the pipette touches the cell (cell-attached configuration) the system performs gigaseal formation (8), then breaks in the cell membrane (9) and automatically starts the electrophysiological measurements (10). When the recording is completed, the operator can decide either to start over the process on a new target cell or continue with one or both of the following manual steps. The nucleus or the cytoplasm of the patched cell can be harvested (11), or the recorded cells can be anatomically reconstructed in the tissue (12).

At the end of the measurements, the implemented pipette cleaning method can be performed or the next patch clamp recording can be started after pipette replacement and from the pipette tip position update step (3). An event logging system collects information during the patch clamp process, including the target locations and the outcome success, and report files can be generated at the end. The report files are compatible with the Allen Cell Types Database[26].

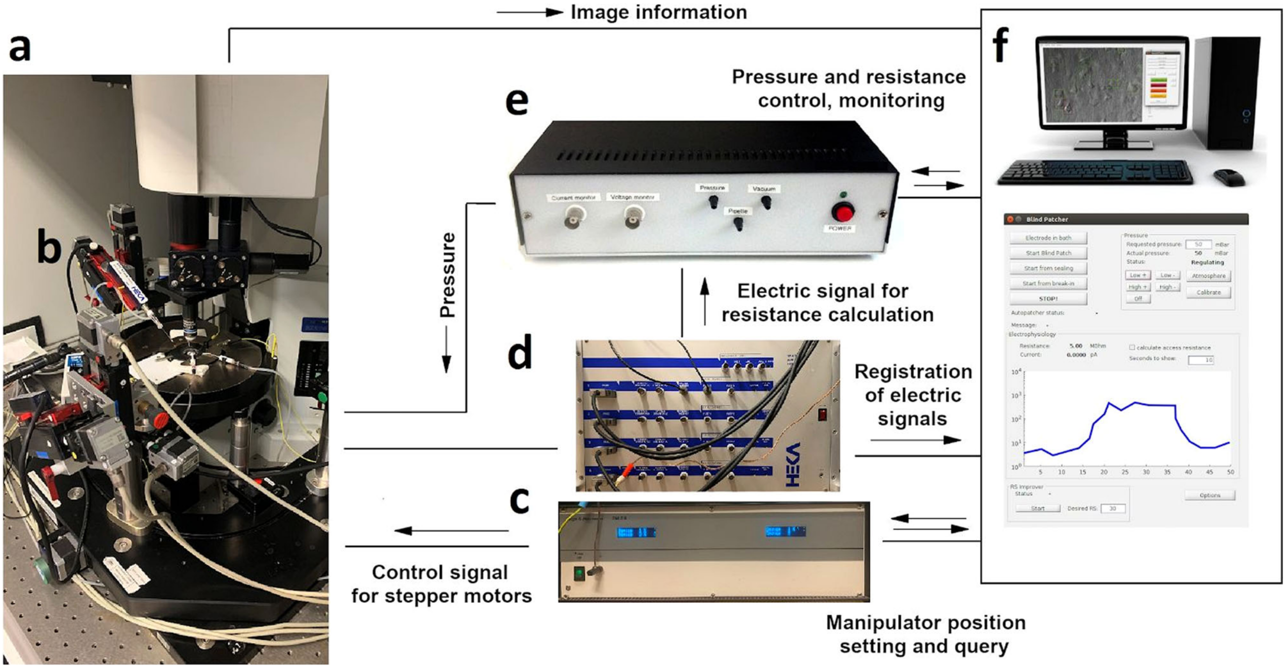
Our system was tested on rodent and human samples in vitro. The quality of the electrophysiological measurements strongly correlates to that made by a trained experimenter. We have used the system for harvesting cytoplasm and nucleus from the recorded cells and performed anatomical reconstruction on the samples. Our system can operate on unstained tissues using deep learning, that reaches the cell detection accuracy of human experts, and that enables the multiplication of the number of recordings while preserving high-quality measurements.

Results

Here, we introduce an automated seek-and-patch system that performs electrophysiological recordings and sample harvesting for molecular biological analysis from single cells on unlabeled neocortical brain slices. Using deep learning, trained on a previously built database of single neurons acquired in 3D, our system can detect most of the healthy neuronal somata in a Z-stack recorded by DIC microscopy from a living neocortical slice. The pipette approaches the target cell, touches it, acquires electrophysiological data, and the cell's nucleus can be isolated for further molecular analysis. Components of the system are a typical electrophysiological setup: IR video microscopy imaging system, motorized microelectrode manipulators, XY shifting table, electrical amplifier, and a custom-designed pressure controller. All these elements were controlled by a custom-developed software (available at https://bitbucket.org/biomag/autopatcher/). The system was successfully applied to perform patch clamp recordings on a large set of rodent and human cells (100 and 74, respectively). The automatically collected cells well represent the wide-range phenotypic heterogeneity of the brain cortex. Subsequent transcriptome profiling and whole-cell anatomical reconstruction confirmed the usefulness and applicability of the proposed system.

Hardware development and control. The hardware setup of the proposed system is shown in Fig. 2. The software system we developed controls each hardware using their drivers on application programming interface (API) level, which makes the system modular and different types of hardware components (e.g., manipulators, biological amplifier, and XZ shifting table) can be attached. The classes which control hardware elements are inherited from abstract classes. Thus, if the software is to be used with a different hardware element then only a few methods should be implemented in a child class that sends commands to that specific device (e.g., to get or set the pipette position or initiate a protocol in the amplifier's software).

The electrophysiological signal from the current monitor output of the amplifier is transferred to the DIGAP software via the analog input channel of the USB digitizer board (National Instruments, USB-6009), which enables real-time resistance measurement.



Pressure
Requested pressure: 50 mBar
Actual pressure: 50 mBar
Status: Regulating
Autopatcher status: -
Message: -
Electrophysiology
Resistance: 5.00 MOhm
Current: 0.0000 pA
□ calculate access resistance
Seconds to show: 10
RS Improver
Status -
Desired RS: 30

Fig. 2 Hardware setup of the DIGAP system. a Microscope with a motorized stage. **b** Micromanipulator. **c** Controller electronics for manipulators. **d** Patch clamp amplifier. **e** Pressure controller module. **f** Computer with the controller software.

To send commands to the amplifier, we used the "batch file control" protocol of HEKA PatchMaster 2×90.3 software (HEKA Elektronik, Germany). To apply different air pressure on the pipette in distinct phases of the patching procedure we built a custom pressure controller detailed in Supplementary Information: Pressure Regulator. Analog pressure sensors are used for monitoring the actual air pressure on the pipette and voltage signals of the sensors were connected in the input channels of the USB digitizer board. The solenoid valves of the regulator are controlled with TTL signals of the digital output channels of the digitizer.

Pipette calibration and automatic detection. Pipette calibration is a one-time process which determines the coordinate system transformation between the pipette and the stage axes. The calibration consists of moving the pipette along its axes with known distances, finding it with the stage and detecting the exact pipette tip position in the camera image. Calibration allows the pipette to be moved at any position of the microscope stage space. Note that no assumptions are made on the orientation or the tilt angles of the pipette.

The glass pipettes usually differ in length, thus the tip position should be updated after a pipette change. To automate this step we have developed algorithms for pipette detection in DIC images. First, we use a fast initialization heuristic and then refine the detection. The refinement step is the extension of our previous differential geometry-based method to three dimensions[21]. The pipette is modeled as two cylinders that have a common reference point and an orientation. The model is updated by the gradient descent method such that it covers dark regions introduced by the pipette in the image. Figure 3a shows the starting and final state of the algorithm from different projections in gradient images for visualization purposes. The detailed description of the algorithms and the equation derivations can be found in Supplementary Information: Pipette Detection System. The algorithm has an accuracy of 0.99 ± 0.55 μm compared to manually selected tip positions, that makes it possible to reliably reach cells of 10 μm diameter (on average) with the pipette when oriented towards their centroids.

Cell detection. We applied a deep learning algorithm in order to detect cells in DIC images and propose them for automatic patch clamp recording. Various software solutions were developed to detect[25,27] or segment[28,29] neurons (and cells in general) in cell cultures or tissues, however, they do not provide satisfactory results on images of contrast-enhancing techniques such as DIC or oblique. To obtain a reliable object detection in brain tissue, we designed a cell detection algorithm, which involved three steps: data annotation, training of the model, and inference.

For acquiring an appropriate set of labeled objects, we created and included a labeling tool into the software (see Supplementary Information: Software Usage) that offers a platform to generate an annotated dataset. Field experts labeled 6344 cells on 265 stacks (184 rat, 81 human). The annotation procedure consisted of putting bounding boxes around the recognized cells over multiple slices in the stack. The stacks consisted of 60–100 slices depending on the image quality in the actual sample. The dimension of the individual slices is 1392 × 1040 pixels (FoV 160.08×119.6 μm). The living cells were labeled on the slices such that a 2D bounding box was put in the 3D center of each object. We also copied the same boxes to the next two slices above and below. This resulted in a bounding box that has five-slices depth. The collected labeled data was converted into the required input format of the deep learning framework we used.

We have tested four different object detection deep learning architectures, including DetectNet[30,31], Faster Region-based Convolutional Neural Network (FRCNN)[32,33], Darknet-ResNeXt[34,35], and Darknet-YOLOv3-SPP[36]. A detailed description and performance comparison is given in (Supplementary Information: Cell Detection System). DetectNet and FRCNN have been implemented into DIGAP software. The former has lower performance but very high efficiency in inference speed, while the latter is the opposite. Users can choose based on requirements and available resources. For this work we used DetectNet.

DetectNet[30,31] architecture was trained using NVIDIA's Deep Learning GPU Training System (DIGITS[37]), which is an extension of Caffe[38], and allows even the non-advanced deep learning users to perform training. The solver used for the training process was adaptive moment estimation[39] (ADAM). The pre-trained weights of the ImageNet dataset were used for the initialization of GoogLeNet to speed up the training process. The number of epochs was 2500 which took 6 days and 15 h.

FRCNN with ResNet50 backbone was also pretrained on ImageNet. The Stochastic Gradient Descent with Momentum (SGDM)[40] was used as the optimizer with cross-entropy loss function. The number of epochs was 6. The initial learning rate was 1e−3, which was dropped every 2 epochs by a factor of 0.2. The training method was set to "end-to-end", that simultaneously trains the region proposal and region classification subnetworks. MATLAB R2019b was used for training, which took 2 days and 11 h. The prediction time of a single image using DetectNet was 0.1 s, while FRCNN required approx. an order of magnitude more time, 0.96 s per image.

By using these tools, the training processes generated models that recognize neurons in their original environment in DIC images (Fig. 3b). We also implemented a procedure that extends the 2D detection by uniting overlapping bounding boxes along the *Z*-axis in the image stacks to complete the object detection in 3D space (Fig. 3c). Bounding boxes of different Z slices are compared and if their intersection is at least 60% of the smaller box then they are united. The following detections are compared iteratively with the intersection region. To compensate for the detection errors when cells are not detected, bounding boxes that are three slices away from each other can still be united even if the two slices in between do not contain detections.

To evaluate the performance of the proposed frameworks we measured precision, recall, and *F*1 score on a validation dataset (Fig. 3d). This dataset consisted of three image stacks (305 images in total) annotated by the same annotator and was not used in the training process. The detected objects were matched with ground truth data automatically if their centroid were at most 5 μm in the lateral plane and 3 μm in the *Z* axis from each other. If a detection could not be matched, it was treated as a false positive (FP). Ground truth objects not paired with a detection were treated as false negatives (FN). Based on these aspects the detection accuracy was calculated as precision $P = TP/(TP + FP)$, recall $R = TP/(TP + FN)$, and *F*1 score $= 2 * P * R/(P + R)$. DetectNet achieved 56.88% *F*1-score (precision = 53.04%, recall = 61.33%). FRCNN architecture provided better results with a 65.83% *F*1-score (precision = 60.73%, recall = 71.88%). Furthermore, the authors of the DeNeRD model[27] showed that simpler neural networks can be used to achieve good accuracy in object detection tasks. Therefore, we have compared the ResNet50 backbone to MobileNetV2[41] combined with FRCNN (Supplementary Information: Cell Detection System). This showed that MobileNetV2 can be a good compromise if hardware limitations or inference speed is an issue.

To test the performance of the annotators we have determined intraexpert and interexpert accuracies. These were measured by

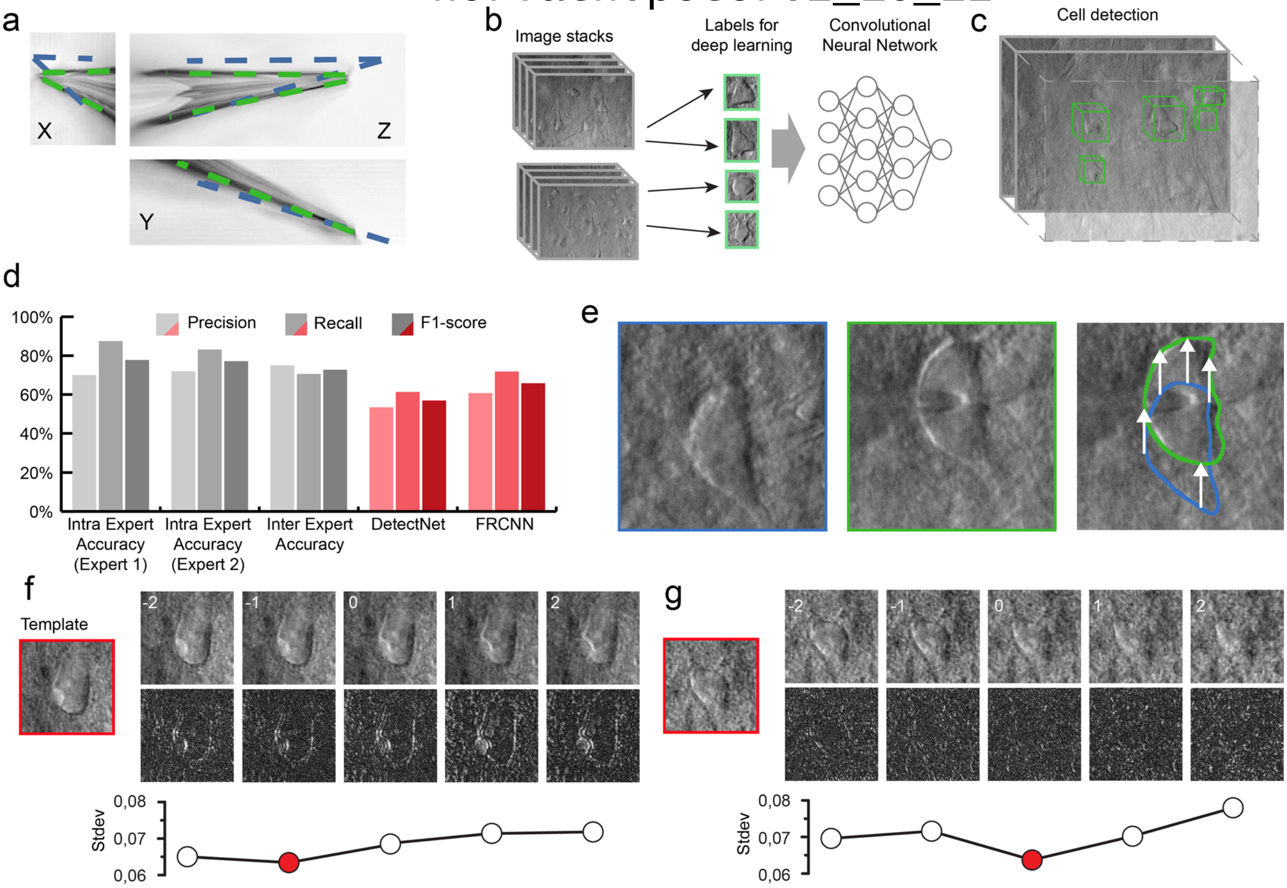


Fig. 3 The developed algorithms for the DIGAP system. a Result of the Pipette Hunter detection model shown in three different projections of the image stack. Initial state (blue contour) and the result (green contour) of our pipette localization algorithm are shown. **b** Training dataset generation: 265 image stacks (60–100 images per stack with 1 µm frame distance along the *Z*-axis) captured from human and rodent neocortical slices with DIC videomicroscopy (left). 31,720 objects as healthy cells (green boxes) labeled on every slice of the image stack by four experts. **c** After the training session, the DIGAP system detects cells in unstained living neocortical tissues. **d** Accuracy of the automated cell detection pipeline. **e** Lateral tracking of the cell movement ($n = 174$). DIC images of the targeted (in blue box) and patched cell (in green box). The cell drifted from its initial location (arrows in the right panel) during the pipette maneuver. **f**, **g** Z-tracking of the cell movement ($n = 174$). The template image was captured at the optimal focal depth (in red boxes) before starting the tracking. During the pipette movement, image stacks were captured from the targeted cell (upper panels) such that the middle slice was taken of the most recent focus position. The bottom row shows the differences between the template and the image of the corresponding *Z* position. The lowest standard deviation value of the difference images (plots) shows the direction of the cell drift in the *Z*-axis. Source Data is available as a Source Data file.

showing the same image stack (102 images) of the validation dataset to two annotators twice within 3 months time shift. The annotators reached 77.12% (precision = 71.91%, recall = 83.12%) and 77.78% *F*1-score (precision = 70%, recall = 87.5%), respectively. To compare the experts, the interexpert accuracy was measured which resulted in 72.73% *F*1-score (precision = 75%, recall = 70.59%) (Fig. 3d).

When the user initiates cell detection in the software, a stack is created and the detected cells are highlighted with bounding boxes (Fig. 3c). The detections are ordered by the confidence value, thus healthier cells are offered earlier. The target cell can also be selected manually based on arbitrary criteria required for the experiment.

Tracking the cell in 3D. Due to the elasticity of the tissue, the movement of the pipette can significantly deform it and change the location of the cell of interest. In order to precisely re-define the pipette trajectory, the location of the target cell needs to be tracked. We have developed an online system that performs tracking in the lateral and *Z* directions (Fig. 3e–g). Both directions require a template image of the target cell which is acquired before starting the patch clamp process when the cell is in the focal plane of the microscope. The lateral tracking is performed in the image of the most recent focal level. It uses the Kanade–Lucas–Tomasi (KLT) feature tracker algorithm[42,43]. The *Z* tracking is based on a focus detection algorithm that operates on a small image stack encompassing the target cell body. The standard deviation of the images of the target cell body is computed and compared to initial images. As a result, the displacement direction of the target cell along the *Z* axis is determined. The whole process was done with stopped pipette to ensure that the cell is not pushed away meanwhile. The detailed explanation of the algorithms with examples can be found in Supplementary Information: Cell Tracking System.

Automated patch clamping steps. After pipette calibration and cell detection the patch clamping procedure can be started. First, the DIGAP software calculates the trajectory of the pipette movement along which the manipulator moves the pipette tip (stepwise, 2 µm) close to the cell while applying medium air pressure (50–70 mbar). The initial trajectory is a straight line along the manipulator's *X* axis. Note that this is tilted (in our case

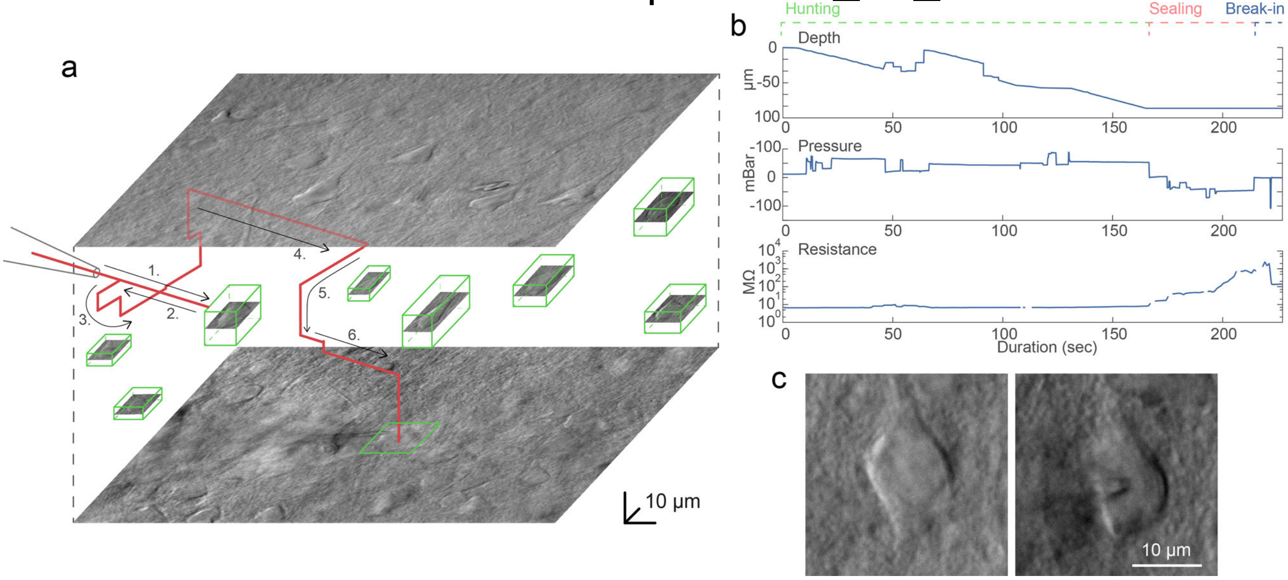


Fig. 4 A representative example of a visual patch clamping procedure. a Trajectory of the pipette tip (red line) with obstacle avoidance (numbered) in the tissue and the spatial location of the detected cells (green boxes). The steps of the avoidance algorithm are the following. 1: The pipette is moved forward in the initial trajectory until an obstacle is hit. 2: The pipette is pulled back. 3: The pipette is moved laterally in a spiral pattern until the resistance is back to normal. 4: The obstacle is passed. 5: The pipette is readjusted to the trajectory. 6: The approaching is continued. **b** Plots of the depth of the pipette tip in the tissue, the applied air pressure, and the measured pipette tip resistance during the approach. **c** Image of a cell before and after performing patch clamp recording on it. Source Data is available as a Source Data file.

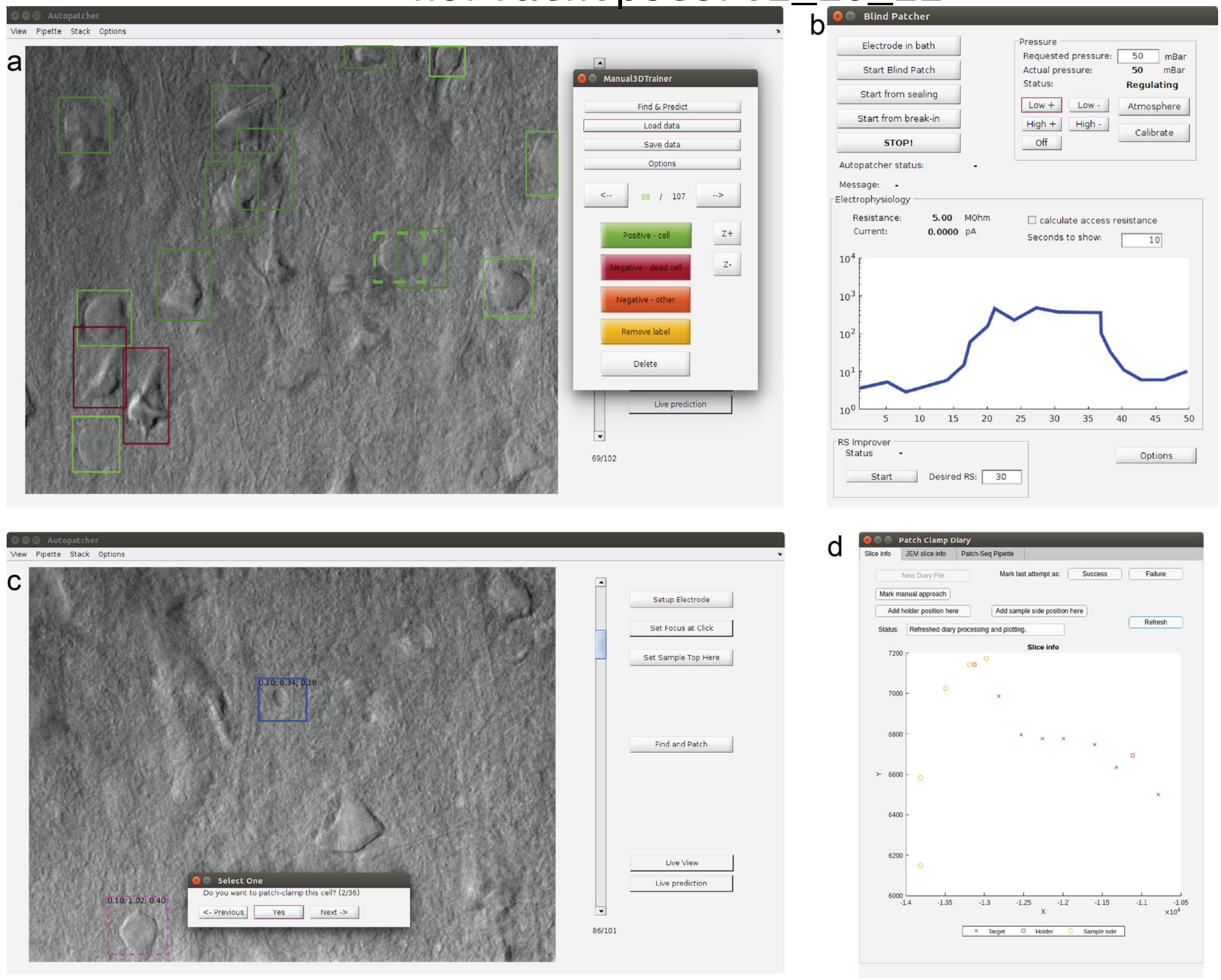
approximately −33 degrees from the horizontal plane) so the movement vector of the pipette is parallel to the longitudinal axis of the pipette. We found that approaching is more reliable if the pipette is first moved a few micrometers above the cell and then finally descending on it. The impedance of the pipette tip is monitored continuously during the movement.

During the movement of the pipette, air pressure is dynamically changed with predefined air pressure values. Air pressures were empirically set for the different phases: hunting, sealing, and breaking. Pipette tip impedance was continuously checked in order to detect phases and apply the task-specific pressure.

Early resistance increase denotes the presence of an obstacle in front of the pipette, e.g., a blood vessel or another cell. If an obstacle is hit, the pipette is pulled back, slightly moved laterally and when the obstacle is passed the pipette is oriented back to the initial trajectory towards the target[15]. Meanwhile, the described 3D tracking algorithm compensates for the movement trajectory due to the possible displacement of the target cell. When the pipette tip reaches the target position above the cell, the pressure is decreased to a low positive value (10–30 mbar). Then the pipette is moved in the Z direction and the resistance of the tip is monitored by 5 ms long −5 mV voltage steps. If the impedance increases more than a predefined value (0.7–1.2 MΩ) the sealing phase is initiated. The cell-attached configuration is set up by the immediate cease of pressure. To achieve tight sealing of the cell membrane into the glass we apply small negative pressure (from −30 to −10 mbar) and the holding potential is set to −60 mV stepwise. If the sealing process is slow and does not reach 1 GΩ ("gigaseal") in 30 s, different protocols are applied. First, the initial vacuum is amplified by 1.5 and 2 times, each for 20 more sec. Then the pipette is moved +/−2 μm in each axis for 2 s. Finally, the pressure is released for 10 s and reapplied for 20 s. If the gigaseal state is reached then suction pulses (−140 to −100 mbar) of increasing length (0.5 + 0.2*attempt sec) are applied for up to 3 min to break-in the membrane. Information about the process, including pipette distance from the target, actual air pressure, and electrical resistance values are continuously monitored and shown in the GUI windows. Description of the steps and the parameter values are described in detail in Supplementary Information: Software Usage. A representative procedure is demonstrated in Fig. 4, and further trajectory, pressure, and resistance data is visualized in Supplementary Information: Representative examples.

Software. The control software is written in MATLAB and the source code is made publicly available at https://bitbucket.org/biomag/autopatcher/. The visual patch clamping process can be started from a user-friendly GUI (Fig. 5) which allows every parameter to be set and the process to be monitored in real-time by the operator. Throughout the session, the Patch Clamp Diary module collects and visualizes information about patch clamping attempts, including their location and outcome status. The user can additionally mark positions in the biological sample that help orientation during the experiment (i.e., boundaries of the brain slice or the parallel strands that keep the tissue secure).

Many utility features are present to help everyday experimenting. Single images or image stacks can be acquired, saved, or loaded from the menu bar. The acquired images can be processed by performing background illumination correction or DIC image reconstruction, which can help in identifying cells and their features. The graphical processing unit (GPU) extension of our reconstruction algorithm[44] can be used for reconstruction, which results in about 1000× speed increase. The software contains a built-in labeling tool that allows image database generation to train deep learning cell recognition. Furthermore, most recent practices from other automation systems have also been implemented for the in vivo usage, including pipette cleaning[16,17] or hit reproducibility check[45]. The XML configuration file makes the adaptation easy between different setups and the software can also operate as a general microscope controller. A logging system is used for maintainability purposes.



Requested pressure: 50 mBar
Actual pressure: 50 mBar
Status: Regulating

Autopatcher status: -

Message: -

Resistance: 5.00 MOhm
Current: 0.0000 pA

□ calculate access resistance

Seconds to show: 10

Status -

Desired RS: 30

Do you want to patch-clamp this cell? (2/36)

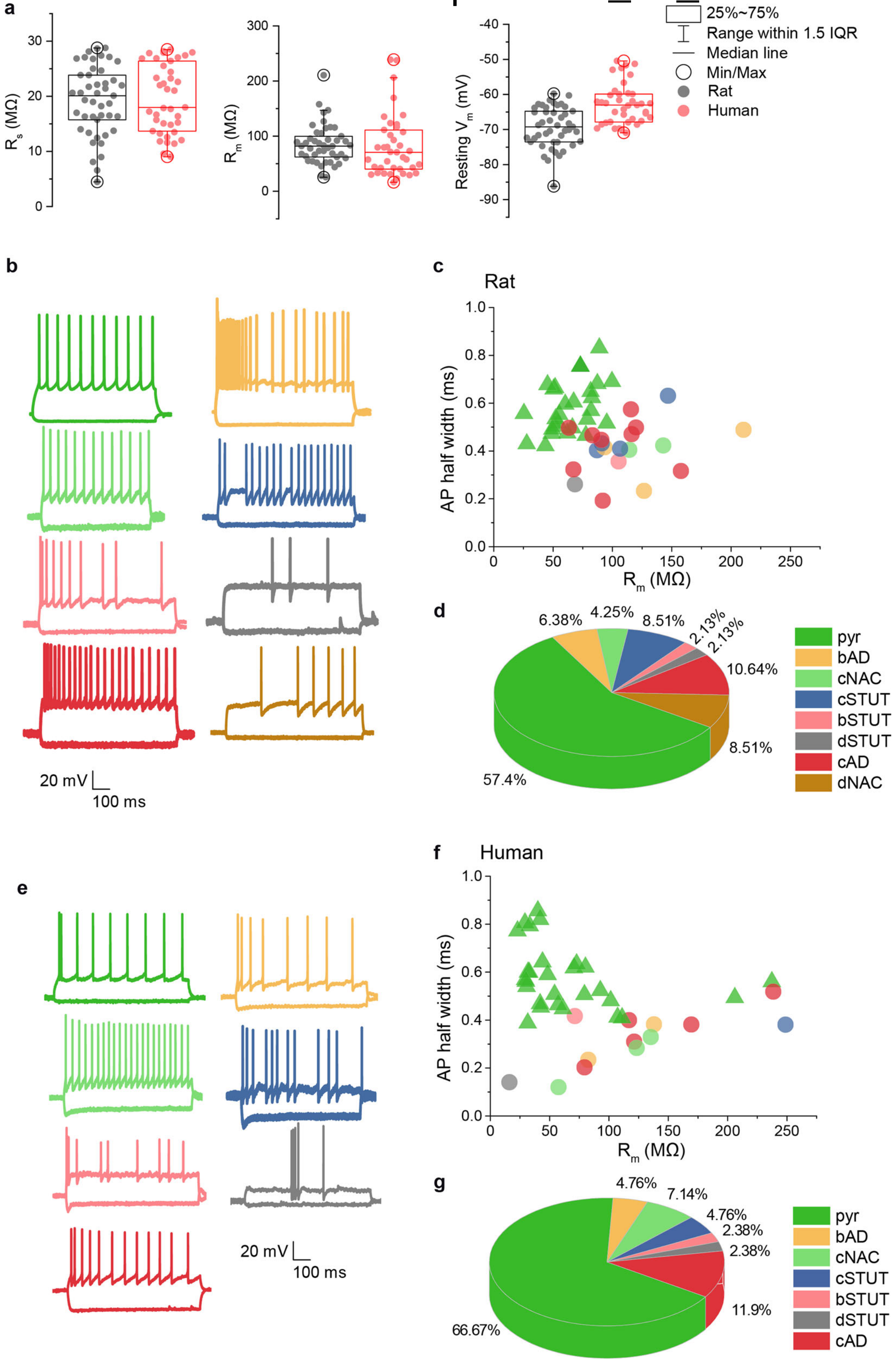
Mark last attempt as:

Status: Refreshed diary processing and plotting.

Fig. 5 GUI of the software. a Main window with an image stack loaded and the built-in labeling tool started. **b** Monitoring window to check the pressure and resistance values. Pressure values can be set here when operating manually, or the measurement can be restarted from different subphases here. **c** Main window when browsing the detected cells, initiated with the Find and Patch button. The measurement can be started by selecting a cell. **d** The Patch Clamp Diary module showing a plot with annotations of a sample and measurements in it.

Application in brain slices. To test the performance and effectiveness of our system we obtained a series of recordings (Supplementary Information: Electrophysiology stimuli for DIGAP) on slice preparation of rat somatosensory and visual cortices ($n = 23$ animals) and human temporal and association cortices ($n = 16$ patients). Successful automatic whole-cell patch clamp trials without experimenter assistance were achieved in a total number of $n = 100$ and $n = 74$ (rodent visual and somatosensory cortices and human cortex, respectively) out of $n = 157$ and $n = 198$ attempts. The data analysis was carried out using Fitmaster 2×73 (HEKA Elektronik, Germany), OriginPro 7.5 (OriginLab, USA), Excel 2016 (Microsoft, USA), and MATLAB R2017a (Mathworks, USA). The quality of recordings was supervised by measuring series resistance (R_s) (Fig. 6). We found a wide range of R_s values within successful attempts in both species: 34.52 ± 18.99 MΩ in rat and 31.39 ± 16.67 MΩ in human recordings. Trials with R_s value exceeding 100 MΩ were noted as unsuccessful attempts. Access resistance in 48.28% of our recordings was under 30 MΩ which we denoted as high quality and used for further analysis. Once the whole cell configuration was formed cells were usually held at most for 15 min to protect neuron viability for further procedures. To test the stability of whole cell configurations, we executed a separate set of experiments and found that half of the trials ($n = 5$ out of 9) could be kept up to 1 h. The average time of experiments during the recording configuration could be maintained was 2729.9 ± 1104.2 s ($n = 9$, min: 928 s, max: 3825 s). During our measurements we were able to detect spontaneous postsynaptic events in the entire length of the recordings. We applied standard stimulation protocol and recorded membrane potential responses to injected currents. Based on the extracted common physiological features and firing patterns we grouped neurons into electrophysiological types (e-types[46]) based on criteria established by the Petilla convention[47]. There were eight e-types in automatic patched rat samples: pyramidal cell (pyr), burst adapting (bAD), continuous non-accommodating (cNAC), continuous stuttering (cSTUT), burst stuttering (bSTUT), delayed stuttering (dSTUT), continuous adapting (cAD), and delayed non-accommodating (dNAC). From the human samples, seven e-types were identified. In our automatically-collected dataset, dNAC type was not represented (Fig. 6).

Electrophysiological recordings were acquired using a biocytin-containing intracellular solution. We performed further anatomical investigation on $n = 44$ experiments with <30 MΩ access resistance and we achieved $n = 18$ ($n = 16$ and $n = 2$ from human and rat, respectively) full and $n = 11$ ($n = 3$ and $n = 8$ from human and rat, respectively) partial recovery



(Fig. 7a, Supplementary Information: Anatomical reconstruction examples).

We next tested if single-cell RNA analysis is achievable from the collected cytoplasm of autopatched neurons. After whole-cell recording of the neurons in the brain slices the intracellular content of the patched cells was aspirated into the recording pipette with gentle suction applied by the pressure regulator unit (−40 mBar for 1 min, then −60 mBar for 2–3 min, and finally −40 mBar for 1 min). The tight seal was maintained and the pipette was carefully withdrawn from the cell to form an outside-out configuration. Subsequently, the content of the pipette was expelled into a low-adsorption test tube (Axygen) containing 0.5 µl SingleCellProtectTM (Avidin Ltd. Szeged, Hungary) solution in order to prevent nucleic acid degradation and to be

Fig. 6 Electrophysiological properties of the cells patched by DIGAP. a Main electrophysiological parameters from the successful automatic patch clamp recordings. The box plots show the series resistance (R_s, left panel), the membrane resistance (R_m, middle panel), and the resting membrane potential (right panel) of all successful measurements ($n = 47$ for rat and $n = 41$ for human samples). The boxes show the median, 25 and 75 percentiles, and min/max values, and the whiskers are 1.5 interquartile ranges. **b** Different cell types are identified according to firing features: pyr pyramidal cell, bAD burst adapting, cNAC continuous non-accommodating, cSTUT continuous stuttering, bSTUT burst stuttering, dSTUT delayed stuttering, cAD continuous adapting, dNAC delayed non-accomodating. **c** Individual neurons' action potential half-widths are presented as a function of the same neuron's R_m. Note the segregation of excitatory and inhibitory neuronal classes. Dataset is recorded from rodent samples (Panel **c** and **d** colors correspond to panel **b**). **d** The proportion of recorded cell types. **e**–**g** Same plots as **b**–**d**, representing the dataset recorded in human neocortical slices. Source Data is available as a Source Data file.

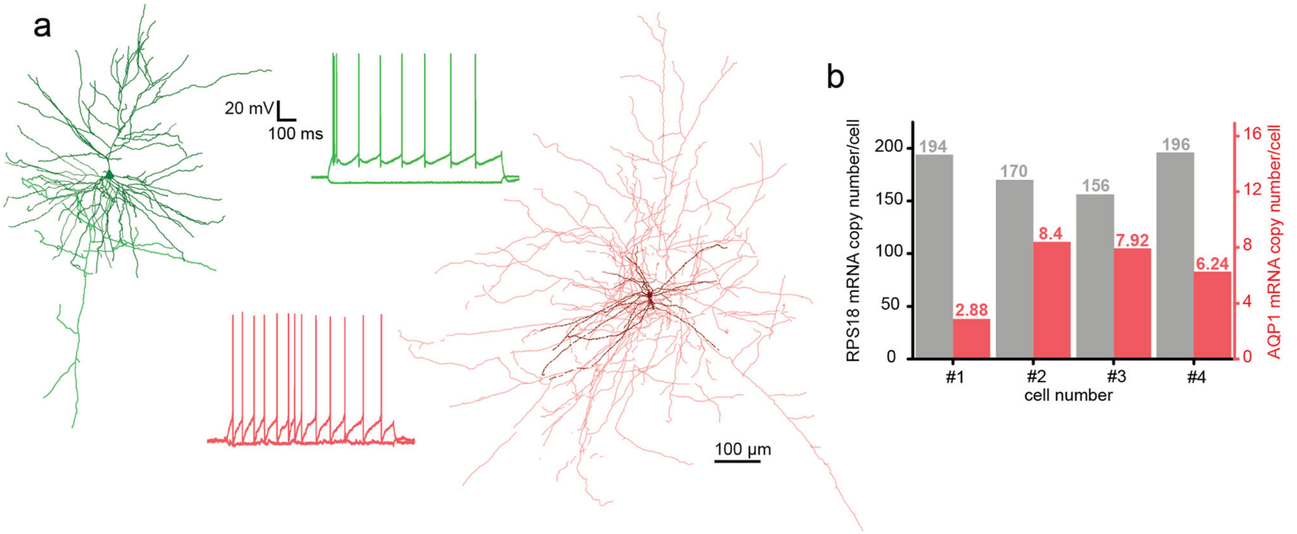


Fig. 7 Anatomical and molecular biological investigation of neurons patched by DIGAP. a Two anatomically reconstructed human autopatched neurons. The darker colors represent somata and dendrites of the pyramidal (green) and the interneuron (red) cells. The brighter color shows the axonal arborization. The firing patterns of the cells are the same color as their reconstructions. **b** mRNA copy numbers of a housekeeping (RPS18, black bars) and the aquaporin 1 (AQP1, red bars) gene from four representative human pyramidal cells. Source Data is available as a Source Data file.

compatible with direct reverse transcription reaction. Then the samples were used for digital polymerase chain reaction (dPCR) analysis to determine the copy number of selected genes. From four single pyramidal cell cytoplasm samples which were extracted from the human temporal cortex, we determined the copy number of a ribosomal housekeeping RPS18 and aquaporin 1 (AQP1) genes (Fig. 7b). The results of the dPCR experiments are in agreement with our previous observations[48,49].

Discussion

The developed DIGAP system is able to fully automatically perform whole-cell patch clamp recordings on single neurons in rodent and human neocortical slices (Supplementary Movie 1, 2, 3). This is a step forward towards characterizing and understanding the phenotypic heterogeneity and cellular diversity of the brain. The presented system has a cell detection module in label-free imaging, which is achieved by deep learning. The system we developed is fully controlled by a single software, including all hardware components, data handling, and visualization. The control software has its highly comprehensive internal logging system, that allows tracking the parameters of each patch clamp recording attempt in addition with the option to store details of the cytoplasm harvesting process. In addition, it can connect to and save database entry records that are compatible with the Allen Brain Atlas single neuron database. In this work, we demonstrated the power of our system that is capable of measuring a large set of rodent and human neurons in the brain cortex. The results show strong correlation to the earlier results in literature in terms of quality and phenotypic composition of cell heterogeneity. Records of measured cells were inserted to the database of the Allen Institute for Brain Science and a subset of the cells was isolated from their tissue environment and single-cell mRNA copy numbers of two selected genes were determined. Furthermore, we successfully demonstrated that autopatched neurons can be anatomically reconstructed.

The main advantage of the proposed system is that it can easily be integrated into any existing setups and although we do not believe that it will fully substitute human experts, it is a great choice for complex specific tasks, allows parallelization and speeds up discovery. It is important to emphasize the need for a standardized and fully documented patch clamping procedure, which is guaranteed by using DIGAP. The choice of advanced image analysis and deep learning techniques made it possible to work with the least harmful imaging modalities at a human expert level of single-cell detection that was impossible so far. Further possibilities are more widespread and potentially enabling or accelerating discoveries. Combining with intelligent single-cell selection strategies of the detected cells, the proposed system can be the ultimate tool to reveal and describe cellular heterogeneity. In multiple patch clamp setup it can be used to describe the connectome at cellular level. We presented DIGAP's application to brain research, but other fields, such as cardiovascular or organoid research will benefit from the system. Based on its nearly complete automation, it can help in education.

Future work includes adding multipipette support to study connections between pairs, triplets, or a higher number of cells at a time. Furthermore, the cell detection can be improved by increasing the size of the training dataset, the diversity of images (by collecting them from various setups), and improving the annotation process, or even extending it to 3D instance segmentation instead of object detection.

Methods

Hardware setup. A customized Olympus BX61 (Olympus, Japan) microscope with a 40× water immersion objective (0.8 NA; FoV 0.6625 mm; Olympus, Japan) with motorized *Z* axis (Femtonics, Hungary) which is controlled by API calls to the software was used for imaging. For moving the pipette and the microscope stage we used Luigs & Neumann Mini manipulators with SM-5 controllers (Luigs & Neumann, Germany). The electrophysiological signals were measured by a HEKA EPC-10 amplifier (HEKA Elektronik, Germany). The signals were digitized at 100 kHz and Bessel filtered at 10 kHz.

In vitro preparation of human and rat brain slices. All procedures were performed according to the Declaration of Helsinki with the approval of the University of Szeged Ethics Committee. Human slices were derived from materials that had to be removed to gain access for the surgical treatment of deep-brain tumors, epilepsy, or hydrocephalus from the association cortical areas with written informed consent of female ($n = 9$, aged 48.2 ± 26.6 years) and male ($n = 7$, aged 48.3 ± 9.9 years) patients prior to surgery. Anesthesia was induced with intravenous midazolam and fentanyl (0.03 mg/kg, 1–2 μg/kg, respectively). A bolus dose of propofol (1–2 mg/kg) was administered intravenously. To facilitate endotracheal intubation, the patient received 0.5 mg/kg rocuronium. After 120 s, the trachea was intubated and the patient was ventilated with a mixture of O_2 and N_2O at a ratio of 1:2. Anesthesia was maintained with sevoflurane at monitored anesthesia care (MAC) volume of 1.2–1.5. After surgical removing blocks of tissue were immediately immersed in ice-cold solution containing (in mM) 130 NaCl, 3.5 KCl, 1 NaH_2PO_4, 24 $NaHCO_3$, 1 $CaCl_2$, 3 $MgSO_4$, 10 d(+)-glucose, saturated with 95% O_2 and 5% CO_2. Slices were cut perpendicular to cortical layers at a thickness of 350 μm with a vibrating blade microtome (Microm HM 650 V, Thermo Fisher Scientific, Germany) and were incubated at room temperature for 1 h in the same solution. The artificial cerebrospinal fluid (aCSF) used during recordings was similar to the slicing solution, but it contained 3 mM CaCl and 1.5 mM $MgSO_4$.

Coronal slices (350 μm) were prepared from the somatosensory cortex of male Wistar rats (P18-25, $n = 23$, RRID: RGD_2312511)[50]. All procedures were performed with the approval of the University of Szeged and in accordance with the Guide for the Care and Use of Laboratory Animals (2011). Recordings were performed at 36 °C temperature. Micropipettes (3.5–5 MΩ) were filled with low [Cl] intracellular solution for whole-cell patch clamp recording: (in mM) 126 K-gluconate, 4 KCl, 4 ATP-Mg, 0.3 GTP-Na_2, 10 HEPES, 10 phosphocreatine, and 8 biocytin (pH 7.20; 300 mOsm).

Molecular biological analysis. After harvesting the cytoplasm of the recorded cells the samples were frozen in dry ice and stored at −80 °C until used for reverse transcription. The reverse transcription (RT) of the harvested cytoplasm was carried out in two steps. The first step took 5 min at 65 °C in a total reaction volume of 5 μl containing 2 μl intracellular solution and SingleCellProtectTM mix with the cytoplasmic contents of the neuron, 0.3 μl TaqMan Assays, 0.3 μl 10 mM dNTPs, 1 μl 5× first-strand buffer, 0.3 μl 0.1 mol/l DTT, 0.3 μl RNase inhibitor (Life Technologies, Thermo Fisher Scientific, Germany) and 100 U of reverse transcriptase (Superscript III, Invitrogen, Thermo Fisher Scientific, Germany). The second step of the reaction was carried out at 55 °C for 1 h and then the reaction was stopped by heating at 75 °C for 15 min. The reverse transcription reaction mix was stored at −20 °C until PCR amplification. For digital PCR analysis the reverse transcription reaction mixture (5 μl), 2 μl TaqMan Assays (Life Technologies, Thermo Fisher Scientific, Germany), 10 μl OpenArray Digital PCR Master Mix (Life Technologies, Thermo Fisher Scientific, Germany) and nuclease-free water (5.5 μl) were mixed in a total volume of 20 μl. The mixture was evenly distributed on an OpenArray plate. RT mixes were loaded into four wells of a 384-well plate from which the OpenArray autoloader transferred the cDNA master mix by capillary action into 256 nanocapillary holes (four subarrays) on an OpenArray plate. Processing of the OpenArray slide, cycling in the OpenArray NT cycler and data analysis was done as previously described[48]. For our dPCR protocol amplification, reactions with CT confidence values below 100 as well as reactions having CT values less than 23 or greater than 33 were considered primer dimers or background signals, respectively, and were excluded from the data set.

Anatomical processing and reconstruction of recorded cells. Following electrophysiological recordings, slices were transferred into a fixative solution containing 4% paraformaldehyde, 15% (v/v) saturated picric acid, and 1.25% glutaraldehyde in 0.1 M phosphate buffer (PB; pH = 7.4) at 4 °C for at least 12 h. After several washes with 0.1 M PB, slices were frozen in liquid nitrogen then thawed in 0.1 M PB, embedded in 10% gelatin, and further sectioned into 60-μm slices. Sections were incubated in a solution of conjugated avidin-biotin horseradish peroxidase (ABC; 1:100; Vector Labs) in Tris-buffered saline (TBS, pH = 7.4) at 4 °C overnight. The enzyme reaction was revealed by 3′ 3-diaminobenzidine tetrahydrochloride (0.05%) as chromogen and 0.01% H_2O_2 as oxidant. Sections were postfixed with 1% OsO_4 in 0.1 M PB. After several washes in distilled water, sections were stained in 1% uranyl acetate and dehydrated in an ascending series of ethanol. Sections were infiltrated with epoxy resin (Durcupan) overnight and embedded on glass slides. Three-dimensional light-microscopic reconstructions were carried out using a Neurolucida system (MicroBrightField, USA) with a 100× objective.

Pipette cleaner. We implemented a pipette cleaning method[16] into our system. The cleaning procedure requires two cleaning agents: Alconox, a commercially available cleaning detergent, and artificial cerebrospinal fluid (aCSF). We 3D printed a holder for two PCR tubes containing the liquids that can be attached to the microscope objective and are reachable by the pipette tip. The cleaning is performed by pneumatically taking up and then removing the agents into and from the pipette. The vacuum strength used for the intake of the liquids is −300 mBar and the pressure used for the expulsion is +1000 mBar. The method consists of three steps. First, the pipette is moved to the cleaning agent bath and vacuum is applied for 4 s. Then, to physically agitate glass-adhered tissue, pressure and vacuum are alternated, each for 1 s and repeated for five times total. Finally, pressure is applied for 10 s to make sure all detergent is removed. In the second step, the pipette is moved to the aCSF bath and any remaining detergent is expelled by applying pressure for 10 s. In the third step, the pipette is moved back to the position near to the biological sample where the cleaning process was initiated. In the original paper, it is shown that these pressure values and the duration of the different steps are more than enough to cycle the volume of agents necessary to clean the pipette tip. We provide a graphical window in our software to calibrate the pipette positions of the tubes containing the cleaning agent and the aCSF and to start the cleaning process.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available in the manuscript, Source Data file, supplementary information and available from the authors upon reasonable request. The annotated image data used for deep learning are available from the corresponding author upon request. Source data are provided with this paper.

Code availability

Source code is available from Bitbucket at https://bitbucket.org/biomag/autopatcher/.

Received: 30 March 2020; Accepted: 18 January 2021;
Published online: 10 February 2021

References

1. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
2. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
3. Zeng, H. et al. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**, 483–496 (2012).
4. Gouwens, N. W. et al. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nat. Neurosci.* **22**, 1182–1195 (2019).
5. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
6. Suk, H.-J., Boyden, E. S. & van Welie, I. Advances in the automation of whole-cell patch clamp technology. *J. Neurosci. Methods* **326**, 108357 (2019).
7. Peng, Y. et al. High-throughput microcircuit analysis of individual human brains through next-generation multineuron patch-clamp. *Elife* https://doi.org/10.1101/639328 (2019).
8. Kodandaramaiah, S. B. et al. Assembly and operation of the autopatcher for automated intracellular neural recording in vivo. *Nat. Protoc.* **11**, 634–654 (2016).
9. Kodandaramaiah, S. B., Franzesi, G. T., Chow, B. Y., Boyden, E. S. & Forest, C. R. Automated whole-cell patch-clamp electrophysiology of neurons in vivo. *Nat. Methods* **9**, 585–587 (2012).
10. Kodandaramaiah, S. B. *Robotics for In Vivo Whole Cell Patch Clamping* (Georgia Institute of Technology, 2012).
11. Suk, H.-J. et al. Closed-loop real-time imaging enables fully automated cell-targeted patch-clamp neural recording in vivo. *Neuron* **96**, 244–245 (2017).
12. Long, B., Li, L., Knoblich, U., Zeng, H. & Peng, H. 3D image-guided automatic pipette positioning for single cell experiments in vivo. *Sci. Rep.* **5**, 18426 (2015).
13. Annecchino, L. A. et al. Robotic automation of in vivo two-photon targeted whole-cell patch-clamp electrophysiology. *Neuron* **95**, 1048–1055 (2017).

14. Desai, N. S., Siegel, J. J., Taylor, W., Chitwood, R. A. & Johnston, D. MATLAB-based automated patch-clamp system for awake behaving mice. *J. Neurophysiol.* **114**, 1331–1345 (2015).
15. Stoy, W. A. et al. Robotic navigation to subcortical neural tissue for intracellular electrophysiology in vivo. *J. Neurophysiol.* **118**, 1141–1150 (2017).
16. Kolb, I. et al. Cleaning patch-clamp pipettes for immediate reuse. *Sci. Rep.* **6**, 35001 (2016).
17. Kolb, I. et al. PatcherBot: a single-cell electrophysiology robot for adherent cells and brain slices. *J. Neural Eng.* **16**, 046003 (2019).
18. Perin, R. & Markram, H. A computer-assisted multi-electrode patch-clamp system. *J. Vis. Exp.* **80**, e50630 (2013).
19. Kodandaramaiah, S. B. et al. Multi-neuron intracellular recording in vivo via interacting autopatching robots. *Elife* **7**, e24656 (2018).
20. Li, L. et al. A robot for high yield electrophysiology and morphology of single neurons in vivo. *Nat. Commun.* **8**, 15604 (2017).
21. Koos, K., Molnár, J. & Horvath, P. Pipette Hunter: patch-clamp pipette detection. *Image Anal.* https://doi.org/10.1007/978-3-319-59126-1_15 (2017).
22. Yang, R. et al. Cell segmentation and pipette identification for automated patch clamp recording. *Robot. Biomim.* **1**, 1–12 (2014).
23. Wu, Q. et al. Integration of autopatching with automated pipette and cell detection in vitro. *J. Neurophysiol.* **116**, 1564–1578 (2016).
24. Moen, E. et al. Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233–1246 (2019).
25. Suleymanova, I. et al. A deep convolutional neural network approach for astrocyte detection. *Sci. Rep.* **8**, 1–7 (2018).
26. Allen Institute for Brain Science. Allen Cell Types Database. Allen Brain Atlas http://help.brain-map.org/display/celltypes.
27. Iqbal, A., Sheikh, A. & Karayannis, T. DeNeRD: high-throughput detection of neurons for brain-wide analysis with deep learning. *Sci. Rep.* **9**, 13828 (2019).
28. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
29. Sommer, C., Straehle, C., Kothe, U. & Hamprecht, F. A. Ilastik: interactive learning and segmentation toolkit. *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* https://doi.org/10.1109/isbi.2011.5872394 (2011).
30. Tao, A., Barker, J. & Sarathy, S. DetectNet: deep neural network for object detection in DIGITS. *NVIDIA Developer Blog* https://developer.nvidia.com/blog/detectnet-deep-neural-network-object-detection-digits/ (2016).
31. Szegedy, C. et al. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2015.7298594 (2015).
32. Ren, S., He, K., Girshick, R., Sun, J. & Faster, R.-C. N. N. Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017).
33. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2016.90 (2016).
34. Xie, S., Girshick, R., Dollar, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2017.634 (2017).
35. Redmon, J. Darknet: open source neural networks in C. http://pjreddie.com/darknet/ (2013–2016).
36. Redmon, J. & Farhadi, A. YOLOv3: an incremental improvement. https://arxiv.org/1804.02767 (2018).
37. Yeager, L., Bernauer, J., Gray, A. & Houston, M. Digits: the deep learning gpu training system. in *ICML 2015 AutoML Workshop* (2015).
38. Jia, Y. et al. Caffe. *Proceedings of the ACM International Conference on Multimedia—MM '14 (2014)* https://doi.org/10.1145/2647868.2654889. (2014).
39. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. https://doi.arxiv.org/1412.6980 (2014).
40. Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. (MIT Press, 2012).
41. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* https://doi.org/10.1109/cvpr.2018.00474 (2018).
42. Tomasi, C. & Kanade, T. Detection and tracking of point features. *Int. J. Comput. Vis.* 137–154 (1991).
43. Shi, J. & Tomasi. Good features to track. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94* https://doi.org/10.1109/cvpr.1994.323794 (1994).
44. Koos, K., Molnár, J., Kelemen, L., Tamás, G. & Horvath, P. DIC image reconstruction using an energy minimization framework to visualize optical path length distribution. *Sci. Rep.* **6**, 30420 (2016).
45. Yang, R., Lai, K. W. C., Xi, N. & Yang, J. Development of automated patch clamp system for electrophysiology. *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)* https://doi.org/10.1109/robio.2013.6739793 (2013).
46. Markram, H. et al. Reconstruction and simulation of neocortical microcircuitry. *Cell* **163**, 456–492 (2015).
47. Petilla Interneuron Nomenclature Group et al. Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nat. Rev. Neurosci.* **9**, 557–568 (2008).
48. Faragó, N. et al. Digital PCR to determine the number of transcripts from single neurons after patch-clamp recording. *Biotechniques* **54**, 327–336 (2013).
49. Faragó, N. et al. Human neuronal changes in brain edema and increased intracranial pressure. *Acta Neuropathol. Commun.* **4**, 78 (2016).
50. Molnár, G. et al. GABAergic neurogliaform cells represent local sources of insulin in the cerebral cortex. *J. Neurosci.* **34**, 1133–1137 (2014).

Acknowledgements

We thank Tímea Tóth and Réka Hollandi for their help in the image labeling, Ádám Szűcs for his work in the early stages of the development, Tamás Szépe for the advice on manipulator control, Nelli Tóth for the anatomical reconstruction, and István Grexa for the 3D printing. This work was supported by NAP-B brain research grant; the NVidia GPU Grant program; the LENDULET-BIOMAG Grant (2018-342); the European Regional Development Funds (GINOP-2.3.2-15-2016-00006, GINOP-2.3.2-15-2016-00026, GINOP-2.3.2-15-2016-00037); the Loránd Eötvös Research Network; the National Research, Development and Innovation Office of Hungary (GINOP-2.3.2-15-2016-00018, KKP_20 Élvonal KKP133807) (G.T.); Ministry of Human Capacities Hungary (20391-3/2018/FEKUSTRAT) (G.T.); from the National Research, Development and Innovation Office (OTKA K128863) (G.T., G.M.); ÚNKP-20-5 - SZTE-681 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund (G.M.); and János Bolyai Research Scholarship of the Hungarian Academy of Sciences (G.M.).

Author contributions

K.K. developed the software. G.O., K.K., M.R., and A.O. built and assembled the hardware. K.K. performed imaging. T.B. and K.K. developed the cell detection system. A.O., G.M., G.O., K.K., N.M., and M.R. performed electrophysiological measurements and analyzed the data. E.T. and J.M. developed the reconstruction models. P.B. provided human samples. G.M., G.T., and P.H. supervised the project. K.K, G.O., T.B., N.M., M.R., A.O., J.M., G.M., G.T., and P.H. contributed to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21291-4.

Correspondence and requests for materials should be addressed to P.H.

Peer review information *Nature Communications* thanks Theofanis Karayannis, Simon Schultz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

nature biotechnology

horvath.peter.2_10_22

ARTICLES

https://doi.org/10.1038/s41587-022-01302-5

Deep Visual Proteomics defines single-cell identity and heterogeneity

Andreas Mund[1,19] ✉, Fabian Coscia[1,2,19], András Kriston[3,4], Réka Hollandi[3], Ferenc Kovács[3,4], Andreas-David Brunner[5], Ede Migh[3], Lisa Schweizer[5], Alberto Santos[1,6,7], Michael Bzorek[8], Soraya Naimy[8], Lise Mette Rahbek-Gjerdrum[8,9], Beatrice Dyring-Andersen[1,10,11], Jutta Bulkescher[12], Claudia Lukas[12,13], Mark Adam Eckert[14], Ernst Lengyel[14], Christian Gnann[15], Emma Lundberg[15,16,17], Peter Horvath[3,4,18] ✉ and Matthias Mann[1,5] ✉

Despite the availabilty 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[1]. 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[2] (Fig. 1). Key challenges turned out to be the accurate definition of 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' (Biology 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

[1] Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [2] Spatial Proteomics Group, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. [3] Synthetic and Systems Biology Unit, Biological Research Centre, Eötvös Loránd Research Network, Szeged, Hungary. [4] Single-Cell Technologies Ltd., Szeged, Hungary. [5] Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [6] Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark. [7] Big Data Institute, Li-Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [8] Department of Pathology, Zealand University Hospital, Roskilde, Denmark. [9] Institute for Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [10] Department of Dermatology and Allergy, Herlev and Gentofte Hospital, University of Copenhagen, Hellerup, Denmark. [11] Leo Foundation Skin Immunology Research Center, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [12] Protein Imaging Platform, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [13] Protein Signaling Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [14] Department of Obstetrics and Gynecology/Section of Gynecologic Oncology, University of Chicago, Chicago, IL, USA. [15] Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden. [16] Department of Bioengineering, Stanford University, Stanford, CA, USA. [17] Chan Zuckerberg Biohub, San Francisco, CA, USA. [18] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [19] These authors contributed equally: Andreas Mund, Fabian Coscia. ✉e-mail: andreas.mund@cpr.ku.dk; horvath.peter@brc.hu; mmann@biochem.mpg.de

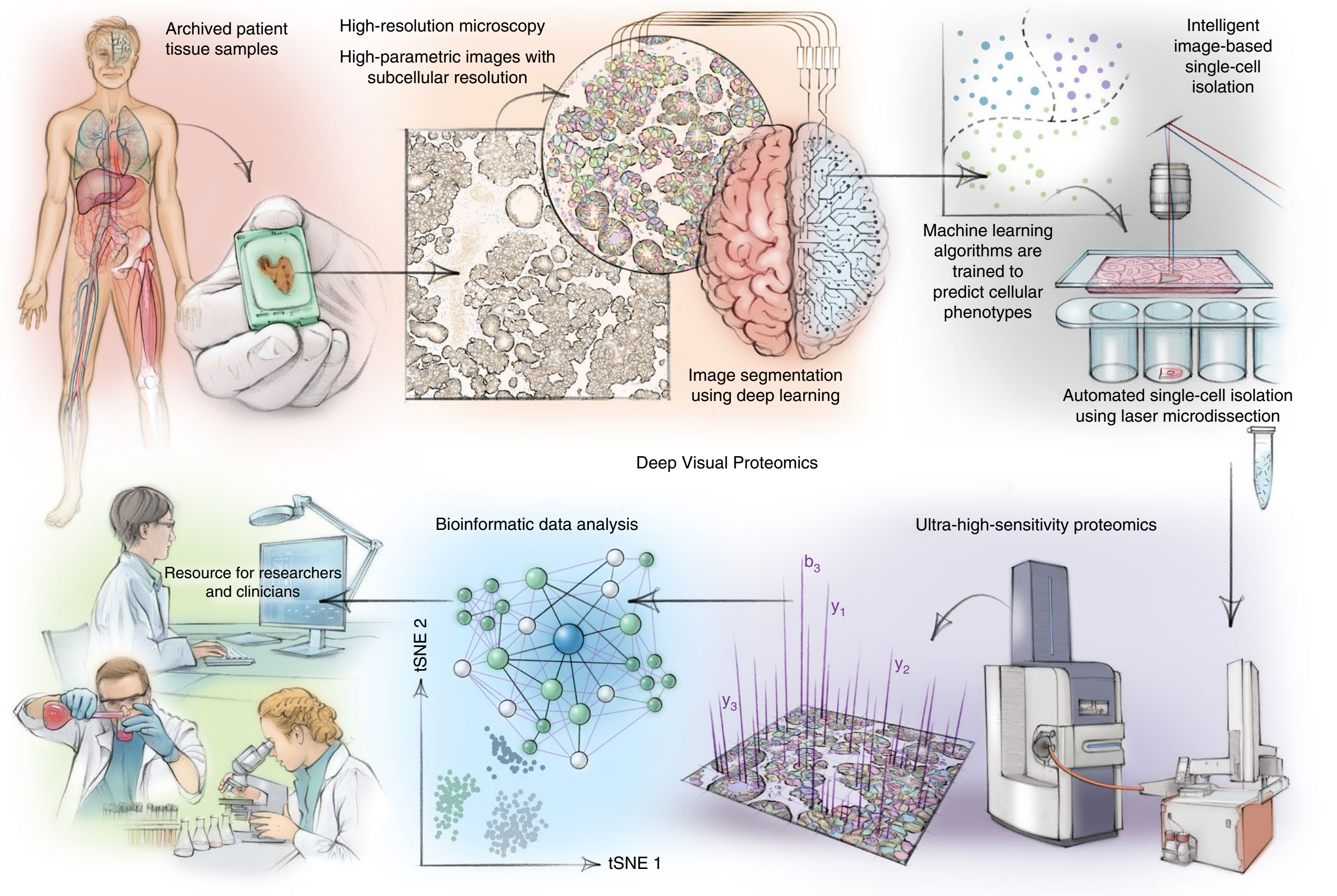

Archived patient tissue samples
High-resolution microscopy
High-parametric images with subcellular resolution
Intelligent image-based single-cell isolation
Machine learning algorithms are trained to predict cellular phenotypes
Image segmentation using deep learning
Automated single-cell isolation using laser microdissection
Deep Visual Proteomics
Bioinformatic data analysis
Ultra-high-sensitivity proteomics
Resource for researchers and clinicians
tSNE 2
tSNE 1
b_3
y_1
y_2
y_3

Fig. 1 | DVP concept and workflow. DVP combines high-resolution imaging, AI-guided image analysis for single-cell classification and isolation with an ultra-sensitive proteomics workflow[2]. DVP links data-rich imaging of cell culture or archived patient biobank tissues with deep-learning-based cell segmentation and machine-learning-based identification of cell types and states. (Un)supervised AI-classified cellular or subcellular objects of interest undergo automated LMD and MS-based proteomic profiling. Subsequent bioinformatics data analysis enables data mining to discover protein signatures, providing molecular insights into proteome variation in health and disease states at the level of single cells. tSNE, t-distributed stochastic neighbor embedding.

segmentation, feature extraction and ML-based phenotype classification. Building on a recent DL-based algorithm for cytoplasm and nucleus segmentation[3], 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[4]. To address this challenge, we used nucleAIzer[3] 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[5]. We trained a deep neural network with these synthetic images for specific segmentation of the cellular compartment of interest (for example, nucleus or cytoplasm; Fig. 2a). We benchmarked it against two leading DL approaches—unet4nuclei[6] and Cellpose[7]—and a widely used adaptive threshold-based and object-splitting-based method[8]. 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[3] where we performed an extensive comparison to additional methods and software (for example, ilastik[9], 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[10]. 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,d and 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[11].

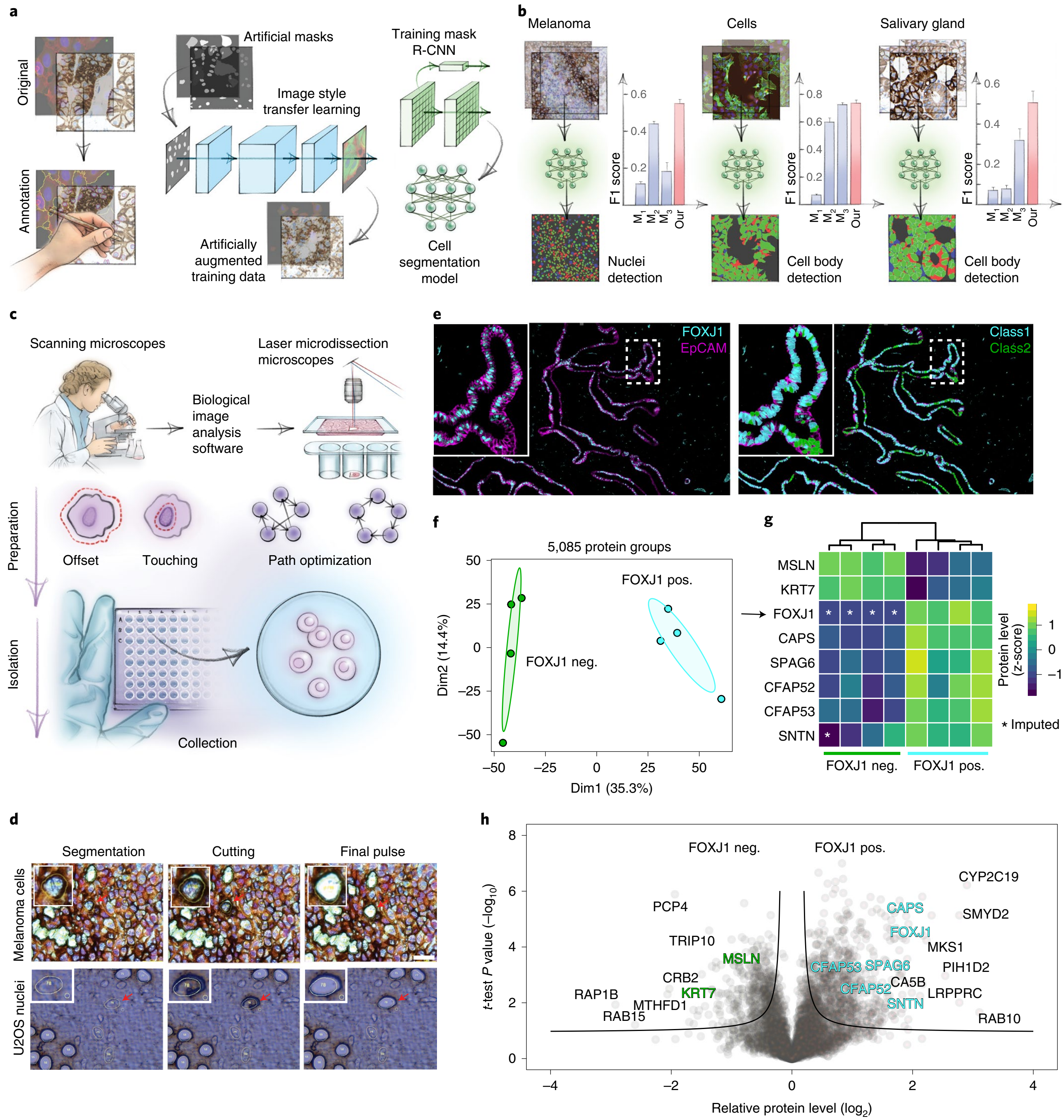


Fig. 2 | BIAS for integrative image analysis and automated LMD single-cell isolation. **a**, AI-driven nucleus and cytoplasm segmentation of normal-appearing and cancer cells and tissue using BIAS. **b**, We benchmarked the accuracy of its segmentation approach using the F1 metric and compared results to three additional methods—M_1 is unet4nuclei[6], M_2 is CellProfiler[8] and M_3 is Cellpose[7]—while OUR refers to nucleAIzer[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**, BIAS serves as the interface between the scanning and an LMD microscope, allowing high-accuracy transfers of cell contours between the microscopes. Illustration of cutting offset with respect to the object of interest and optimal path finding. **d**, Practical illustration of the functions in the upper panel. **e**, 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 = ×387. **f**, PCA of FOXJ1-positive and FOXJ1-negative cell proteomes. **g**, 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. **h**, 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, $s_0 = 0.1$, $n = 4$ biological replicates).

Table 1 | Mean F1 scores of the compared segmentation methods on our samples

Sample	Method			
	M_1	M_2	M_3	OUR
U2OS cyto	0.0667* ± 0.0075	0.5994 ± 0.0262	0.7205 ± 0.0152	**0.7336** ± 0.0218
Melanoma nuc	0.1126 ± 0.0151	0.4386 ± 0.0157	0.1801 ± 0.0504	**0.5498** ± 0.0231
Melanoma cyto	0.0058* ± 0.0021	0.0549 ± 0.0083	0.4859 ± 0.0354	**0.5536** ± 0.0625
Salivary gland nuc	0.0797 ± 0.0138	0.6488 ± 0.0430	0.0338 ± 0.0145	**0.7684** ± 0.0316
Salivary gland cyto	0.0714* ± 0.0151	0.0793 ± 0.0167	0.3174 ± 0.0588	**0.5051** ± 0.0586
Melanoma (pink) nuc	0.0682 ± 0.0183	0.2999 ± 0.0599	0.0364 ± 0.0238	**0.5079** ± 0.0392
Melanoma (pink) cyto	0.0261* ± 0.0070	0.0865 ± 0.0213	0.2659 ± 0.0429	**0.2839** ± 0.0229
Fallopian tube nuc	0.0006 ± 0.0009	0.3121 ± 0.0501	0.3160 ± 0.0631	**0.4724** ± 0.0683
Fallopian tube cyto	0.0016* ± 0.0023	0.0671 ± 0.0208	**0.4566** ± 0.0530	0.3455 ± 0.0473

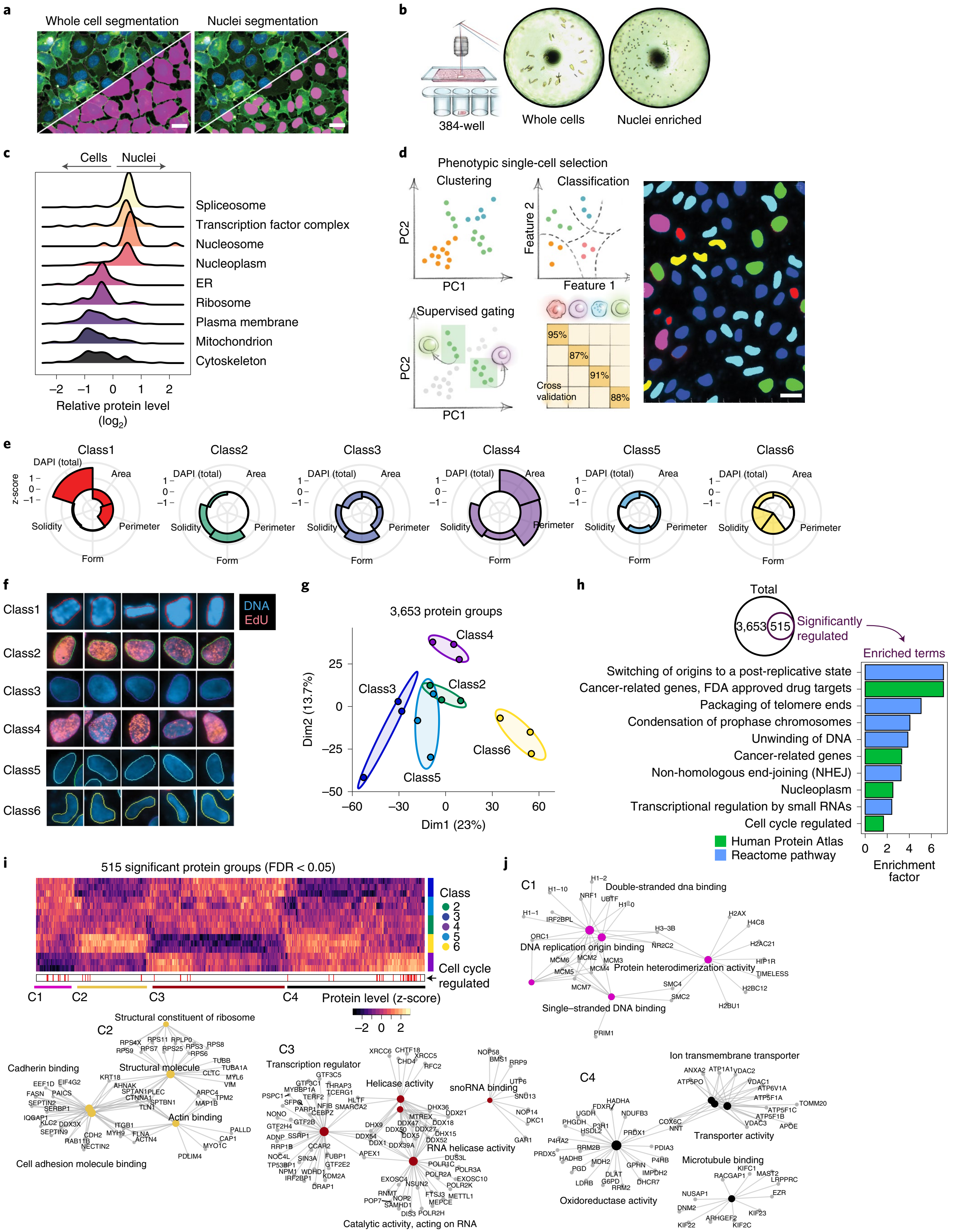
The methods are as follows: M_1 is unet4nuclei[6], M_2 is CellProfiler[8], M_3 is Cellpose[7] and OUR refers to nucleAIzer[3] (implemented in BIAS). High scores are highlighted in bold. Asterisks (*) mark that M_1 is intended for nucleus segmentation but was applied to segment cytoplasm. s.e.m. is displayed 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[12]—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 can now be applied to study the early onset of the disease without admixing unrelated cell types[13].

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[14]). 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[15], which we addressed in turn. We processed samples using our recently developed workflow for ultra-low sample input[2,16], 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[2,17]. 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[18] (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) $< 10^{-5}$) (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

Fig. 3 | DVP defines single-cell heterogeneity at the subcellular level. a, Segmentation of whole cells and nuclei in BIAS of DNA (DAPI)-stained U2OS cells. Scale bar, 20 µm **b**, Automated LMD of whole cells and nuclei into 384-well plates. Images show wells after collection. **c**, Relative protein levels (*x* axis) of major cellular compartments between whole cell ($n=3$ biological replicates) and nuclei ($n=3$ biological replicates) specific proteomes. *y* axis displays point density. **d**, Left: conceptual workflows of the phenotype finder model of BIAS for ML-based classification of cellular phenotypes. Right: results of unsupervised ML-based classification of six distinct U2OS nuclei classes based on morphological features and DNA staining intensity. Colors represent classes. Scale bar, 20 µm. **e**, Phenotypic features used by ML to define six distinct nuclei classes. Radar plots show z-scored relative levels of morphological features (nuclear area, perimeter, solidity and form factor) and DNA staining intensity (total DAPI signal). **f**, Example images of nuclei from the six classes identified by ML. Blue color shows DNA staining intensity, and red color shows EdU staining intensity to identify cells undergoing replication. Represented nuclei are enlarged for visualization and do not reflect actual sizes. **g**, PCA of five interphase classes based on 3,653 protein groups after data filtering. Replicates of classes ($n=3$ biological replicates) are highlighted by ellipses with a 95% confidence interval. **h**, Enrichment analysis of proteins regulated among the five nuclei classes. Significant proteins (515 ANOVA significant, FDR < 0.05, $s_0=0.1$) were compared to the set of unchanged proteins based on Gene Ontology Biological Process (GOBP), Reactome pathways as well as cell cycle and cancer annotations derived from the Human Protein Atlas (HPA)[20]. A Fisher's exact test with a Benjamini–Hochberg FDR of 0.05 was used (Supplementary Table 3). **i**, Unsupervised hierarchical clustering of all 515 ANOVA significant protein groups (Supplementary Table 4). Cell-cycle-regulated proteins reported by the HPA are shown in the lower bar. Nuclei classes ($n=3$ biological replicates) are shown in the row bar. C1–C4 show clusters upregulated in the different nucleus classes. **j**, Network analysis of enriched pathways for protein clusters C1-C4. Pathway enrichment analysis was performed with the ClusterProfiler R package[36]. ER, endoplasmic reticulum; PC, principal component.


a
Whole cell segmentation
Nuclei segmentation
b
384-well
Whole cells
Nuclei enriched
c
Cells
Nuclei
Spliceosome
Transcription factor complex
Nucleosome
Nucleoplasm
ER
Ribosome
Plasma membrane
Mitochondrion
Cytoskeleton
−2 −1 0 1 2
Relative protein level ($\log_2$)
d
Phenotypic single-cell selection
Clustering
Classification
PC2
PC1
Feature 2
Feature 1
Supervised gating
95%
87%
91%
88%
Cross validation
e
Class1
Class2
Class3
Class4
Class5
Class6
DAPI (total)
Area
Perimeter
Form
Solidity
z-score
1
0
−1
f
Class1
Class2
Class3
Class4
Class5
Class6
DNA
EdU
g
3,653 protein groups
Class4
Class3
Class2
Class6
Class5
Dim2 (13.7%)
Dim1 (23%)
25
0
−25
−50
−60 −30 0 30 60
h
Total
3,653
515
Significantly regulated
Enriched terms
Switching of origins to a post-replicative state
Cancer-related genes, FDA approved drug targets
Packaging of telomere ends
Condensation of prophase chromosomes
Unwinding of DNA
Cancer-related genes
Non-homologous end-joining (NHEJ)
Nucleoplasm
Transcriptional regulation by small RNAs
Cell cycle regulated
🟩 Human Protein Atlas
🟦 Reactome pathway
0 2 4 6
Enrichment factor
i
515 significant protein groups (FDR < 0.05)
Class
2
3
4
5
6
Cell cycle regulated
C1
C2
C3
C4
Protein level (z-score)
−2 0 2
j
C1
Double-stranded dna binding
DNA replication origin binding
Protein heterodimerization activity
Single-stranded DNA binding
H1–10 H1–2 NRF1 UBTF H1–0 H1–1 IRF2BPL ORC1 H3–3B NR2C2 H2AX H4C8 H2AC21 HIP1R TIMELESS MCM6 MCM2 MCM3 MCM5 MCM4 MCM7 SMC4 SMC2 H2BC12 H2BU1 PRIM1
C2
Structural constituent of ribosome
Cadherin binding
Structural molecule
Actin binding
Cell adhesion molecule binding
RPS4X RPS11 RPLP0 RPS9 RPS7 RPS25 RPS3 RPS8 RPS6 TUBB TUBA1A KRT18 CLTC MYL6 EEF1D EIF4G2 AHNAK VIM FASN PAICS SPTAN1 PLEC ARPC4 TPM2 SEPTIN2 CTNNA1 SPTBN1 MAP1B SERBP1 TLN1 IQGAP1 KLC2 DDX3X ITGB1 PALLD SEPTIN9 CDH2 MYH9 FLNA CAP1 RAB11B ACTN4 MYO1C NECTIN2 PDLIM4
C3
Transcription regulator
Helicase activity
snoRNA binding
RNA helicase activity
Catalytic activity, acting on RNA
XRCC6 CHTF18 XRCC5 CHD4 RFC2 NOP58 BMS1 RRP9 GTF3C5 GTF3C1 THRAP3 MYBBP1A TCERG1 UTP6 PSPC1 TERF2 SNU13 SFPQ HLTF NONO PARP1 NFIB SMARCA2 CEBPZ DHX36 DDX21 NOP14 DKC1 GTF2B MTREX GTF2H4 DHX9 DDX47 DDX18 DDX50 DDX27 ADNP SSRP1 DDX5 DHX15 DDX54 DDX52 RRP1B CCAR2 DDX1 DDX39A GAR1 APEX1 NOC4L SIN3A FUBP1 DUS3L TP53BP1 GTF2E2 POLR1C POLR3A NPM1 WDHD1 KDM2A POLR2A EXOSC10 IRF2BP1 EXOSC4 NSUN2 POLR2K DRAP1 RNMT NOP2 FTSJ3 METTL1 POP7 SAMHD1 MEPCE DIS3 POLR2H
C4
Ion transmembrane transporter
Transporter activity
Oxidoreductase activity
Microtubule binding
ANXA2 ATP1A1 VDAC2 ATP5PO VDAC1 ATP6V1A ATP5F1A TOMM20 ATP5F1C ATP5F1B VDAC3 APOE HADHA FDXR NDUFB3 UGDH PHGDH COX6C NNT HSDL2 P3H1 PRDX1 P4HA2 RRM2B PDIA3 PRDX5 HADHB MDH2 GPHN P4HB PGD DLAT IMPDH2 LDHB G6PD DHCR7 RRM2 KIFC1 MAST2 RACGAP1 LRPPRC NUSAP1 EZR DNM2 ARHGEF2 KIF23 KIF22 KIF2C

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[19], we found significant enrichment in our regulated proteins ($\text{FDR} < 10^{-6}$). 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 nucleoli of classes 2, 4 and 3 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[19], we investigated C7orf50 in the human pathology atlas[20] where high expression was associated with favorable outcomes in pancreatic cancer (Extended Data Fig. 3g; $P<0.001$). 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. 3f,h).

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[21,22]. 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.

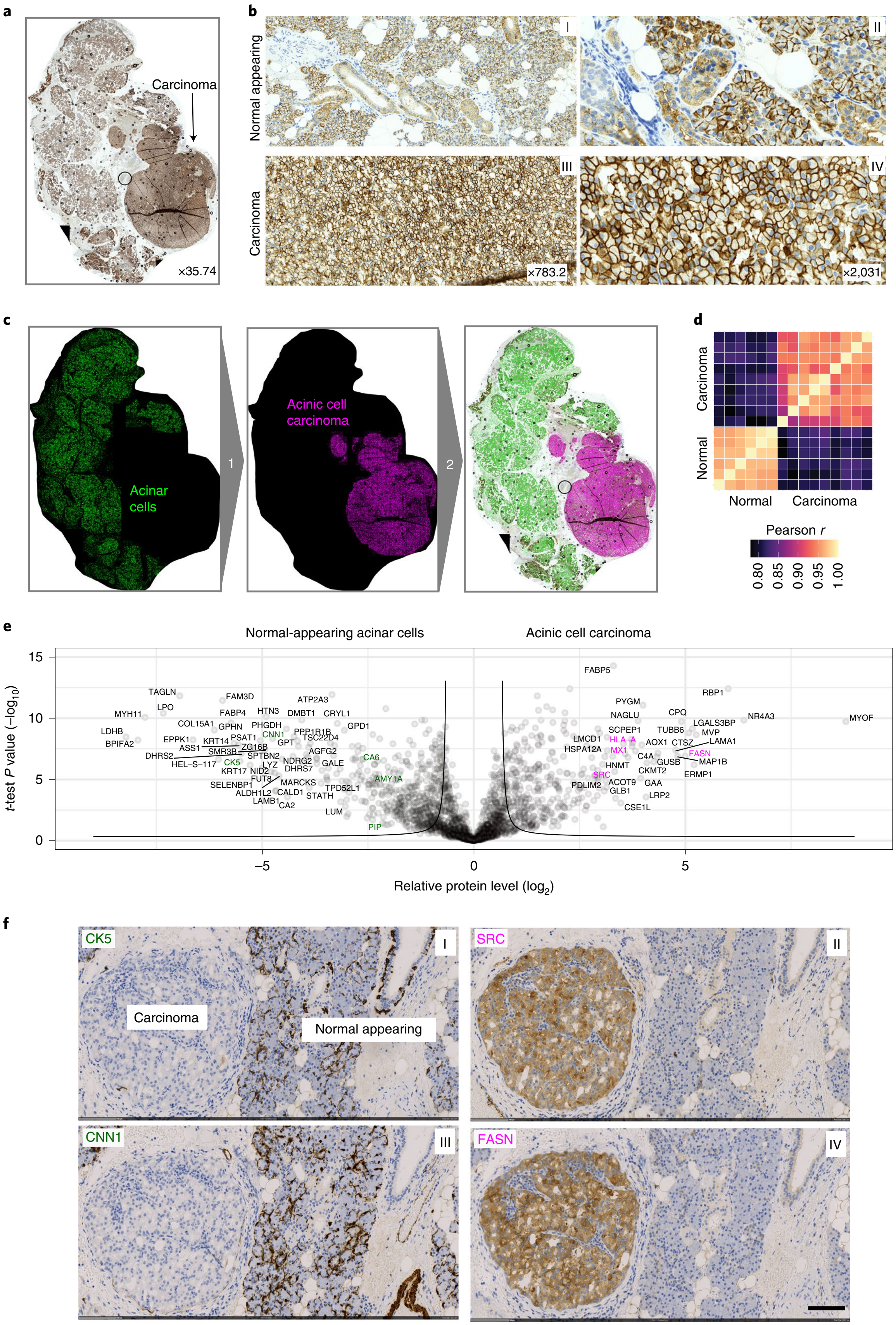
DVP applied to cancer tissue heterogeneity. Billions of patient samples are collected routinely during diagnostic workup and stored in the archives of pathology departments around the world[23]. The precise proteomic characterization of single cells in their spatial and subcellular context from tissue slides could have a tremendous clinical effect, complementing the emerging field of digital pathology[24]. 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 predominatly uniform tumor cells with round nuclei and abundant basophilic cytoplasm (Fig. 4a,b).

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. 4c 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[25] (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[25]. 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

Fig. 4 | DVP applied to archived tissue of a rare salivary gland carcinoma. a, IHC staining of an acinic cell carcinoma of the salivary gland using the cell adhesion protein EpCAM. **b**, Representative regions from normal-appearing tissue (upper panels I and II) and acinic cell carcinoma (lower panels III and IV) from **a**. **c**, DVP workflow applied to the acinic cell carcinoma tissue. DL-based single cell detection of normal-appearing (green) and neoplastic (magenta) cells positive for EpCAM. Cell classification based on phenotypic features (form factor, area, solidity, perimeter and EpCAM intensity). **d**, Proteome correlations of replicates from normal-appearing (normal, $n=6$) or cancer regions (cancer, $n=9$). **e**, Volcano plot of pairwise proteomic comparison between normal and cancer tissue. t-test significant proteins (two-sided t-test, $\text{FDR}<0.05$, $s_0=0.1$, $n=6$ biological replicates for normal and $n=9$ for cancer) are highlighted by black lines. Proteins more highly expressed in normal tissue are highlighted in green on the volcano's left, including known acinic cell markers (AMY1A, CA6 and PIP). Proteins more highly expressed in the acinic cell carcinoma are on the right in magenta, including the proto-oncogene SRC and interferon response proteins (MX1 and HLA-A; Supplementary Table 6). **f**, IHC validation of proteomic results. CNN1, SRC, CK5 and FASN are significantly enriched in normal or cancer tissue. Scale bar, 100 µm.

actionable therapeutic targets[26] (Fig. 4e). We validated the proteomic findings using IHC analysis of significantly enriched proteins in either normal-appearing or cancererous 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



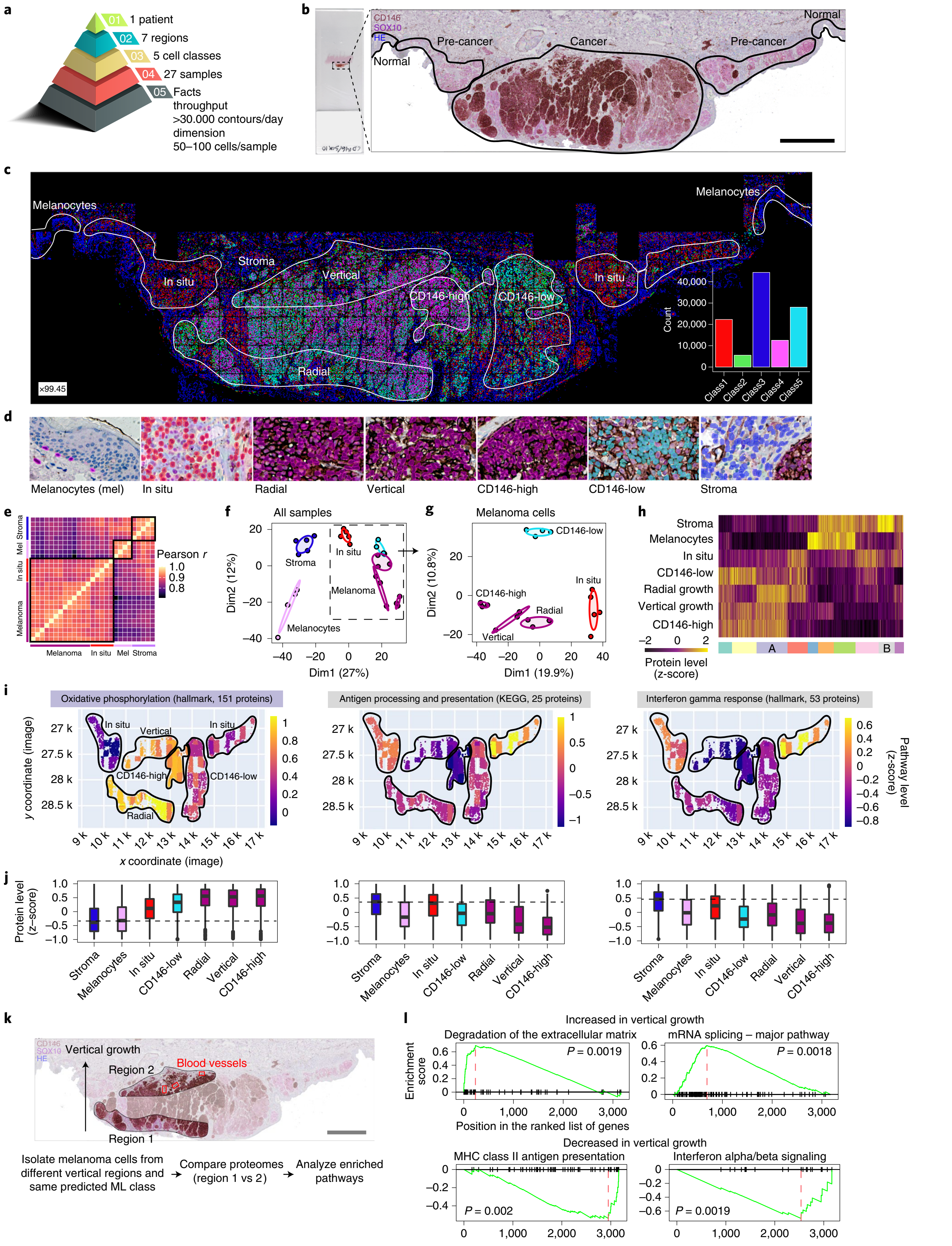

a
01 1 patient
02 7 regions
03 5 cell classes
04 27 samples
05 Facts
throughput
>30.000 contours/day
dimension
50–100 cells/sample
b
CD146
SOX10
HE
Normal
Pre-cancer
Cancer
Pre-cancer
Normal
c
Melanocytes
Melanocytes
In situ
Stroma
Vertical
CD146-high
CD146-low
In situ
Radial
×99.45
Count
40,000
30,000
20,000
10,000
0
Class1
Class2
Class3
Class4
Class5
d
Melanocytes (mel)
In situ
Radial
Vertical
CD146-high
CD146-low
Stroma
e
Stroma
Mel
In situ
Melanoma
Pearson r
1.0
0.9
0.8
Melanoma
In situ
Mel
Stroma
f
All samples
Stroma
In situ
Melanoma
Melanocytes
Dim2 (12%)
Dim1 (27%)
g
Melanoma cells
CD146-low
In situ
CD146-high
Radial
Vertical
Dim2 (10.8%)
Dim1 (19.9%)
h
Stroma
Melanocytes
In situ
CD146-low
Radial growth
Vertical growth
CD146-high
−2 0 2
Protein level (z-score)
A
B
i
Oxidative phosphorylation (hallmark, 151 proteins)
Antigen processing and presentation (KEGG, 25 proteins)
Interferon gamma response (hallmark, 53 proteins)
In situ
Vertical
In situ
CD146-high
CD146-low
Radial
y coordinate (image)
x coordinate (image)
27 k
27.5 k
28 k
28.5 k
9 k 10 k 11 k 12 k 13 k 14 k 15 k 16 k 17 k
Pathway level (z-score)
j
Protein level (z-score)
Stroma
Melanocytes
In situ
CD146-low
Radial
Vertical
CD146-high
k
CD146
SOX10
HE
Vertical growth
Region 2
Blood vessels
Region 1
Isolate melanoma cells from different vertical regions and same predicted ML class
Compare proteomes (region 1 vs 2)
Analyze enriched pathways
l
Increased in vertical growth
Degradation of the extracellular matrix
P = 0.0019
mRNA splicing – major pathway
P = 0.0018
Enrichment score
Position in the ranked list of genes
Decreased in vertical growth
MHC class II antigen presentation
P = 0.002
Interferon alpha/beta signaling
P = 0.0019
0 1,000 2,000 3,000

Fig. 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[36]. Enriched terms with an FDR < 0.05 are shown. MHC, major histocompatibility complex.

this highly metastatic disease. With pathogenic mutations in melanoma largely catalogued[27–29], we set out to directly study spatially resolved proteomes of distinct cellular phenotypes of melanoma progression (Fig. 5a,b and Extended Data Fig. 5a,b). We co-stained FFPE-embedded primary tumor material preserved for 17 years with two markers, SOX10 and CD146, to map melanoma cells. As overexpression of CD146 is implicated in melanoma progression, and immunotherapy against CD146 targets metastasis[30], we used CD146 as a disease progression marker in our analysis. ML predicted five classes with clearly defined spatial distribution: class 1, melanoma in situ; class 2, predominantly tumor; class 3, cells of the tumor microenvironment; class 4, enriched in CD146-high regions; and class 5, enriched in CD146-low regions. We used high-content imaging to determine the required number of cells to identify statistically and analytically robust cellular phenotypes for precise cell type and state isolation within a spatial region. For this reason, we typically collected around 100 cells per sample (Methods). Including replicates, we isolated and profiled 27 different samples obtained from seven unique regions of the same tissue section, including normal melanocytes, melanoma in situ and primary melanoma from the radial and vertical growth phases (Fig. 5a–d). We found high quantitative reproducibility among biological replicates, resulting in disease state and region-specific proteomes (Fig. 5e–g). Pre-cancerous (melanoma in situ) and primary melanoma showed differences in proteins involved in immune cell signaling and cell metabolism and coincided with reduced melanogenesis (Supplementary Table 7 and Extended Data Fig. 5d). The advanced stages (radial and vertical melanoma growth phase) showed well-defined activation of metabolic activation along with disease progression, a known hallmark of human cancers[31]. Expression of proteins involved in oxidative phosphorylation and mitochondria function gradually increased from melanocytes, melanoma in situ to invasive melanoma, indicating a dependency on mitochondrial respiration in the advanced tumor stages (Fig. 5h–j, Extended Data Fig. 5c and Supplementary Tables 7–9). Conversely, proteins involved in antigen presentation and interferon response were downregulated when compared to melanoma in situ (Fig. 5h–j and Supplementary Tables 7–9), in line with immune evasion strategies in melanoma[32].

Melanoma progression is a stepwise process involving radial and vertical growth phases. The direct comparison of these spatially defined regions of the same phenotype (class 4 cells) further highlighted critical features of cancer metastasis, such as extracellular matrix (ECM) remodeling (for example, collagen degradation) and upregulated PDGF signaling[33] (Fig. 5k,l, Extended Data Fig. 5e and Supplementary Table 10). These tumor-driven changes support growth, increase migration of tumor cells and remodel the ECM to facilitate metastasis to distant organs via adjacent blood vessels[33]. DVP also discovered a significant upregulation of mRNA splicing in the vertical compared to the radial growth phase. Pro-oncogenic alternative splicing has recently become a therapeutic strategy in oncology[34], and these tumors often present immunogenic neoantigens[35]. The increase in splicing coincided with a significant downregulation of immune-related signaling (interferon signaling and antigen presentation) (Fig. 5l and Supplementary Table 10), suggesting the transition from an immunogenic 'hot' to a 'cold' tumor zone in the vertical growth phase within the same tumor section. Clearly, DVP spatially resolved tumor heterogeneity by localizing tumor-related mRNA splicing, immune responses and ECM remodeling pathways in different regions.

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

1. Hériché, J.-K., Alexander, S. & Ellenberg, J. Integrating imaging and omics: computational methods and challenges. *Annu. Rev. Biomed. Data Sci.* **2**, 175–197 (2019).

2. Brunner, A. et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol. Syst. Biol.* **18**, e10798 (2022).
3. Hollandi, R. et al. nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.* **10**, 453–458 (2020).
4. Smith, K. & Horvath, P. Active learning strategies for phenotypic profiling of high-content screens. *J. Biomol. Screen.* **19**, 685–695 (2014).
5. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. Preprint at https://arxiv.org/abs/1611.07004 (2016).
6. Caicedo, J. et al. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat. Methods* **16**, 1247–1253 (2019).
7. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2020).
8. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
9. Berg, S. et al. ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods* **16**, 1226–1232 (2019).
10. Conrad, C. et al. Micropilot: automation of fluorescence microscopy-based imaging for systems biology. *Nat. Methods* **8**, 246–249 (2011).
11. Zhao, T. et al. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).
12. Lengyel, E. Ovarian cancer development and metastasis. *Am. J. Pathol.* **177**, 1053–1064 (2010).
13. Kurnit, K. C., Fleming, G. F. & Lengyel, E. Updates and new options in advanced epithelial ovarian cancer treatment. *Obstet. Gynecol.* **137**, 108–121 (2021).
14. Sakaue-Sawano, A. et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008).
15. Altelaar, A. M. & Heck, A. J. Trends in ultrasensitive proteomics. *Curr. Opin. Chem. Biol.* **16**, 206–213 (2012).
16. Coscia, F. et al. A streamlined mass spectrometry-based proteomics workflow for large-scale FFPE tissue analysis. *J. Pathol.* **251**, 100–112 (2020).
17. Meier, F. et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
18. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).
19. Mahdessian, D. et al. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* **590**, 649–654 (2021).
20. Uhlen, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419–1260419 (2015).
21. Venturini, V. et al. The nucleus measures shape changes for cellular proprioception to control dynamic cell behavior. *Science* **370**, eaba2644 (2020).
22. Arias-Garcia, M., Rickman, R., Sero, J., Yuan, Y. & Bakal, C. The cell–cell adhesion protein JAM3 determines nuclear deformability by regulating microtubule organization. Preprint at https://www.biorxiv.org/content/10.1101/689737v2.full (2020).
23. Kokkat, T. J., Patel, M. S., McGarvey, D., Livolsi, V. A. & Baloch, Z. W. Archived formalin-fixed paraffin-embedded (FFPE) blocks: a valuable underexploited resource for extraction of DNA, RNA, and protein. *Biopreserv. Biobank* **11**, 101–106 (2013).
24. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *Lancet Oncol.* **20**, e253–e261 (2019).
25. Zhu, S., Schuerch, C. & Hunt, J. Review and updates of immunohistochemistry in selected salivary gland and head and neck tumors. *Arch. Pathol. Lab. Med.* **139**, 55–66 (2015).
26. Kim, L. C., Song, L. & Haura, E. B. Src kinases as therapeutic targets for cancer. *Nat. Rev. Clin. Oncol.* **6**, 587–595 (2009).
27. Shain, A. H. et al. The genetic evolution of melanoma from precursor lesions. *N. Engl. J. Med.* **373**, 1926–1936 (2015).
28. Pollock, P. M. et al. High frequency of *BRAF* mutations in nevi. *Nat. Genet.* **33**, 19–20 (2003).
29. Raamsdonk, C. D. V. et al. Frequent somatic mutations of *GNAQ* in uveal melanoma and blue naevi. *Nature* **457**, 599–602 (2009).
30. Wang, Z. et al. CD146, from a melanoma cell adhesion molecule to a signaling receptor. *Signal Transduct. Target Ther.* **5**, 148 (2020).
31. Kumar, P. R., Moore, J. A., Bowles, K. M., Rushworth, S. A. & Moncrieff, M. D. Mitochondrial oxidative phosphorylation in cutaneous melanoma. *Br. J. Cancer* **124**, 115–123 (2021).
32. Eddy, K. & Chen, S. Overcoming immune evasion in melanoma. *Int. J. Mol. Sci.* **21**, 8984 (2020).
33. Winkler, J., Abisoye-Ogunniyan, A., Metcalf, K. J. & Werb, Z. Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Nat. Commun.* **11**, 5120 (2020).
34. Zhang, Y., Qian, J., Gu, C. & Yang, Y. Alternative splicing and cancer: a systematic review. *Signal Transduct. Target Ther.* **6**, 78 (2021).
35. Frankiw, L., Baltimore, D. & Li, G. Alternative mRNA splicing in cancer immunotherapy. *Nat. Rev. Immunol.* **19**, 675–687 (2019).
36. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

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

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 51-year-old male and located at the left upper chest. TNM stage at diagnosis was T3aN1M0. The histological subtype was superficial spreading melanoma; the Clark level was 4; and the Breslow thickness was 2.27 mm. Tumor immune infiltration was categorized as non-brisk. 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 of a 29-year-old male. There was no sign of mitosis, necrosis de-differentiation or perineural or intravascular growth. The tumor cells were positive in EpCAM, CK7, DOG1 and SOX10. Mammaglobin was negative. The sample was 4 years old, and the patient is currently disease-free. The study was carried out in accordance with institutional guidelines under approval by the local Medical Ethics Review Committee (SJ-742) and the Data Protection Agency (REG-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-year-old female and was macroscopically and histologically normal appearing. All patients consented before surgery. Patient-derived tissues were obtained fresh or paraffin-embedded according to an approved institutional review board protocol (13372B) 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.

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 U2OS FUCCI cells were kindly provided by Atsushi Miyawaki[14]. These cells are endogenously tagged with two fluorescent proteins fused to the cell cycle regulators CDT1 (mKO2-hCdt1$^+$) and geminin (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% CO_2 humidified environment in McCoy's 5A (modified) medium GlutaMAX supplement (Thermo Fisher Scientific, 36600021) supplemented with 10% FBS (VWR) without antibiotics.

U2OS cells stably expressing a membrane-targeted form of eGFP were generated by transfection with plasmid Lck-GFP (Addgene, 61099 (ref. [37])) and culturing in selection medium (DMEM medium containing 10% FBS, penicillin–streptomycin and 400 μg ml^{-1} of Geneticin) under conditions of limited dilution to yield single colonies. A clonal cell line with homogenous and moderate expression levels of Lck-eGFP at the plasma membrane was established from a single colony.

All cell lines were tested for mycoplasma (MycoAlert, Lonza) and authenticated by STR profiling (IdentiCell).

IHC staining on membrane slides. Membrane PEN slides 1.0 (Zeiss, 415190-9041-000) were treated with UV light for 1 hour and coated with APES (3-aminopropyltriethoxysilane) using VECTABOND reagent (Vector Labs, SP-1800-7) according to the manufacturer's protocol. FFPE tissue sections were cut (2.5 µm), air dried at 37 °C overnight and heated at 60 °C for 20 minutes to facilitate better tissue adhesion. Next, sections were deparaffinized, rehydratrated and loaded wet on the fully automated instrument Omnis (Dako). Antigen retrieval was conducted using Target Retrieval Solution pH 9 (Dako, S2367) diluted 1:10 and heated for 60 minutes at 90 °C. Single stain for EpCAM (Nordic BioSite, clone BS14, BSH-7402-1, dilution 1:400) and sequential double stain for SOX10/CD146 (SOX10, Nordic BioSite, clone BS7, BSH-7959-1, dilution 1:200; CD146, Cell Marque, clone EP54, AC-0052, dilution 1:400) was performed, and slides were incubated for 30 minutes (32 °C). After washing and blocking of endogenous peroxidase activity, the reactions were detected and visualized using EnVision FLEX, High pH kit (Dako, GV800 and GV809/GV821) according to the manufacturer's instructions. In the double stain, EnVision DAB (Dako, GV825) and EnVision Magenta (Dako, GV900) substrate chromogen systems were used for visualization of CD146 and SOX10, respectively. Finally, slides were rinsed in water, counterstained with Mayer's hematoxylin and air dried without mounting.

IHC staining for validation of DVP studies. FFPE tissue sections were cut (2.5 µm), placed on coated slides (Agilent/Dako, K8020) and air dried vertically before heating at 60 °C for 20 minutes to facilitate tissue adhesion. Next, slides were loaded on the fully automated instrument Omnis. Sections were dewaxed, and antigen retrieval was conducted using Target Retrieval Solution High pH (Agilent/Dako, GV804, diluted 1:50) at 97 °C for 24 minutes. Subsequently, the sections were incubated with the primary antibodies. We selected antibodies assessed and approved by a board-certified consultant pathologist. Proto-oncogene tyrosine protein kinase SRC/c-Src (Cell Signaling Technology, clone 36D10, 2109, dilution 1:3,200), fatty acid synthase/FASN (Cell Signaling Technology, clone C20G5, 3180, dilution 1:100), calponin-1/CNN1 (Cell Marque, clone EP63, AC-0060, dilution 1:300) and cytokeratin 5/CK5 (Leica Biosystems, clone XM26, NCL-L-CK5, dilution 1:200) for 30 minutes at 32 °C. After washing and blocking of endogenous peroxidase activity, the reactions were detected and visualized using EnVision FLEX, High pH kit (Agilent/Dako, GV800 and GV809/GV821) 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-ethynyl-2′-deoxyuridine (EdU) for 20 minutes and then fixed for 5 minutes at room temperature with 4% paraformaldehyde (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 GB mount (GBI Labs, E01-18).

For validation experiments (Extended Data Fig. 3), 96-well glass-bottom plates (Greiner SensoPlate Plus, Greiner Bio-One) were coated with 12.5 μg ml^{-1} of human fibronectin (Sigma-Aldrich) for 1 hour at room temperature. Immunocytochemistry was carried out following an established protocol[38]. Then, 8,000 U2OS cells were seeded in each well and incubated in a 37 °C and 5% CO_2 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 3×5 minutes. Rabbit polyclonal HPA antibodies targeting the proteins of interest were diluted in blocking buffer (PBS + 4% FBS) at 2–4 μg ml^{-1} along with primary marker antibodies (see below) and incubated overnight at 4 °C. Cells were washed with PBS for 4×10 minutes and incubated with secondary antibodies (goat anti-rabbit Alexa Fluor 488 (A11034, Thermo Fisher Scientific), goat anti-mouse Alexa Fluor 555 (A21424, Thermo Fisher Scientific) and goat anti-chicken Alexa Fluor 647 (A21449, Thermo Fisher Scientific)) in blocking buffer at 1.25 μg ml^{-1} for 90 minutes at room temperature. Cells were counterstained in 0.05 μg ml^{-1} of DAPI for 15 minutes, washed with for 4×10 minutes and mounted in PBS.

Primary antibodies used were as follows:

For C7orf50 cell cycle validation: mouse anti-ANLN at 1.25 μg ml^{-1} (amab90662, Atlas Antibodies)

Mouse anti CCNB1 at 1 μg ml^{-1} (610220, BD Biosciences)

Rabbit anti-C7orf50 at 1 μg ml^{-1} (HPA052281, Atlas Antibodies)

For human fallopian tube tissue, FFPE tissue sections (2.5 µm) were mounted and pre-processed as described above. Thereafter, tissue was dewaxed by washing 2×2 minutes in 100% xylene, followed by a series of 100%, 95% and 70% ethanol for 1 minute, respectively, and 3×1 minute in ddH_2O. 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 1:200, 14452, Cell Signaling Technology) were diluted in 10% goat serum and incubated overnight at 4 °C in a humidified chamber. Tissue specimens were washed 5× in TBST and secondary antibodies for the visualization of FOXJ1 (Alexa Fluor 647 goat anti-mouse, dilution 1:200, A21235, 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 5× in TBST, followed by 2× 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. Bright-field acquisition was obtained using the VIS LED light source and a CCD 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 C7orf50 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 Cell Profiler software[8] using DAPI for nuclei 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 C7orf50 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 ×63/0.70 (tissue) or ×40/0.60 (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 µm^3; BNID 100434), in agreement with estimations in spatial transcriptomics analysis[39].

Leica LMD7 cutting accuracy (Leica R&D, patent EP1276586)
For ×150 objective: $\frac{10}{150} = 0.07$ µm

Segmentation methods and accuracy evaluation. nucleAIzer[3] models were integrated into BIAS and customized for these experiments by retraining and refining the nucleus and cytoplasm segmentation models. First, style transfer[5] learning was performed as follows. Given a new experimental scenario such as our melanoma or salivary gland tissue sections stained immunohistochemically, the acquisition of which produces such an image type that no annotated training data exist for, 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[40] (generative adversarial networks): foreground to mimic, for example, nuclei, and background for surrounding, for example, tissue structures. Parallelly, artificial masks of either nucleus 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 segment the target compartments.

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 M$_1$ in Fig. 2a and S1)[6] and Cellpose (M$_3$)[7]—alongside a widely used, conventional adaptive threshold-based and object splitting-based application (M$_2$)[8]. We note that M$_1$ is not intended for cytoplasm segmentation (see details in ref. [6] and below). Segmentation accuracy according to the F1 metric is displayed as bar plots (Fig. 2b, Extended Data Fig. 1a, Table 1 and Supplementary Table 1), and visual representation in a color-coded manner is also provided.

unet4nuclei[6] is optimized to segment nuclei on cell culture images; Cellpose[7] is an approach intended for either nucleus or cytoplasm segmentation on various microscopy image types; and CellProfiler[8] is a conventional threshold-based and object splitting-based software broadly used in the bioimage analysis community. unet4nuclei, as its name suggests, is primarily intended for nucleus segmentation and uses a U-Net-based network after pre-processing of input images and then post-processes detected objects. Cellpose uses a vector flow representation of instances, and its neural network (also based on U-Net) predicts and combines horizontal and vertical flows. unet4nuclei has successfully been applied in nucleus segmentation of cell cultures, whereas Cellpose is able to generalize well on various image modalities even outside microscopy and can be used to segment nuclei and cytoplasms. However, as most segmentation methods, neither is able to adapt to a new image domain, such as a particular experiment type (for example, IHC salivary gland tissue), without re-training on newly created ground truth annotations. On the contrary, our segmentation algorithm (nucleAIzer[3]) is able to do so via the image style transfer approach mentioned above. Obviously, conventional algorithms cannot adapt either; thus, they need to be re-parameterized for each experiment. For the evaluation, an expert CellProfiler user was asked to optimize a pipeline for each sample type to the best possible segmentation result, and then all images per sample type were segmented with one pipeline (corresponding to the given sample).

We evaluated our segmentation performance (and comparisons) according to the F1 score metric calculated at the 0.7-IoU (intersection over union) threshold. IoU, also known as Jaccard index, was calculated from the overlapping region of the predicted (segmented) object with its corresponding ground truth (real) object at a given threshold (see formulation below). True-positive (TP), false-positive (FP) and false-negative (FN) objects were counted accordingly, if they had an IoU greater than the threshold *t* (in our case, 0.7), to yield the F1 score at this threshold (see formulation below). Segmentation evaluation was performed on 10–20 randomly selected 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 experts using AnnotatorJ[41]. We included images from all relevant regions of 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[3] 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).

$$Jaccard\ index = \frac{|x \cap y|}{|x \cup y|} = \frac{|x \cap y|}{|x| + |y| - |x \cap y|}$$

$$precision(t) = \frac{TP(t)}{TP(t) + FP(t)}$$

$$recall(t) = \frac{TP(t)}{TP(t) + FN(t)}$$

$$F1\ score(t) = 2 \cdot \frac{precision(t) \cdot recall(t)}{precision(t) + recall(t)}$$

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 pevious study[3] 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, M$_2$ and ilastik[9].

We also note that previous methods, such as CellProfiler or ilastik, can perform accurate segmentation of cells; moreover, the performance of M$_2$ on tissue nucleus segmentation is remarkable. On the other hand, robust methods (for example, DL-based) offer the convenience of not needing to reset most 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, 0030129547). 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 µm^2 per sample was collected for an estimated number of 100–200 cells based on the average HeLa cell volume of 2,000 µm^3 (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 2,000*g* 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 5× 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 LysC was added (pre-diluted in ultra-pure water to 4 ng µl^{-1}) 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^{-1}) 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 2,000*g*. 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[42]. Next, 30 µg of purified peptides was separated on a 30-cm C18 column in 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 minutes, followed by an increase to 60% for 5 minutes and a 10-minute wash in 95% buffer B at 300 nl min^{-1}. 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. 4) or data-independent (diaPASEF) 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 $1/K_0 = 1.6$ V s cm^{-2} to 0.6 V s cm^{-2}. Precursor ions for MS/MS analysis were isolated with a 2-Th window for $m/z < 700$ and 3-Th for $m/z > 700$ 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 $1/K_0 = 1.6$ V s cm^{-2} to 20 eV at $1/K_0 = 0.6$ V s cm^{-2}. Singly 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 eV at $1/K_0 = 1.6$ V s cm^{-2} to 20 eV at $1/K_0 = 0.6$ V s cm^{-2}. We used the ddaPASEF method for library generation.

Data analysis of proteomic raw files. Mass spectrometric raw files acquired in ddaPASEF mode (Fig. 4) were analyzed with MaxQuant (version 1.6.7.0)[43,44]. The UniProt database (2019 release, UP000005640_9606) was searched with a peptide spectral match and protein-level FDR of 1%. A minimum of seven amino acids was required, including N-terminal acetylation and methionine oxidation as variable modifications. Due to omitted reduction and alkylation, cysteine carbamidomethylation was removed from fixed modifications. Enzyme specificity was set to trypsin with a maximum of two allowed missed cleavages. First and main search mass tolerance was set to 70 p.p.m. and 20 p.p.m., respectively. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs (MBR) with a 0.7-minute retention time match window and a 0.05 $1/K_0$ ion mobility window. Label-free quantification was performed with the MaxLFQ algorithm[45] and a minimum ratio count of 1.

For diaPASEF measurements (Figs. 2, 3 and 5), raw files were analyzed with DIA-NN[46] (version 1.8). To generate a project-specific spectral library, a 24-fraction high-pH reversed-phase fractionated precursor library was created from the same tissue specimen and acquired in ddaPASEF mode, as described above. Raw files were analyzed with MSFragger[47] under default settings (with the exception that cysteine carbamidomethylation was removed from fixed modifications) to generate the library file used in DIA-NN. The library consisted of 90,056 precursors, 79,802 elution groups and 7,765 protein groups.

Bioinformatic analysis. Proteomics data analysis was performed with Perseus[48] 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_0 value[49] of 0.1 was used for the pairwise proteomic comparison in Figs. 2h and 4e. 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[36] (Supplementary Tables 7 and 10), the ReactomePA package[50] and the WebGestalt gene set analysis toolkit (WebGestaltR)[51], 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 aggregated 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.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository[52] 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 database[53] (https://www.ebi.ac.uk/biostudies/) with the accession number S-BSST820. 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-BSST820), containing all features applied in the described workflows. Several major components of our work are available in open-source repositories (Supplementary Table 11).

References

37. Benediktsson, A. M., Schachtele, S. J., Green, S. H. & Dailey, M. E. Ballistic labeling and dynamic imaging of astrocytes in organotypic hippocampal slice cultures. *J. Neurosci. Methods* **141**, 41–53 (2005).
38. Stadler, C., Skogs, M., Brismar, H., Uhlén, M. & Lundberg, E. A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* **73**, 1067–1078 (2010).
39. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).
40. Goodfellow, J. P.-A. I. J. & Bengio, Y. Generative adversarial networks. *Proc. International Conference on Neural Information Processing Systems* 2672–2680 (2014).
41. Hollandi, R., Diosdi, A., Hollandi, G., Moshkov, N. & Horvath, P. AnnotatorJ: an ImageJ plugin to ease hand annotation of cellular compartments. *Mol. Biol. Cell* **31**, 2179–2186 (2020).
42. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics*. *Mol. Cell Proteomics* **16**, 694–705 (2017).
43. Prianichnikov, N. et al. MaxQuant software for ion mobility enhanced shotgun proteomics*. *Mol. Cell Proteomics* **19**, 1058–1069 (2020).
44. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
45. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteomics* **13**, 2513–2526 (2014).
46. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
47. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
48. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
49. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
50. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2015).
51. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
52. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
53. Sarkans, U. et al. The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* **46**, D1266–D1270 (2017).
54. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

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 F. Schlaudraff from Leica for technical support; T. Danka and M. Kovács 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 846795 (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 161L0222. 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 (ERAPERMED-COMPASS, ERAPERMED-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

Conceptualization: A.M. F.C., P.H. and M.M.; Methodology: A.M., F.C., A.D.B., M.B., B.D.A. and M.M.; Software: R.H., F.K., A.K. and P.H.; Investigation: A.M., F.C. and R.H.; Formal analysis: A.M., F.C. and R.H.; Writing—original draft: A.M., F.C., P.H. and M.M.; Writing—review and editing: all authors; Resources: all authors.; Data curation: L.M.R.G., M.B., S.N., A.M., F.C., R.H., F.K., A.K., A.S., E.M., L.S., M.A.E., E. Lengyel and P.H.; Visualization: A.M., F.C., A.S. and R.H.; Project administration: A.M. and P.H.; Supervision: M.M.; Funding acquisition: F.C., P.H., E. Lundberg and M.M.

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.