Bioimage Informatics for Phenomics

Thomas Walter

Habilitation à diriger des recherches

Sorbonne Université

UFR 919: Faculté d'ingénierie

Bioimage Informatics for Phenomics

Thomas Walter

Researcher, CR1

Centre for Computational Biology

Mines ParisTech

Habilitation soutenue le 24 Juin 2020 à Paris devant le jury composé de:

Isabelle BLOCH	*Présidente*
Charles KERVRANN	*Rapporteur*
Xavier DESCOMBES	*Rapporteur*
Andres SANTOS	*Rapporteur*
Franck PEREZ	*Examinateur*
Peter HORVATH	*Examinateur*
Daniel RACOCEANU	*Examinateur*

Abstract

While we have the technologies and computational tools to analyze entire genomes, transcriptomes and proteomes, the computational description of phenotypes resulting from this molecular basis is still lagging behind. Yet, the quantitative description of the diverse aspects of the phenome is a prerequisite for understanding the complex genotype-phenotype relationships in living systems.

High Content Screening (HCS) allows to systematically explore many different aspects of the phenome, in particular cell morphology, the dynamics of cellular behavior and the spatial distribution of transcripts and proteins inside cells. Monitoring and analyzing the changes in these aspects upon perturbation by gene silencing or drug treatment have the potential to unravel the relationship between these cellular properties and the molecular mechanisms that regulate them. Similarly, the analysis of stained tissue slides allows to study architectural changes depending on disease related variables.

Large-scale imaging approaches, such as HCS and histopathology, thus provide information that is complementary to information at the molecular level, traditionally studied in bioinformatics. In order to make best use of these challenging and complex large-scale image data sets, we need robust and sophisticated methods capable of integrating a large set of image features in order to reach a biologically meaningful description of the data. For this reason, computer vision is the method of choice for computational phenotyping.

This manuscript describes my contributions to the field of computational phenotyping by computer vision. After an introduction to the field of Bioimage Informatics as well as some background on High Content Screening, I will describe a number of different projects I have been working on over the last years exemplifying the different types of information that can be studied with images: (1) analysis of morphological phenotypes by supervised learning, (2) analysis of temporal information, in the form of phenotypic and spatial trajectories, (3) analysis of localization patterns, i.e. the spatial distributions of biomolecules inside cells and (4) analysis at the tissular scale.

These methods have been applied to large-scale screens on cell division and migration, namely the first genome-wide screen by time-lapse microscopy in a human cell line. Some more recent applications include the study of the spatial aspects of gene expression, where we aim at understanding the patterns according to which RNA localize inside cells, as well as the field of digital pathology, where we wish to predict clinical variables, such as outcome or response to treatment, from large stained images of diseased tissue. While often used for diagnostic purpose, histopathology data is also informative about cellular phenotypes and therefore allows to bridge the gap between phenotypic analysis at the cellular level and implications for disease (at the patient level).

The most striking evolution in this field is the advent of deep learning, that has revolutionized computer vision over the last years. I will discuss the role deep learning is going to play in the near future with respect to the different applications mentioned above, and the methodological or conceptional developments that are most promising for each application in turn.

Altogether, one of the major challenges in bioinformatics today is to establish relationships between the molecular level (e.g. the level of genetic mutations or transcripts) and the level of an entire biological system (e.g. cell or even patient level) by analysis of large-scale omics data sets. These associations need to cross 9 orders of magnitude. The projects and methods I am showing in this manuscript will contribute to bridging this gap.

Résumé

Nous possédons aujourd'hui les technologies expérimentales et computationnelles pour analyser des génomes, transcriptomes et protéomes dans leur intégralité. Cependant, les méthodes permettant la description et l'analyse computationnelle des phénotypes — résultant de cette base moléculaire — sont encore insuffisantes. Or, la description quantitative des divers aspects du phénome est une condition sine qua non pour la connaissance approfondie des relations complexes entre génotype et phénotype.
Le criblage à haut débit nous permet d'explorer de façon systématique un grand nombre d'aspects du phénome, en particulier en ce qui concerne la morphologie cellulaire, le comportement dynamique des cellules et la distribution spatiale des transcrits et protéines à l'intérieur des cellules. L'observation et l'analyse du changement de ces propriétés en réponse à des perturbations (e.g. l'extinction d'un gène ou un traitement pharmacologique) font apparaître le lien fonctionnel entre les perturbations agissant au niveau moléculaire et les variables phénotypiques. De façon similaire, l'analyse des coupes histologiques permet d'étudier les changements architecturaux qui interviennent au niveau tissulaire en réponse à une maladie.
Le criblage à haut débit et l'histopathologie génèrent donc de grands jeux de données qui donnent accès à une information complémentaire à celle portant sur l'échelle moléculaire, traditionnellement étudiée en bioinformatique. Afin de pleinement exploiter le potentiel de ces données complexes et massives, nous avons besoin de méthodes robustes et sophistiquées, capables d'intégrer une grande quantité de caractéristiques d'images. Pour cette raison, la vision par ordinateur s'impose comme choix méthodologique afin d'arriver à une description biologiquement pertinente.
Ce manuscrit décrit mes contributions au domaine du phénotypage computationnel par des méthodes de vision par ordinateur. Après une introduction générale qui portera sur les problématiques actuelles en bio-imagerie et le criblage à haut contenu, je décrirai les travaux que j'ai menés ces dernières années au regard des différents types d'information que l'on peut étudier avec des images : (1) L'analyse de phénotypes morphologiques par l'apprentissage supervisé (2) L'analyse de l'information temporelle sous forme de trajectoires phénotypiques et spatiales (3) L'analyse des distributions spatiales de biomolécules à l'intérieur des cellules (4) L'analyse de l'échelle tissulaire.
En termes d'applications biologiques, je présenterai l'analyse que j'ai réalisée du premier crible à l'échelle génomique par vidéo-microscopie dans des cellules humaines, qui a permis d'étudier la division et la migration cellulaires. Mes projets plus récents incluent l'étude des aspects spatiaux de l'expression génétique, où l'on veut comprendre la répartition spatiale intra-cellulaire des ARNs, ainsi que la pathologie digitale. C'est un sujet en plein essor, dans lequel on souhaite prédire des variables cliniques, comme la réponse au traitement des données histopathologiques.
L'évolution la plus remarquable dans ce domaine ces dernières années est la montée de l'apprentissage profond, qui a révolutionné la vision par ordinateur. Je discuterai du rôle que l'apprentissage profond jouera en analyse d'images biologiques, ainsi que des développements méthodologiques et conceptuels les plus prometteurs pour chacune des applications mentionnées plus haut.
Pour conclure, l'un des défis majeurs en bioinformatique est d'étudier les relations entre le niveau moléculaire (e.g. les mutations génétiques) et le niveau d'un système biologique intégral (e.g. le patient) par l'analyse des données omics. Relier l'un à l'autre nécessite de franchir 9 ordres de grandeur. Les projets et méthodes que je présente dans ce manuscrit contribueront à combler ce fossé.

Contents

1 Introduction

In this introductory chapter, I will present the context of my research project. Bioimage Informatics is an emerging discipline at the interface between image analysis and computational biology, aiming at providing the computational tools to answer biological questions from the analysis of image data. This covers a broad range different problems and methodological approaches, ranging from image reconstruction and inverse problems to image classification and data integration. I will start by giving a systematic overview over the different problems and approaches in the field. Then, I will focus in some more detail on Computational Phenotyping, where we aim at quantifying the visual readout provided by microscopy in order to infer biologically relevant information. Given this overall context, I will finish this chapter by giving an overview of my trajectory and the projects I had the opportunity to contribute to.

1.1 Bioimage Informatics - an emerging discipline

In the last two decades, technological developments, such as next-generation sequencing, have triggered revolutionary changes in the life sciences. In particular, the many new technologies have reinforced the quantitative aspects of biology and encouraged ambitious, large-scale projects involving a large number of research institutions worldwide. Examples include fundamental research projects aiming at deciphering the human genome (International Human Genome Sequencing Consortium 2001; International Human Genome Sequencing Consortium 2004) and studying its variability (ENCODE Project Consortium 2007; ENCODE Project Consortium 2012; 1000 Genomes Project Consortium 2010). Unlike most traditional research projects in life sciences in the past, where the main contribution consisted in the written publication, these large-scale projects reach their main impact by the generation of large data repositories, open to the scientific community, and thus building a precious resource for future research projects. The emergence of these enormous efforts to collect and maintain high-quality data describing the molecular aspects of life has transformed biology to become - to some extend - a data science. This shift in focus, as well as the wealth and amount of data currently acquired in studies of all scales, has resulted in an ever increasing importance of the development of sophisticated and robust computational methods and tools, thus initiating the emergence of bioinformatics as a discipline, which is now a well-established and recognized field of research.

Long before these technical breakthroughs imaging has been and still is one of the most important and most widely used experimental techniques in biology. Importantly, imaging approaches are complementary to most molecular approaches in several ways. In particular, they allow one:

1. To investigate the morphological properties of biological entities, often informative about biological function.

2. To explore the temporal dimension. Microscopy is certainly among the most efficient techniques when it comes to studying the behavior of biological systems over time.

3. To explore the spatial dimension of living systems, such as the spatial arrangements of proteins inside cells, or the organization of cells in a developing organism.

4. To study living systems at different scales of organization: the molecular, cellular, tissular and organism scale are accessible to imaging approaches (see Figure 1.1). In many cases, we can even have access to two or more of these scales at the same time.

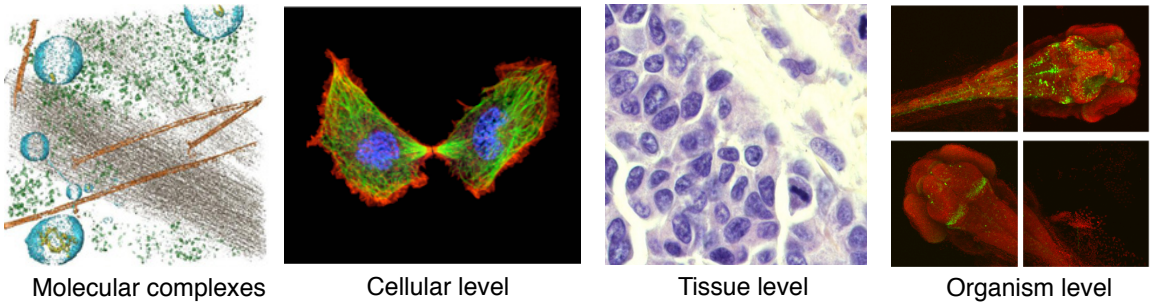

Figure 1.1: Bioimages give access to multiple scales, allow to study cellular phenotypes and to explore the spatial dimension of living systems.

On the downside, imaging also faces limitations, as compared to molecular techniques.

1. While imaging approaches have become much more powerful with the advent of fluorescence microscopy (see also Annexe A), the number of marked biomolecules in the same sample is necessarily limited to only a handful. In contrast to sequencing technologies, imaging approaches are thus not comprehensive; only a small fraction of the biomolecules present in a sample is usually imaged. While there are some notable exceptions — discussed in section 4.3 — the lack of comprehensiveness remains a key limitation for most imaging approaches in biology.

2. While sequencing (and many other molecular techniques) has become much cheaper over the last years, the price of imaging systems has not decreased so dramatically. Consequently, large imaging studies can represent an important financial burden for research laboratories.

3. The information contained in an image does not allow for a direct and straightforward representation usable in an analysis workflow. Indeed, images were initially not meant for quantitative analysis and rather suggested subjective and qualitative interpretation; unlike for expression or sequence data, it is often a difficult issue to extract meaningful information from an image.

The need for quantification schemes optimized for microscopy data has been recognized already in the seventies and eighties (Young 1972; Meyer 1986). While computational analysis of images has been an exception at that time, image quantification and analysis have now become an essential part of microscopy-based assays and also a requirement for publication in most journals (Swedlow et al. 2003). This evolution has been further amplified by the emergence of imaging techniques that require sophisticated algorithms for image generation, such as super-resolution or light-sheet microscopy, and of large-scale imaging projects, where the size and complexity of image data make manual analysis unpractical. In this context, **bioimage informatics** has emerged as a new discipline (Peng 2008; Peng et al. 2012; Danuser 2011; Myers 2012; Cardona et al. 2012; Kervrann et al. 2016; Jug et al. 2014). Bioimage informatics is the discipline that covers all methodological and software aspects that are necessary to answer biological questions by computational analysis of image data. Bioimage informatics is therefore at the interface

between bioinformatics, biology, microscopy, image analysis, image processing, data handling, mining and visualization.

Tasks for Bioimage Informatics There are many reasons justifying the use of image processing and analysis for biology, and consequently the diversity of concrete tasks and application areas is huge. Here, I give a short overview over the different categories of approaches in this field.

Image reconstruction aims at generating one image from a series of acquired raw images. Each of these individual raw images cannot be used by itself to interpret the biological information in the sample. For instance, in PALM (Betzig et al. 2006) and STORM (Rust et al. 2006), the two most popular techniques for super-resolution microscopy, a series of images is taken, each resulting from the stochastic emission of a subset of fluorophores present in the image (for more details, see Annex A). Only processing the entire series of images and their combination into one super-resolved image of the sample can reveal structures inside the cells at high resolution. The generated image, i.e. the image that is presented to the biologist, therefore involves already heavy processing. Similar situations arise in other microscopy techniques, such as Optical Coherence Tomography (Sharpe et al. 2002) or light-sheet microscopy (Huisken et al. 2004), where each single image corresponds to the projection of the sample in one direction, and the ultimate image to be generated is the 3D volume (see also Annex A).

Image enhancement and image restoration aims at estimating from the observation an image that is closest in some sense to the (unknown) original object by using prior knowledge about the kind of perturbation (such as the optical system or the corrupting noise). This is known as an *inverse problem*: the rationale is to estimate an original image X, from an observed image $Y = f(X)$, where $f(\cdot)$ models the kind of perturbation we are trying to revert. This is achieved by minimizing a cost function $J(X, Y) = D(f(X), Y) + \mathcal{R}(X)$, where D is a measure for the divergence between observed and estimated data and $\mathcal{R}$ expresses constraints on the original object (Sarder et al. 2006).

Image quantification / Image analysis extracts biologically meaningful information from images or image series, i.e. the input is an image or a series of images, and the output is a measurement or a distribution of measurements. This category includes a large panel of methodologically different problems. While often the actual measurements may seem trivial to perform (such as intensity measurements, cell size, number of microtubules or particle speed), the underlying workflows can be complex and often involve classical but challenging image analysis problems, such as object detection, segmentation and tracking. Segmentation and tracking are thus core problems in Bioimage Informatics, and many of the challenges organized in this domain tackle these questions (e.g. (Chenouard et al. 2014; Vladimir 2017)).

Combining experiments One of the limitations of imaging approaches in biology mentioned above is the limited number of markers that can be used in experiments. Some recently developed experimental techniques to partly overcome this limitation (Schubert et al. 2006; Chen et al. 2015) are only applicable in specific setups and challenging to perform routinely. Another strategy is to address this problem computationally by mapping images taken with different markers to a common reference volume. This strategy is called atlas based registration, and has been applied to developing embryos (Fowlkes et al. 2008) and full grown organs (Peng et al. 2011). At the cellular level, this is even more challenging due to the large morphological variability between cells, but has also been addressed by the use of generative models (Zhao et al. 2007). The most recent development in this category relies on the hypothesis that standard microscopy images contain much more information than what can be grasped by our visual system. It is therefore possible to predict fluorescence images without actually acquiring them

(Christiansen et al. 2018). Another challenging application is the combination of images taken under different modalities, for instance by light and electron microscopy (Paul-Gilloteaux et al. 2017).

Computer Vision and computational phenotyping The phenotype of a biological system, such as a cell or an organism, is the set of its observable properties. Computational phenotyping refers to automatically inferring a quantitative description of a phenotype from images. This is related to image quantification and analysis in that we aim at transforming an image into a set of numbers, but the objective is to jointly analyze a relatively large number of properties allowing thus to reach an understanding of the image in biological terms. The computational discipline that is concerned with the understanding of images rather than analyzing one aspect is called *Computer Vision*. For this reason, computational phenotyping and computer vision are intimately related. Examples include the detection of subcellular protein localization patterns (Boland et al. 1998) or classification of cellular phenotypes (Walter et al. 2010a). In section 1.2, I will explain these approaches in more detail.

Visualization of the increasingly complex and massive data sets has also become an important issue in Bioimage Informatics (Walter et al. 2010b). One difficulty in visualization stems from the fact that image data sets are carrying more and more information, either in the form of additional dimensions (space, time, channels) or just in size, as it is often the case for tissue imaging. Such data sets tend to be huge (tens of GB to TB of data), and can contain detailed information on hundreds of thousands of individual cells. Typical visualization problems are (i) how to navigate through massive raw data, (ii) how to display multi-dimensional analysis results on top of the image data and (iii) how to summarize trends in the data in a concise way that can be communicated to other researchers. In particular for 3D volumes, an additional question arises when we want to annotate objects or regions in an intuitive and effective way (Alessandro et al. 2016). An additional problem occurs when the number of experiments is large, e.g. in High Content Screening (HCS), where a single data set can consist in thousands of single experiments. Here, the task is to provide concise overviews of the quantitative data derived from the images, ideally with some capacity to link back to the initial image data (Jones et al. 2008; Antal et al. 2015).

Image data bases have also become increasingly important over the last decade, and not only for large-scale projects with tens of thousands of experiments. In the frame of reproducible research it is essential to store not only the actual images, but also the exact experimental settings under which the image was taken, as well as the measurements that were taken from the image and the underlying computational workflow. The first step towards such a system was actually to find a common image format supported by virtually all microscope vendors. This was a huge effort that brought together academic and industrial partners, who eventually agreed on a common format that coexists today with proprietary formats and allows independent analysis groups to access the data (Linkert et al. 2010). Today, we also see a number of database solutions specialized in microscopy images (Goldberg et al. 2005; Linkert et al. 2010; Williams et al. 2017).

Bioimage Informatics and Medical image analysis Analysis of medical images has a long tradition and is perhaps the closest relative of Bioimage Informatics, as in both cases, the imaged objects are living systems. While there are indeed some common conferences (such as ISBI, MICCAI) and journals, the two communities remain overall relatively separated. There is of course an important overlap in terms of methodologies, but there are also major differences in the nature of the data and the underlying scientific questions:

1. Medical Imaging naturally focuses on the human body, often at the scale of the organ. In bioimaging, there is a large variety of scales at which imaging experiments might be performed; the object of interest might be protein complexes, organels, cells, tissues and evolving or full-grown organisms.

2. The image acquisition modes in the bioimaging field are less standardized than in medical imaging. In the medical field, images are taken in clinical routine with clear and strict protocols. In biology, acquisition techniques are rapidly evolving and protocols tend to change relatively often for a variety of reasons. Changes in the experimental protocol or the acquisition techniques imply important differences in the image characteristics, but provide researchers also with an additional degree of flexibility in order to make computational analysis easier by improving the image quality.

3. Moreover, even if the same object (e.g. a cell) is imaged with the same microscope, the signal characteristics and image analysis tasks will entirely depend on what part of the cell is highlighted with fluorescence markers. This makes bioimages highly variable.

4. Generally, medical imaging aims at providing a large variety of diagnostic tools; bioimaging aims at answering scientific questions in biology. This difference has a huge impact on the computational methods in the two domains: while the actual objective of medical image analysis is mostly to improve the diagnostic power of medical examinations and is therefore clearly defined and standardized, the objective of bioimage informatics can be very variable and usually evolves during a project.

5. As a consequence of the aforementioned considerations, there are more large-scale annotated datasets in the medical field than in the domain of bioimaging. This has an important impact on methodological choices, too, in particular with respect to the deep learning methodology, which is today the predominant technique used in medical imaging, in particular in computer aided diagnosis.

Despite these differences, Medical Image Analysis has had a huge impact on Bioimage Informatics, both in terms of methodological developments and in terms of community activities and organization. Also, the differences mentioned above only describe general trends with notable exceptions. For instance, it is of course also possible to answer fundamental scientific questions with medical imaging approaches, e.g. in computational neuroscience, where MRI data are often analyzed in view of understanding the functioning of the human brain rather than to perform diagnostic tasks. Moreover, there are also imaging modalities that are used both in fundamental biology and for diagnostic purposes (e.g. histology).

Altogether, Bioimage Informatics deals with all computational aspects of bioimaging, covering methodological developments in diverse fields and ultimately aims at supporting microscopy and inferring biological knowledge from image data.

1.2 Phenomics and computational phenotyping

The phenotype of a living system describes its observable physical and biochemical characteristics that result from the complex interaction of the genetic code and environmental factors (Houle et al. 2010). In analogy to *genomics* where we study the entirety of the genetic code, the term *phenomics* suggests that we aim at studying the full set of all possible characteristics of a living system. While this is of course not possible, as we have never access to all variables of

a living system, we understand by *phenomics* the systematic study of complex phenotypes in form of a high-dimensional set of observable properties as a function of changes in the genome, gene expression or environmental factors, as opposed to over-simplified approaches where we characterize a living system by only one or very few variables (such as survival).

Imaging approaches are in general well suited to phenomics studies, as they give access to a large number of fundamental properties of cells, tissues and organisms, while maintaining their spatial organization and morphological integrity. In particular, the use of fluorescent markers (Chalfie et al. 1994), a technique widely used in biology since the nineties, allows to highlight predefined sets of biomolecules. This has been instrumental for microscopy, because it allows one to either study the spatial distribution of these molecules, thereby linking imaging technologies to the molecular techniques, or to infer properties of the marked biological structures (structures where the marked protein is known to localize), such as cytoskeletal arrangements or nuclear morphologies. While by using imaging techniques, we can obtain rich information on many properties of the biological system informative on many biological processes, we must be aware that most of the system's variables remain hidden. This is very different from molecular techniques in the field of phenomics, such as transcriptomics.

High Content Screening for Phenomics In order to study phenotypes systematically as a function of genetic mutations, changes in gene expression or chemical perturbations, we need to perform a large number of experiments under controlled conditions. The set of experimental techniques allowing to perform a large number of biological experiments in parallel relying on a high degree of automation, is referred to as High Throughput Screening (HTS). Traditional HTS usually provides little information in each experiment, such as a single number informative on the number of cells or the overall cytotoxicity. In High Content Screening (HCS), each one of the experiments is characterized by high information content, as provided by imaging experiments. For cell based assays for instance, this means that single cells can be distinguished and analyzed individually. In HCS, large numbers of different conditions can be tested with respect to their effect on cells and organisms. We can thus obtain systematic views on the phenotypic space living systems explore in response to these perturbations (see Figure 1.2). HCS is therefore an excellent tool for phenomics.

High Content Screening in fundamental research In fundamental research, genetic screening aims at deciphering the molecular basis of cellular processes. For this, the expression of genes is altered, e.g. the expression of genes can be increased (overexpression) or reduced (knock-downs, e.g. by RNA interference) or completely prevented (knock-out). The consequence of this change in expression on the cellular phenotypes is monitored by microscopy employing fluorescent reporters tailored to the biological question. The rationale of these approaches is that the analysis of the loss- or gain-of-function phenotypes will provide us with information on the function of the down-regulated or overexpressed gene (Pepperkok et al. 2006; Carpenter et al. 2004; Echeverri et al. 2006). Such large-scale studies have been successfully applied to unravel the molecular basis of cellular processes as diverse and as fundamental as cell migration (K. J. Simpson et al. 2008), protein secretion (J. C. Simpson et al. 2012), endocytosis (Collinet et al. 2010) and cell division (Neumann et al. 2010). Another example for large-scale imaging studies are localization screens where the objective is to understand the spatial distribution of proteins or RNA inside cells and organisms, the localization mechanisms (such as active transport) and functional impact of localization. The subcellular localization of proteins for instance has been the subject of systematic studies for many years (Boland et al. 1998; Glory et al. 2007; Coelho et al. 2010; Regev et al. 2017; Thul et al. 2017) and triggered the development of may computational methods in the field. Localization studies are also very popular at the

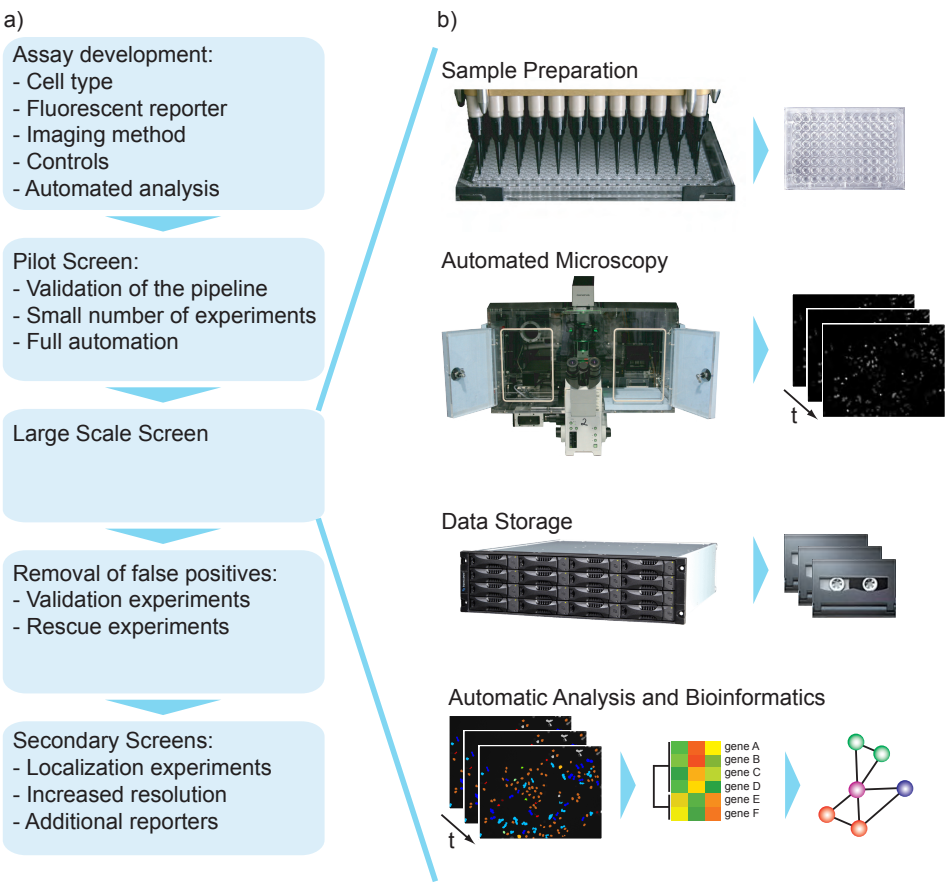

Figure 1.2: Project workflow for high-throughput screening. (A) After the design of the assay is completed, a pilot screen is performed, applying the entire assay pipeline on a small scale to identify and correct potential short- comings. The subsequent large-scale screen then typically results in a list of candidate hits that must be validated. In secondary screens with higher spatial, temporal, and phenotypic resolutions, more information on these validated hits can be collected and integrated. (B) The typical workflow for a large-scale screen based on time-lapse microscopy consists of sample preparation, automated time-lapse microscopy, professional data storage and back- up and automatic analysis, and bioinformatics. Figure is taken from (Terjung et al. 2010)

level of tissues (Uhlen et al. 2015) and organisms (Lécuyer et al. 2007; Lécuyer et al. 2008). Both functional and localization screens provide us with functionally relevant information on gene products in cells, tissues and organisms and are therefore important tools for functional genomics.

High Content Screening in drug screening Similar experimental and computational workflows are being used for the discovery and the phenotypic characterization of drugs, as they constitute an interesting compromise between throughput and richness of functional information on the drug effect on living systems (Perlman et al. 2004; Loo et al. 2007). In comparison to more traditional high-throughput screens, such as cytotoxicity or viability assays, they provide a complex picture of drug-induced phenotypes, bearing information on the perturbed biological processes and pathways. Importantly, they also allow scoring for phenotypic heterogeneity of the drug response (D. K. Singh et al. 2010; Slack et al. 2008), which has major implications for the efficiency of drugs.

High Content Screening data sets HCS generates large, systematic and comprehensive data sets, which have the potential to become a scientific resource that ideally complements other omics data; they allow for the exploration of the phenotypic space (Yin et al. 2013) and the phenotypic annotation of genes, thereby contributing to the understanding of complex genotype-phenotype relationships. In order to make best use of these rich data sets and allow meta-studies combining several screens (on different processes, in different models, etc.), it is essential to develop tools to adequately present these data to the scientific community (Neumann et al. 2010; Antal et al. 2015; Williams et al. 2017) and to standardize the formats of phenotypic description (Hoehndorf et al. 2012). These technical developments will allow - on the long run - to re-mine and integrate different existing data sets and to perform meta-analyses (Schoenauer Sebag et al. 2015b; Schoenauer Sebag et al. 2015a; Suratanee et al. 2014; Pau et al. 2013). Today, screening approaches are no longer limited to a few specialized laboratories, and access to screening platforms is currently organized by large national and supranational initiatives, such as Euro Bioimaging (http://www.eurobioimaging.eu) at the European and France Bioimaging (FBI, http://france-bioimaging.org) at the national level in France. Consequently, many more large-scale HCS data sets are going to be acquired in the next years complementing and extending the existing genome-wide surveys in different ways.

Cellular phenotyping in their tissular context High Content Screening traditionally focuses on cells in culture, i.e. cells that are heavily modified in order to allow for controlled experiments and optimal imaging conditions. While such approaches have been instrumental to elucidate the molecular basis of many fundamental biological processes and to infer the mechanism of action of drugs, they are limited to these rather artificial conditions and also do not allow to study the more complex interplay of cells in a tissue or an organism, as well as the role of their spatial organization. In order to address these questions, researchers have developed a plethora of assays in developing organisms (Donà et al. 2013; Jug et al. 2014; Guirao et al. 2015). In addition, thanks to experimental advances in digital pathology and large-scale initiatives like *The Cancer Genome Atlas (TCGA)* (International Cancer Genome Consortium 2010), huge image data bases of stained tumor tissue sections are available bearing information both at the cellular and the tissue level. While stained tissue sections are usually complex to analyze automatically due to the specificities of the staining procedure and the tissue variability, these large-scale datasets represent an exciting resource bridging the gap between the cellular and the tissular scale. Each tissue slide tends to be very large and may contain information on hundreds of thousands of cells, each of which needs to be segmented and analyzed, which is a formidable challenge.

In all of these approaches, acquired image data sets are becoming increasingly massive and complex. In both basic research and drug discovery, as well as tissue phenotyping, there is therefore a strong need for increasingly sophisticated, yet robust and systematically studied methods to analyze these data and for software tools allowing the scientific community to apply these methods to newly acquired data sets.

1.3 From clinical applications to fundamental research and back

This manuscript gives an overview over my research activities in the years 2006-2020. My research focuses on the development of methods for computational phenotyping. Nevertheless, I would like to mention the research I have been concentrating on before and that initiated my interest in this exciting field.

In my PhD thesis at the Centre of Mathematical Morphology (CMM) at Mines ParisTech, I

developed methods for the Computer Aided Diagnosis of Diabetic Retinopathy, a severe and frequent eye disease, related to diabetes (Walter et al. 2001; Walter et al. 2002a; Walter et al. 2002b; Walter et al. 2005; Walter et al. 2007). My work focused on the detection of lesions (such as microaneurysms, hemorrhages and exudates) and anatomical landmarks (such as the vascular tree, the optic disk and the macula) in color images of the human retina (see Figure 1.3) in view of building a system for Computer Aided Diagnosis (CAD) of Diabetic Retinopathy, a severe and frequent eye disease. Today, 15 years later, DR has been the first disease for which the U.S. Food and Drug Administration (FDA) has approved a diagnostic device based on Computer Vision, whose predictions are not controlled by an expert (Abràmoff et al. 2018).

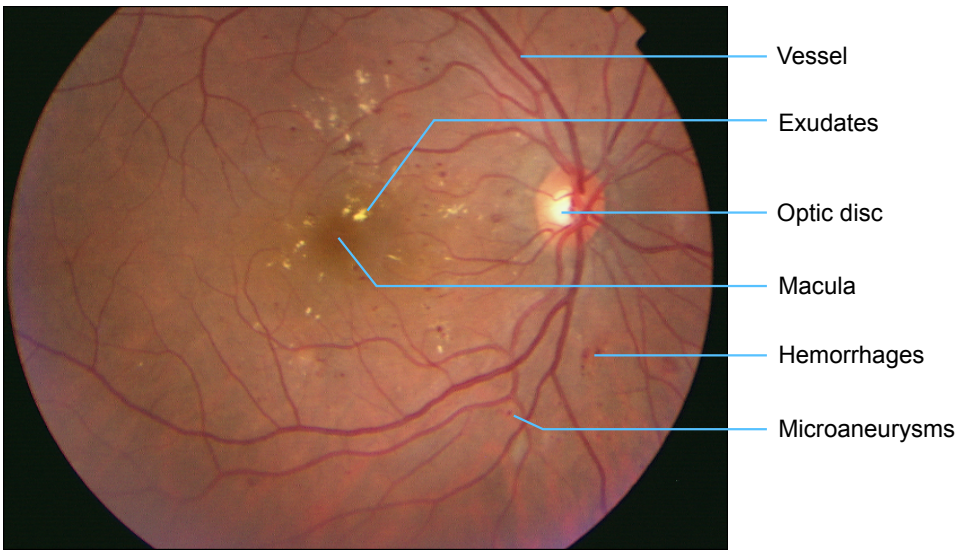

Figure 1.3: A color image of the human retina. The main lesions (microaneurysms, hemorrhages, exudates) and the main anatomical landmarks (vascular tree, optic disk, macula) are indicated.

While I continued to be interested in medical applications, I wanted to slightly shift the focus of my research from a more clinical application, where the ultimate goal is to support diagnosis by automation and/or quantification, to a more experimental field, where the objective is to make discoveries using image based assays. My first step in this direction was to start a project with the team of Pierre Corvol at the Collège de France. The objective of the project was to investigate the level of angiogenesis in response to drug treatment (Sihn et al. 2007). Angiogenesis is the main process of blood vessel formation and therefore plays an important role in tissue growth and development, but it is also a major hallmark of tumor progression. Blocking angiogenesis is therefore one important strategy in cancer treatment. In this work, I developed tools to quantify angiogenesis in chicken embryos in order to assess the anti-angiogenic properties of drugs by quantifying first and second order vessels, bifurcation points and overall vascular density (see Figure 1.4).

From a methodological point of view, these two projects were dominated by Mathematical Morphology, but I also started to get interested in Computer Vision and Machine Learning, namely in the context of lesion detection (Walter et al. 2007). Both Mathematical Morphology and Machine Learning remained the most important computational fields for my future work.

With this first experience in analyzing experimental rather than patient data, I decided to move to applications of image analysis in fundamental research in biology. In 2006, I moved to the European Molecular Biology Laboratory (EMBL) in Heidelberg, where I joined the group of Jan Ellenberg. Here, I started to work in the field of Computational Phenotyping and High Content

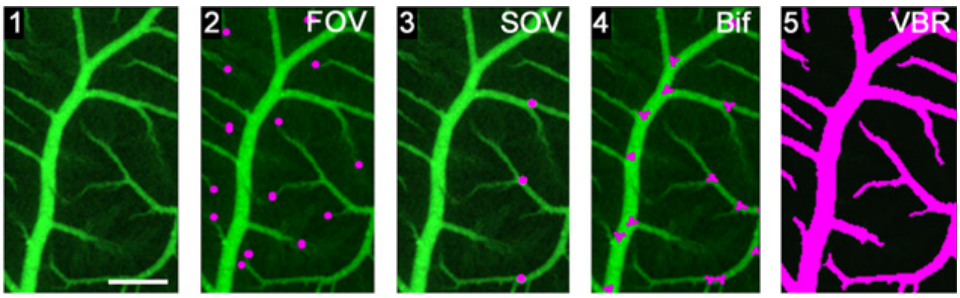

Figure 1.4: Quantification of angiogenesis in chicken embryos: (1) Detail of an angiography, (2) Detection of First Order Vessels (FOV), the vessels in vicinity to the capillaries, (3) Detection of Second Order Vessels (SOV), the vessels in vicinity to the first order vessels, (4) Bifurcation points, (5) Vessel segmentation, allowing to measure the overall vascular density.

Screening. I had the chance to join the Mitocheck-Project, a European project that aimed at a systematic study of cell division involving a large-scale screen generating 200.000 videos that needed to be analyzed. This gave me the opportunity to apply and extend many of the concepts I had learned during my PhD thesis, but it also shifted the focus from pure image analysis to a mix of computer vision, data mining and computational biology.

In 2012, I joined the Centre for Computational Biology, Mines ParisTech, where I started to build a team on Bioimage Informatics. There have been three main axes of research. First, I continued to work on live cell imaging data in view of analyzing mitotic and migration phenotypes. Concretely, I was interested in developing methods for the analysis of Secondary Screening data with higher temporal and spatial resolution, still in view of analyzing mitotic phenotypes. In addition, my team developed a methodological framework for the analysis of migration phenotypes from large-scale microscopy data. More recently, we proposed new methods for drug screening in multiple cell lines, each corresponding to different molecularly defined subtypes of a disease. The challenge was to compare phenotypes even if the baseline morphologies differ. The second research axis is dedicated to the emerging field of Spatial Transcriptomics, where we develop methods to analyze the spatial aspects of gene expression. This involves detection of individual transcripts in cells and classification of localization patterns. In the third axis, we develop methods to analyze stained tissue slides. Histopathology links information on cellular phenotypes with clinical relevance, and therefore closes the loop I started during my PhD thesis.

These three axes cover the complementary aspects of biological systems that can be explored with images: morphology, time, physical space and multi-scale organization. The organization of the manuscript follows roughly this organization rather than being strictly chronological: in chapter 2, I describe the methods I and my team have been developing for the analysis of morphological phenotyping. Chapter 3 is dedicated to the analysis of the temporal aspects of phenotypes. Chapter 4 deals with current and future projects in the field of RNA localization and spatial transcriptomics and chapter 5 focuses on our work on tissue phenotyping.

2 Computational methods for morphological phenotyping

As introduced in section 1.2, the phenotype of a cell refers to a potentially large set of observable parameters. The difficulty in the computational analysis of such data is thus to capture all of the interesting properties, to describe them quantitatively and to infer biological knowledge from these descriptors.

In the following, I will first discuss basic strategies in section 2.1 and review the relevant literature. In section 2.2, I discuss in detail methods that I have developed together with my collaborators for the computational analysis of morphological phenotypes and in particular the detection of mitotic phenotypes. Section 2.3 introduces the methodological extensions we developed in order to screen multiple cell lines. In section 2.4 I discuss the articles I published on this topic and detail the perspectives for the field to which I plan to contribute.

2.1 Overview

In this section, I will give an overview over different strategies for the computational analysis of phenotypes from large-scale screening data; see also excellent reviews (S. Singh et al. 2014; Caicedo et al. 2017) and comparative analyses (Kümmel et al. 2011; Ljosa et al. 2013) on this topic. Given the many possible imaging techniques, the presentation cannot be exhaustive, but I hope to cover the main options.

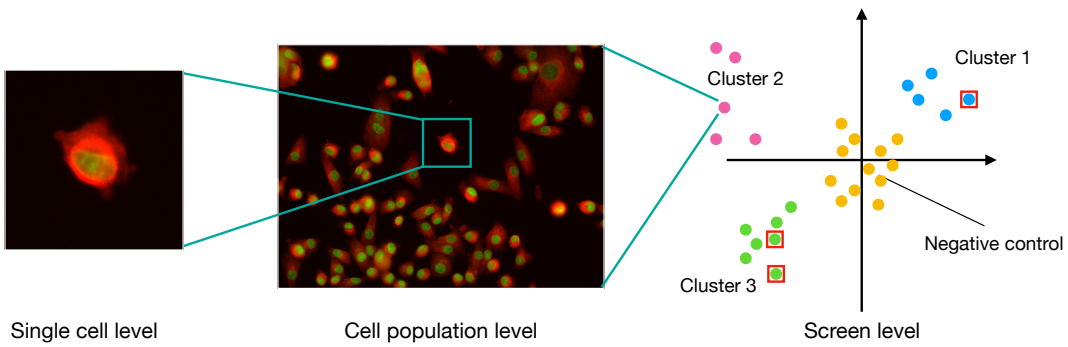

Figure 2.1: Simplified view on the different levels of computational phenotyping: (1) the phenotype of individual cells can be quantified (left). (2) The phenotypic description of individual cells can be aggregated to build the phenotypic profile of the cell population (middle). (3) Phenotypic profiles of cell populations can be further analyzed in order to identify groups of experimental conditions with similar effect (right panel). Different colors indicate different condition groups, orange corresponds to the negative controls (no effect). Red squares indicate experimental conditions with known effects. They can be used for interpretation of the result ("guilt by association").

The problem of computational phenotyping can be stated as follows: we start from a large

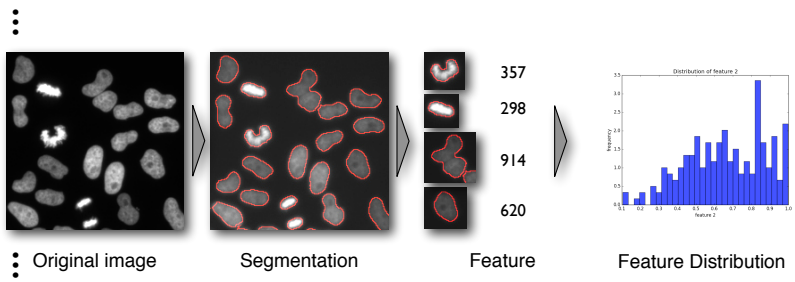

Figure 2.2: Phenotype description by a single feature distribution: for each segmented object, a single feature is calculated. The population is thus described by a univariate feature distribution.

number of microscopy experiments, each consisting in a single image with potentially several channels[1]:

$$\begin{aligned} I : E \subset \mathbb{Z}^D &\rightarrow \mathbb{R}^C \\ u &\rightarrow I(u) \end{aligned} \tag{2.1}$$

where D is the number of spatial dimensions (2 or 3) and C the number of channels, which is traditionally between 1 and 5, but which can - with some techniques - also reach several hundreds (Eng et al. 2019). Alternatively, images can also be defined as tensors of order 4: $I \in \mathbb{R}^{D_1 \times D_2 \times D_3 \times C}$, where $D_{1,2,3}$ are the spatial dimensions and C the number of channels. In this presentation, there is no formal difference between spatial and channel dimensions. Each of these images displays cellular populations under varying experimental conditions. Importantly, depending on the imaging modality and the choice of fluorescent markers, the part of the cell which is visible and thus amenable to computational analysis, might differ from project to project. We assume however, that there is at least one channel that allows us to identify individual nuclei or cells.

The first step usually consists in segmenting cells or cellular compartments from the images. This step can be rather trivial or exceedingly complicated, depending on the objects that are to be studied, the performance we need to achieve and the imaging modalities and markers that are used. As a result of the segmentation step, we obtain a partition of the image plane E into disjoint regions, one of which representing the background, and all others different objects. We denote the set of object pixels $\zeta = \bigcup_{i=1..N_S} S_i$, with S_i the connected components of ζ and thus corresponding to the individual objects of interest (typically cells or nuclei).

There are different levels of phenotyping (see Figure 2.1): at the cellular level, we aim at quantifying the phenotype of individual cells. At the population level, we want to quantify (or profile) the phenotype of a cellular population, potentially by aggregating the single cell descriptions. Population profiles are then further analyzed to identify experimental conditions that deviate from the negative controls (**hit detection**) and eventually to identify groups of experimental conditions with similar population phenotypes.

2.1.1 Phenotype description by univariate feature distributions

In the conceptually easiest case, there is one feature that fully describes the biological process under study. Obvious examples of such features include cell size or overall fluorescence intensity, e.g. in the DAPI channel as a proxy for the cell cycle phase (Carpenter et al. 2006). Furthermore,

[1] We do not consider live cell imaging experiments in this section and refer the reader to section 2.2

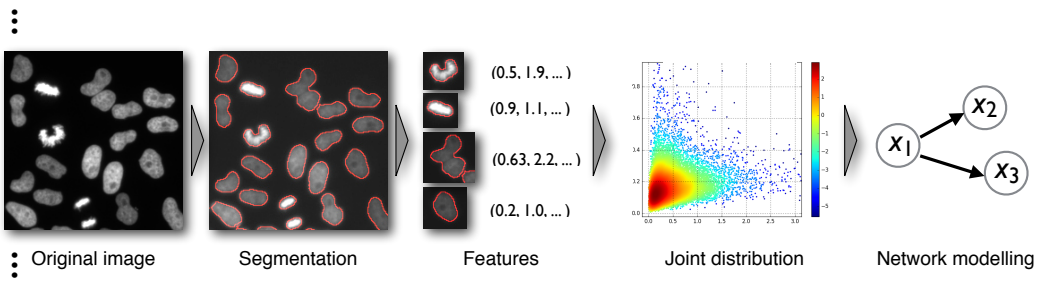

Figure 2.3: Phenotype description by a multivariate feature distribution

presence of fluorescent reporters can be measured by intensity or ratio measurements at specific locations in the cell, and thus inform on more complex phenotypes such as intra-cellular transport (J. C. Simpson et al. 2012) (in this case the amount of reporter proteins accumulating in the cell membrane is an indicator of the correct functioning of the transport mechanism). The feature might also be more complex than just an intensity, ratio or size measurement. DNA damage for instance can be quantified by detection and counting of spots from a γ-H2AX marker that highlights individual breaks in DNA (Ivashkevich et al. 2012; Garcia-Canton et al. 2013) or by using texture features allowing to omit the detection step (J. Boyd et al. 2018).

In all these cases, the analysis at the single cell level sums up to determining this one feature for each cell and the population phenotype is then described by a univariate feature distribution (see figure 2.2). Further processing steps include spatial normalization to remove systematic bias due to the location of the experiment inside the well plate, batch normalization to remove any systematic inter-plate variation and statistical testing in order to identify those genes that differ significantly from the control conditions (Bray et al. 2012; S. Singh et al. 2014).

While limited in many aspects, it is the simplicity in the description of the phenotype by a simple feature that makes the subsequent analysis steps more straightforward. This also includes a well-known parameter to judge the "quality" of a screen (called Z'-factor) which evaluates the difference between feature distributions of negative and positive controls[2]. For these reasons and despite the limitations discussed below, this modeling strategy is still very popular, in particular in pharmaceutical companies.

2.1.2 Phenotype description by multivariate feature distributions

In most cases, there is more than one relevant feature that can be measured for a cell. Indeed, even if it is possible to quantitatively describe the biological process under study with just one feature (as described in section 2.1.1), it is usually interesting to relate this feature to other features measured for the same cells. In this case, each object i is represented by a vector of P feature values $x \in \mathbb{R}^P$. These features can be calculated from the set of pixels S_i in which case they are called *shape features* or on the corresponding image values $I(x)$ with $x \in S_i$, in which case they are called *intensity* or *texture features*.

A population of cells is thus modeled by a joint distribution of feature values (see figure 2.3). Further analysis of these multidimensional distributions can take different forms, depending on whether the features are biologically interpretable or general features:

[2] A negative control is an experiment from which we expect that there is no effect, while a positive control is an experiment from which we expect a strong effect.

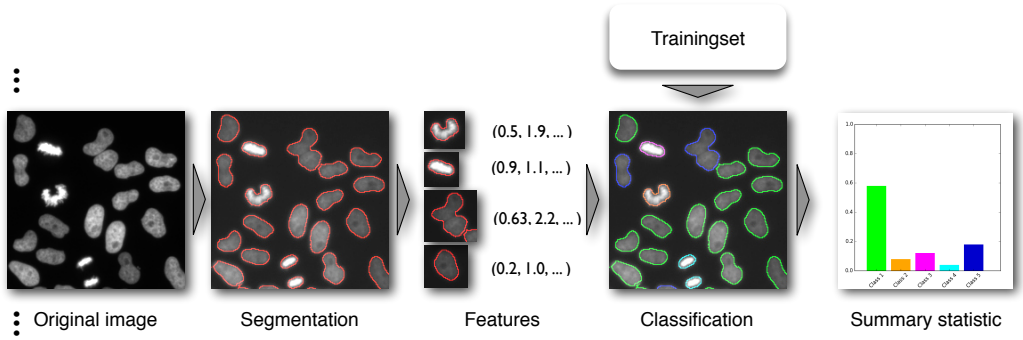

Figure 2.4: Phenotype description by single cell classification

- **General features:** One simple approach consists in describing the population phenotype by a vector of average feature values (Adams et al. 2006), other statistical descriptors of univariate distributions (Collinet et al. 2010) or test statistics resulting from a comparison to the control condition (Perlman et al. 2004). These and similar strategies disregard potential connections between features and thus model the multivariate feature distribution of P features as P independent feature distributions. Instead of using average values of the raw features as phenotypic profiles, it is also possible to project the features to a lower-dimensional space, e.g. by Principal Component Analysis or Factor Analysis (Ljosa et al. 2013).

- **Biologically interpretable features:** It is also possible to model the joint distribution of features by graphical models. This is particularly interesting if the features are biologically interpretable. Such models thus allow to investigate potential dependencies between different cellular properties and the modifications of these dependencies upon perturbation (Graml et al. 2014) or the investigation of causal relationships (J. Boyd et al. 2018) between the biologically interpretable features, allowing thus to distinguish between primary and secondary phenotypes or to remove the effect of confounders.

It is important to note that if features are not interpretable by themselves, biological interpretability needs to be injected at a later stage of the workflow. This is typically done in the last step in a *guilt by association* approach (see Figure 2.1): similar phenotypic profiles are grouped together, and the biological meaning of such a group with similar population phenotype is inferred from a priori biological knowledge about one or several of its members.

2.1.3 Phenotype description by single cell classification

In most cases, the number of raw image features that allow for a biological interpretation is relatively low. As a consequence, it is likely that many aspects of a cellular phenotype are typically not accounted for, when describing each cell by such a low-dimensional, yet interpretable feature vector. Furthermore, there are visually striking features, such as texture or morphology, that do not easily translate into a single biologically meaningful image feature. On the other hand, it is possible to extract a high dimensional feature vector that is likely to describe most aspects of the appearance of a cell under the microscope. But such a profile - albeit a good and potentially comprehensive description - is difficult to interpret. In many cases, we wish to interpret the result of our analysis not only in a final map of condition clusters, as shown in Figure 2.1, but actually at a single cell level. Moreover, the description of a population profile by a joint feature distribution in a high-dimensional space is not necessarily very practical for

further processing.

In such a case, it makes sense to use supervised or unsupervised learning in order to assign to each segmented object one morphological class (see figure 2.4):

$$\begin{aligned} f : \mathbb{R}^P &\rightarrow \mathbb{Y} = \{y^{(1)}, y^{(2)}, \dots\ y^{(K)}\} \\ x &\rightarrow y = f(x) \end{aligned} \tag{2.2}$$

f is thus the function that predicts a phenotypic class given a feature vector. This function is learned from the data. We distinguish two important settings:

- **Supervised Learning.** In this case f is inferred from a set of annotated samples called the training set $T = \{(x_i, y_i)\}_{i=1..N}$. The advantage of supervised learning is that we can guarantee that the output is biologically meaningful, as we predefine the classes according to our a priori knowledge on the biological system. The drawback is that we might not know all the phenotypic classes in a large-scale data set, and for this reason novelty detection (at the single cell level) is not possible in this setting. Methods used in this context include Support Vector Machines (SVM) (Walter et al. 2008; Walter et al. 2010a), AdaBoost, Random Forests (Dao et al. 2016) and — more recently — Deep Learning (Dürr et al. 2016; Pawlowski et al. 2016; Sommer et al. 2017). In contrast to other classification methods, Deep Learning does not rely on hand-crafted features, i.e. the representation vector x is also learned from the training data.

- **Unsupervised Learning.** In this case, there are no annotations, and the classes are derived from the data. The advantage is that this analysis is to a certain degree unbiased and allows for the detection of new phenotypes (Yin et al. 2008). In addition to being computationally challenging for large amounts of data, the drawback of unsupervised learning is that the biological sense is not guaranteed: two cells might be visually different but biologically very close and vice versa. In contrast to supervised learning, there is no way of imposing "biological meaningfulness". As unsupervised learning does often not give the desired results, one often seeks a compromise between supervised and unsupervised methods (Sommer et al. 2017).

The population of cells is thus described by a vector of classification results or a summary statistic thereof, typically the percentages of cells in all classes. It is also possible to concatenate classification results and descriptors of the feature description (Fuchs et al. 2010). And finally, depending on the biological process under study, the spatial distribution of cells of different types of states might also be relevant (Snijder et al. 2009).

2.1.4 Omitting the cellular level: population phenotyping without cell segmentation

Finally, there is also the option to classify images directly without segmenting and analyzing individual cells. This can be done by either calculating large feature vectors (Orlov et al. 2008; Uhlmann et al. 2016) with typically $P > 1000$ and applying supervised or unsupervised learning algorithms on these feature vectors. Deep Learning is also well suited to this kind of approach (Godinez et al. 2017). Alternatively, images can be partitioned into nested superpixel representations and unsupervised learning and projection techniques can be applied to find suitable low dimensional representations of images with the benefit of being independent from the number of cells (Rajaram et al. 2012).

The benefit of these methods is that image segmentation is omitted altogether. Indeed, image segmentation often is — depending on the fluorescent markers used — the bottleneck of the entire analysis workflow, as the adaptation of existing algorithms and their validation is often time-consuming and cumbersome. The price to pay is that the single cell resolution gets lost, and one of the most formidable advantages of using images in the first place, i.e. the access to the single cell level and the investigation of phenotypic heterogeneity, needs to be sacrificed. As a consequence, biological interpretation is only possible at the very last step of the analysis workflow (*guilt by association*, see Figure 2.1). This therefore removes any intermediate validation step that might be time-consuming, but which are also useful in terms of algorithmic design (modularity).

This strategy is particularly promising in the case of drug screens: in this case, ground truth is available at the level of the experiment and often little is known about the expected cellular phenotypes. As stated in (Rajaram et al. 2012), this approach is also not incompatible with a more detailed cellular analysis: indeed, it can be used for a coarse primary analysis, complemented by a more detailed secondary analysis at the cellular level.

2.2 Morphological phenotyping in action: detecting mitotic phenotypes in a genome-wide screen

I started to work in the field of High Content Screening in 2006, when I joined the group of Jan Ellenberg at the EMBL in Heidelberg. The project I was working on between 2006 and 2010 was the Mitocheck-project, where we aimed at identifying all human genes required for cell division. It was the first RNAi screen by live cell imaging in human cells, and the data generated during this screen was unprecedented in terms of size and complexity at that time. Many of the papers I have cited in the previous section, were not yet published at that time, and both method and software tools to analyze such large-scale screening data were scarce. In the following section, I will describe the methods and tools we have developed in the context of this project. Even though this one project initially triggered most of our developments, they have been refined, further developed and applied in the context of several targeted live cell imaging screens at small or medium scale (often referred to as "secondary screens"). Finally, I would like to mention that these developments were the fruit of a joint effort of the entire team I was part of. The input of experimental biologists was instrumental to the success of the entire project, but also the design of the algorithms. In terms of pure methodological development, I would like to highlight the fruitful collaboration with Michael Held, with whom I shared both ideas and code that ultimately was the basis of our common open-source software CellCognition (`https://www.cellcognition-project.org`).

2.2.1 A genome-wide RNAi screen for the identification of genes required for cell division

Since entire genomes of many species have been sequenced, including the human genome (International Human Genome Sequencing Consortium 2001; International Human Genome Sequencing Consortium 2004), one of the major tasks today is to understand how biological function arises from genomic information. This field of research is usually referred to as **Functional Genomics**.

Loss-of-function screens are important tools in this field: individual genes are down-regulated

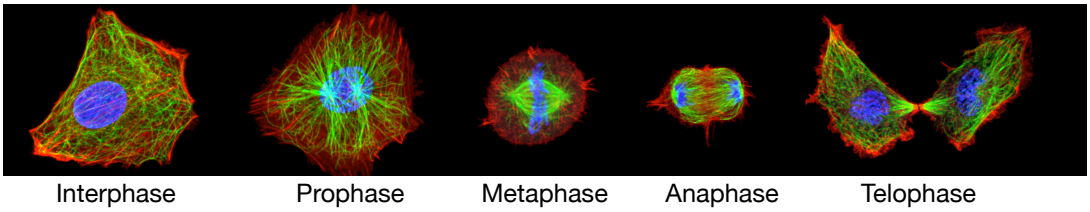

Figure 2.5: Mitotic phases (Blue: DNA, Red: Actin, Green: Microtubuli). 5 mitotic phases are shown: Interphase (corresponding to G_1, S and G_2), Prophase (caracterized by nuclear envelope breakdown and chromosome condensation), Metaphase (microtubuli form the mitotic spindle and chromosomes are aligned at the metaphase plate), Anaphase (chromosome sets are distributed to the two daughter cells), Telophase (chromosome decondense, the nuclear envelope reforms and cytokinesis concludes cell division).

or knocked out and the phenotype in response to the loss of the gene product is observed by microscopy. There are several techniques that can be used to reduce the expression of a gene. RNA interference (RNAi) for instance is a natural process to regulate gene expression, which can be hijacked in order to downregulate (i.e. reduce the expression level of) a specific target gene. From the imaged phenotype, it is often possible to derive hypotheses on the function of the gene. The specificity of this hypothesis depends on the choice of the imaging method and the employed marker: if for instance, we observe cell death without any other modification, we can only conclude that the gene is essential, i.e. the gene product is required for survival. If we observe a characteristic shape modification, we can hypothesize on a more specific role in the regulation of cell shape.

In our project, we wanted to identify the genes required for mitosis (cell division). Cell division involves a dramatic rearrangement of virtually all cellular compartments, but the different mitotic phases can actually be inferred from the chromosome configurations alone. Cell division is a relatively rare event. At any given time point, only 5% of the cells are undergoing mitosis. In addition, it is a highly dynamical process. We therefore reasoned that in order to capture the phenotypic dynamics and to collect data relevant for mitosis from most cells inside the cellular population, we would need to acquire live cell imaging data for each knock-down experiment. Taken together, our approach to identify all human genes required for cell division was to perform a genome-wide RNAi screen by live cell imaging in HeLa[3] cells stably expressing the chromosome marker H2B-GFP[4]. With at least 6 experiments for each of the $\sim$ 20000 protein coding genes and the necessary control experiments, we finally acquired a data set of more than 190000 videos of 48 hours (time-lapse of 30 minutes).

2.2.2 Morphological Phenotyping and detection of mitotic phases

As described in section 2.1.3, morphologies cannot be easily represented by a low number of interpretable features; on the other hand, it is usually relatively easy to provide examples for different morphological categories. Supervised Learning is therefore the method of choice, when it comes to morphological profiling. This is particularly true for mitosis, as the changes in morphology are rather dramatic and the different mitotic phases are clearly defined and distinguishable (see Figure 2.5). Our strategy was therefore to first segment nuclei and to extract a

[3]HeLa cells: immortalized cell line that is widely used in biological research. The name comes from Henrietta Lacks, from whom the original cells were taken.

[4]H2B-GFP refers to a fused protein consisting in the histone H2B and the Green Fluorescent Protein (GFP). The histone H2B is part of the nucleosomes that help organizing, packaging and maintaining chromosomes. The amount of H2B is supposed to be roughly proportional to the amount of DNA. Stable expression means that the cells are genetically modified in order to express the fused protein H2B-GFP.

large number of features for each segmented nucleus describing both shape and texture. Furthermore, we defined a set of classes, annotated examples for each of these classes and trained a Support Vector Machine in order to classify unseen nuclei into one of these predefined classes.

Segmentation While segmentation can often be the bottleneck in image analysis, it was reasonably straightforward in this project, where nuclei appear as bright, well-contrasted and rarely overlapping objects on a dark background. They can be reasonably well detected with standard approaches (prefiltering, background subtraction and thresholding). The most frequent problem of this technique is that sometimes close objects are segmented as one single object. One traditional way of separating close roundish objects is to apply the Watershed transform on the inverse distance transform of the initial segmentation result (Beucher et al. 1979). This strategy however corresponds to a strong shape prior: objects must not have prominent notches. If they have they are split into several objects. As this genome-wide screen contained many and actually unpredictable morphologies, we went for a more conservative strategy: we used Toggle Mappings (Meyer et al. 1989; Soille 2003) to prefilter images and to prevent close, but still separated objects from being segmented together (Walter et al. 2010a). In this case, we would accept that overlapping objects are segmented together, simply because in this context, we have no way of deciding whether they are separated in reality and just segmented together or truly connected, which can happen for instance due to segregation problems after division. We will come back to the segmentation of nuclei under more difficult conditions in chapter 5.

Object Features In order to describe shape, intensity and texture of objects, we used several families of descriptors, including basic shape and intensity features (e.g. size, elongation, circularity, etc.), classical texture features (Haralick features (Haralick et al. 1973), moment-based features (Reeve et al. 1992)), but also less widely used features (statistical geometric features (Walker et al. 1996), morphological granulometries (Serra 1983)) and newly defined features based on the convex hull and on morphological dynamics of the distance map. Altogether, we encoded each individual cell by a feature vector $x \in \mathbb{R}^P$ with $P = 190$. Most of these features had no evident biological sense.

Class definition and generation of a training set The next step was to apply supervised learning to assign a class label to each of the nuclei. For this, we needed (1) a predefined set of classes and (2) a set of manually annotated examples for each of these classes.

1. **Class definition:** When defining morphological classes, one would first be guided by the literature and try to identify morphologies that have been previously reported. In our case, this was reasonably straightforward: nuclear morphologies corresponding to the different mitotic phases are well known (Interphase, Prophase, Prometaphase, Metaphase, Anaphase and Telophase, see Figure 2.5). However, a compromise needs to be found between what is technically possible and what is essential for the biological interpretation. For instance, depending on the spatial resolution of the microscope, it can be very difficult to detect early prophase, as the change between interphase and prophase concerns subtle changes which are not clearly visible at low resolution. Second, there might be morphologies resulting from perturbation experiments that we do not know a priori. This means, that we need to use unsupervised learning (clustering or novelty detection) in order to identify these categories, which is problematic for several reasons as discussed in section 2.1.3.

2. **Finding examples for each class:** When classes are defined, we need to annotate

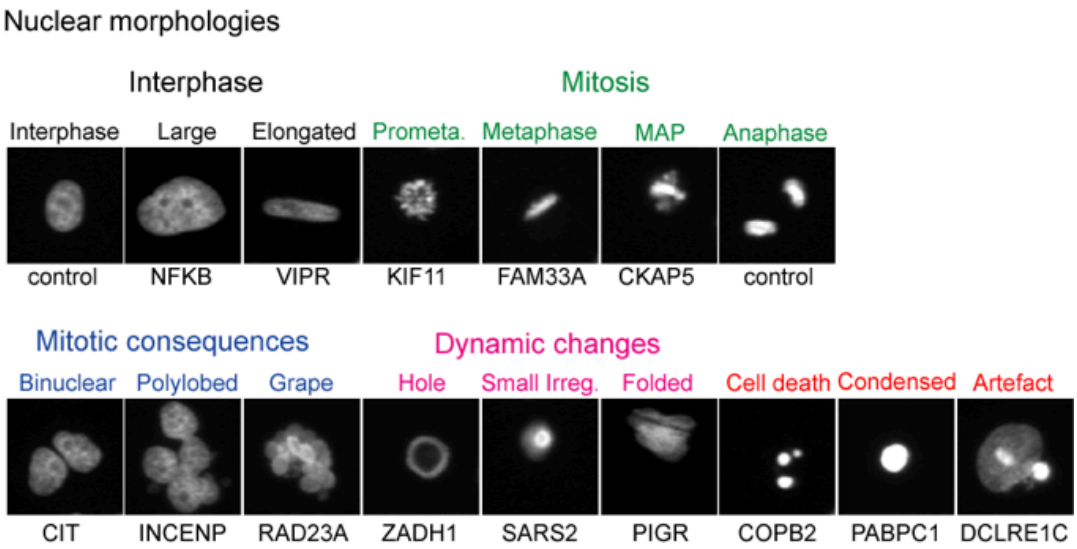

Figure 2.6: Classes of nuclear morphologies from the Mitocheck screen (Neumann et al. 2010). On top of each image: name of the morphological class. Below each image: the name of the gene, whose down-regulation can lead to the corresponding morphology (images were taken from the corresponding RNAi-experiment). Color indicates broader categories: black indicates variations of single interphase nuclei, green indicates normal and abnormal mitotic phases, blue indicates interphase morphologies consisting in several nuclei as a consequence of a failure in division, magenta indicates nuclei with severe, but non-mitotic phenotypes, red indicates variants of cell death.

samples that fall into these categories. While this might seem straightforward for some categories, there can be a number of practical problems: (1) there can be borderline cases that are hard to annotate, (2) rare morphologies can be tricky and time-consuming to find in large scale data sets, (3) it is difficult to evaluate how many samples need to be annotated to cover the morphological variability inside a class.

These two steps were actually the bottleneck of the computational analysis, and even today, 14 years after the start of my involvement in the project, there have been few publications actually targeting this problem.

Classification For classification, we used Support Vector Machines with an RBF kernel, where the parameters are found with grid search and the reported performance determined with nested cross validation. The reasons for using SVM were twofold: first, it was one of the most powerful methods for classification at that time, second it compared favorably to other methods tested and third, the training does not need any particular care by the user, as both training and parameter tuning are well defined procedures. We also used Recursive Feature Elimination (Guyon et al. 2003) to reduce the set of features, but finally, there was no clear advantage of using a smaller set of features: the accuracy did not improve with smaller feature sets and in terms of computation time, there was also little to be gained, as computation time depends rather on the number of feature families than the number of actual features (e.g. if a co-occurrence matrix is already calculated, it does not matter whether the full set of Haralick features is calculated or just a subset). It is important to emphasize, that the actual problem of applying supervised learning to computational phenotyping is not so much the classification itself. The major bottleneck was clearly the definition of the classes and the generation of the training set. It must be also noted that in practice, there is often a time-consuming back-and-forth between class definition, annotation and classification. Altogether, we trained a classifier to detect the morphologies

illustrated in Figure 2.6 with an accuracy of 86% (Walter et al. 2010a; Neumann et al. 2010). For smaller screens with less morphological variability and in particular fewer morphological classes, the same method reaches an accuracy of > 95% for the detection of mitotic phases (Held et al. 2010). Classification of chromosome figures was the key component of the analysis workflow applied on the mitosis screen introduced in section 2.2.1, but was complemented by an analysis of the dynamic behavior detailed in section 3.1.

2.3 Screening multiple cell lines

As explained in section 1.2, HCS is not limited to applications in functional genomics, but also provides us with a powerful tool for identifying potential drugs effective against a particular disease (Perlman et al. 2004). In drug screens, we test a panel of drugs against a cell line that serves as a model for the studied disease in order to discover new drugs active against the disease and to get insights into the mechanism of action (MOA) of the drugs by relating the observed phenotypes to the perturbed pathways.

However, as a proxy for diseased cells, a single cell line cannot be thought of as a perfect model. Many diseases feature a significant molecular heterogeneity. Consequently, a drug may be effective against one molecular subtype of a disease, but less so against another. To characterize drugs with respect to their effect not only on one cell line but on a consensus of several is therefore a promising strategy to streamline the drug discovery process. Nevertheless, this is not an easy task in morphological screening, as different cell lines usually have distinct archetypal morphologies even prior to perturbation. It is therefore conceptually difficult to characterize and compare drug effects across cell lines.

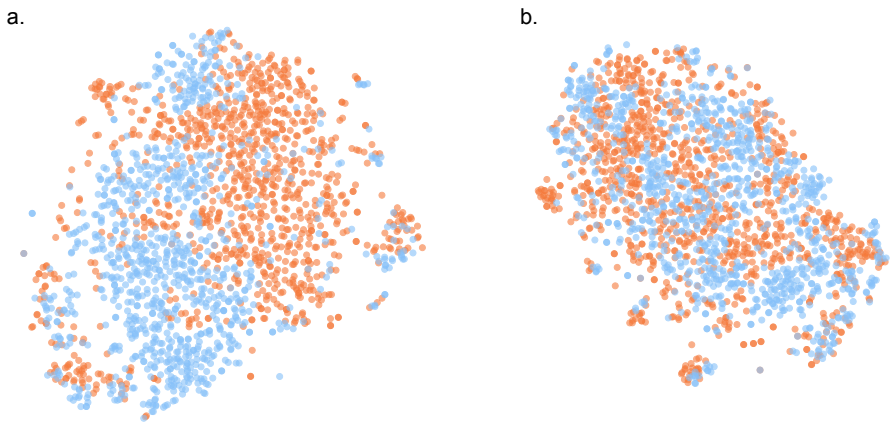

Figure 2.7: t-SNE embeddings of encodings from autoencoder (left) and domain-adversarial autoencoder (right), with cell lines distinguished by colour, and mean silhouette scores of 0.11 and 0.01 respectively.

In order to address this issue, my team has benchmarked a large panel of traditional and deep learning methods for MOA prediction in triple negative breast cancer (TNBC) cell lines and developed methods based on domain adaptation by adversary training in order to compare drug effects in two different cell lines (J. C. Boyd et al. 2019). The idea is to find a representation that allows to predict drug MOA (e.g. the molecular target or the targeted biological process) and at the same time does not allow to predict the cell line (the domain). In other words, the network parameters are optimized such as to minimize the loss for MOA prediction, and to maximize the loss for cell line prediction. This can be done elegantly by reversing the gradient layer. The

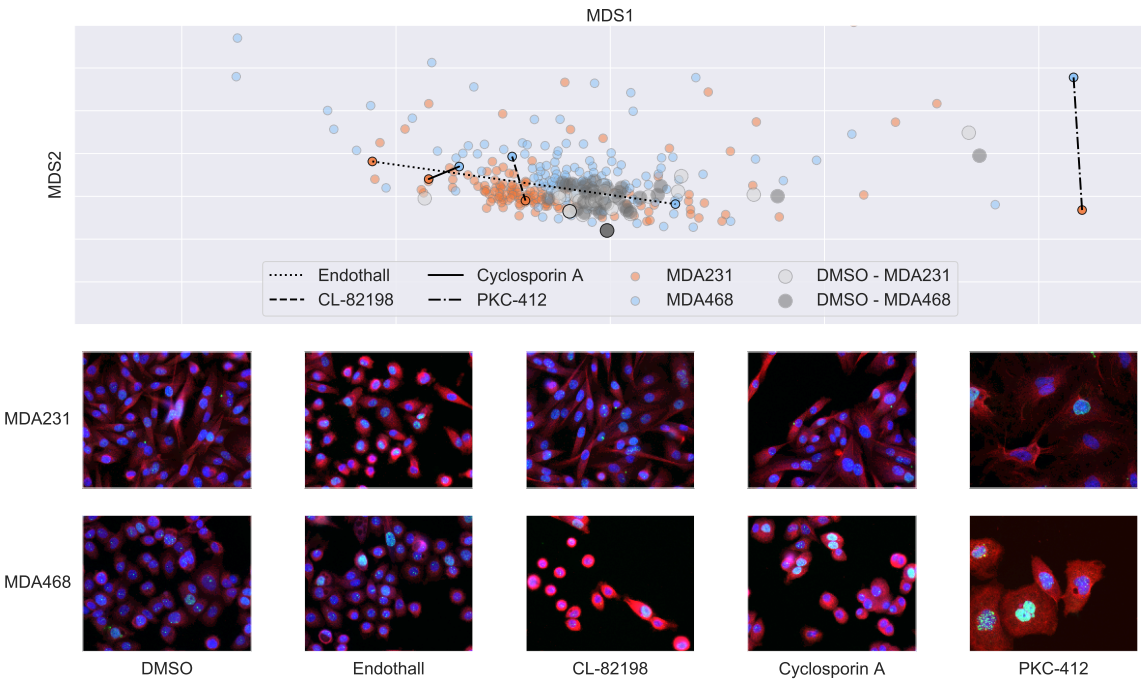

Figure 2.8: MDS embedding of drug effect profiles for MDA231 and MDA468 cell lines. Detection of differential drug effects between cell lines with examples for each category below (MDA231 top, MDA468 bottom). From left to right: no drug effect in either cell line (negative control); drug effect in MDA231 cell line only; drug effect in MDA468 cell line only; similar drug effects in both cell lines; differentiated drug effects in both cell lines. Shown are example images, blue: DAPI, red: microtubules, green: DSB.

effect of domain adaptation on cellular features is illustrated in figure 2.7.

Application of these methods to a pilot screen allowed us to categorize drugs into four different groups: (1) drugs that act only on one of the cell lines, (2) drugs that act on both cell lines in a similar way, (3) drugs that act on both cell lines, but differently, (4) inactive drugs. MDS projection of the domain invariant features and examples are illustrated in figure 2.8.

While only applied to a small pilot screen, I believe that there is a lot of potential in the developed method: multi-cell-line-screens are becoming increasingly popular, as they represent an excellent environment to develop methods for precision medicine Costello et al. 2014; Eduati et al. 2015. While these toxicogenetic data sets were so far limited to single numbered outputs, it is obvious that a richer readout, such as images, can provide a much better estimation of drug effect similarities, which in turn are key for models aiming at predicting the effect of a drug to a cell-line from its genomic and transcriptomic descriptions.

2.4 Conclusion and perspectives

2.4.1 Supervision and Publications

Section 2.2 was dedicated to the classification of cell morphologies, applied to a genome-wide RNAi screen by live cell imaging, described in section 2.2.1. The analysis workflows I have developed in this context therefore contained both the automatic recognition of cellular morphologies and the analysis of the time-resolved data, which will be discussed in detail in the next chapter.

For this reason, some of my publications from this time deal with both aspects (Erfle et al. 2007; Walter et al. 2008; Walter et al. 2010a; Neumann et al. 2010; Walter et al. 2010b; Held et al. 2010; Terjung et al. 2010; Conrad et al. 2011) and are therefore listed again in section 3.3).

In collaboration with the Carazo group at the Gurdon institute, Camebridge, I also contributed to a genetic screen in yeast, where we investigated the interplay between several biological processes and modeled the interdependencies between features by a Bayesian network (Graml et al. 2014). I have been also involved in two competitions on multi-cell-line drug screens without images, where the objective was to predict the drug effect, as measured by a single value, from the drug and the genetic features of the cell-line, leading to three publications (Costello et al. 2014; Eduati et al. 2015; Bernard et al. 2017). Our experience in these two challenges was that a single number can be hardly sufficient to represent the complex drug-induced phenotypes. This triggered the idea of performing multi-cell-line HCS, in particular for drug screening, where the cell lines could represent molecularly defined subtypes of a disease. This work was done by Joseph Boyd, a PhD student in my group, who will defend his thesis in 2020. So far, this project has lead to 2 publications (J. Boyd et al. 2018; J. C. Boyd et al. 2019).

2.4.2 Perspectives

Computational analysis of morphological phenotypes is a well developed field. The standard method consisting in segmentation, feature extraction and classification is versatile and powerful enough for many real-world applications. For instance, classification of mitotic phenotypes reaches accuracies of more than 95% with these methods (Held et al. 2010). While we could potentially improve on that with newer methods, in particular deep learning, the usefulness of such developments would certainly be questionable. For this reason, it is probably more interesting to target questions that we cannot answer with traditional methods or for which the current results are either poor or in some other way not effective.

Segmentation of cells and cellular structures by fully convolutional neural networks One of the bottlenecks of the workflows discussed in section 2.1 is the segmentation of cells and cellular components that can be difficult. Segmentation of nuclei is often a simple task, but can get difficult depending on the cell line and the imaging modality (see for instance the detection of nuclei in stained tissues in section 5.2). Segmentation of the cytoplasmic region can still be complicated in some cases, and in general segmentation tasks are often time consuming and complex, in particular for label-free microscopy, such as bright-field or phase contrast microscopy. For this reason, novel and generally applicable segmentation methods are currently developed and can have a great impact on HCS, as we can extend the applicability of HCS to cell lines and imaging conditions that are far from being ideal.

Deep Learning with experimentally generated ground truth Beyond image segmentation, deep learning is also the state of the art method for image classification. Two major use cases:

1. Classification of entire cell populations, i.e. classification at the experiment level, can be interesting for drug screening (Godinez et al. 2017), but in many cases, we still wish to keep the cellular level: images provide us with the advantage that we can observe and quantify cellular heterogeneity; it is suboptimal to give away this interesting property.

2. Classification of single cell morphologies by convolutional neural networks is an interesting alternative to the traditional approaches, in particular for morphologies that are so far

difficult to classify. However, it is well known that neural networks - while more powerful - tend to require more annotated data than traditional classification methods (e.g. SVM, RF).

As discussed in section 1.1, it is challenging to get large annotated data sets for individual biological projects, and annotations from one project are often not easily transferable to another. Beyond the use of pretrained networks, one of the major issues is therefore how to get larger annotated data sets. One interesting option is to use smart experiment design to actually generate ground truth data at a cellular level. This can be achieved by the use of fluorescent labels for training that are indicative for certain phenotypes (such as cell division or cell death). We are currently investigating this strategy to analyze a large series of live cell phase contrast imaging experiments. The objective of this project is to study engineered T-cells that selectively attack cancer cells. The relevant cell classes can be marked with fluorescent markers, but there are several reasons which make the use of live dyes in this context impractical, such as cost and long term stability. However, they can be used in a calibration phase to derive a ground truth from which a Deep Neural Network can be trained and applied to the phase contrast data. In contrast, the experiments can be run without any fluorescent marker.

In silico labeling With in silico labeling, we refer to the prediction of fluorescence images from bright field or phase contrast images (Christiansen et al. 2018; Weigert et al. 2018; Ounkomol et al. 2018; Belthangady et al. 2019). The subtle difference to the generation of ground truth by experiment design is that we are not predicting class labels but images of a different modality. The main use case of this technique is that we can predict the position and shape of organelles from phase contrast or bright field images, with the advantage that we do not need to fluorescently label these structures. This evolution is likely to revolutionize HCS, as we can now have a panel of landmark proteins marked computationally, while we can keep the fluorescent channels for other proteins. We are going to use and further develop these techniques in the frame of localization screens introduced in chapter 4.

3 Exploring the temporal dimension: recognition of dynamic phenotypes

In this chapter, I will describe my developments for the analysis of the temporal aspects of cellular phenotypes. In section 3.1, I briefly describe our efforts to describe time-resolved morphological phenotypes, starting with clustering and modeling approaches of phenotype dynamics for the Mitocheck screen, described in section 2.2.1, followed by our work on secondary screening data, where we apply Hidden Markov Models (HMM) for error correction and in silico alignment of mitotic phenotypes. These projects were mostly developed during my time at the EMBL in close collaboration with experimental and computational researchers from EMBL. In section 3.2, I describe the developments of my team to analyze nuclear movements in the mitocheck data involving clustering of spatial trajectories. Hence, while section 3.1 is concerned with morpho-temporal analysis, section 3.2 is dedicated to spatio-temporal analysis and therefore represents a transition to the next chapter dedicated to the analysis of spatial patterns.

3.1 Analysis of morphological phenotypes over time

In the last section, we assigned to each object (nucleus) first a feature vector $x \in \mathbb{R}^P$ and to each x a class $y \in \mathbb{Y} = \{y^{(1)}, y^{(2)}, \ldots, y^{(K)}\}$ from the set of K predefined classes. We can therefore describe an initial image sequence $\{I_t\}_{t=1..T}$ by a sequence of relative count vectors $\{c_t\}_{t=1..T}$, where for each timepoint t the vector of fractions of objects in each of the K predefined classes is denoted by $c_t \in [0,1]^K$. This sequence of vectors (or equivalently, this multidimensional time series) is now our description of the time-resolved phenotype.

With these time-series, we can now find those experiments that differ from negative controls and thus identify the "hits" of the screen (Neumann et al. 2010), i.e. genes with notable differences in the classes relevant for mitosis. In addition to this, we can study the dynamics of the phenotypes. For this, we developed methods to give a concise overview over the order in which phenotypic events occur (Event Order Maps, EOM). The strategy is to find an order of events that is in agreement with most pairwise event orders from the different replicates (Walter et al. 2010a).

3.1.1 Clustering of multi-dimensional time series

Beyond pure visualization to study major causes and consequences of phenotypes at the population level, we wanted also to cluster experiments according to phenotypic similarity using the full time-resolved profiles. This involves calculating similarities between multidimensional time-series. An additional complication is that phenotypic onsets can be very different between RNAi experiments against different gene products, but have little value for functional similarity (the phenotypic onset is typically influenced by factors such as protein stability or the required level of the targeted protein). We therefore developed a dissimilarity measure that is independent of phenotypic onsets (given that the phenotypes are in the observation window, i.e. that the

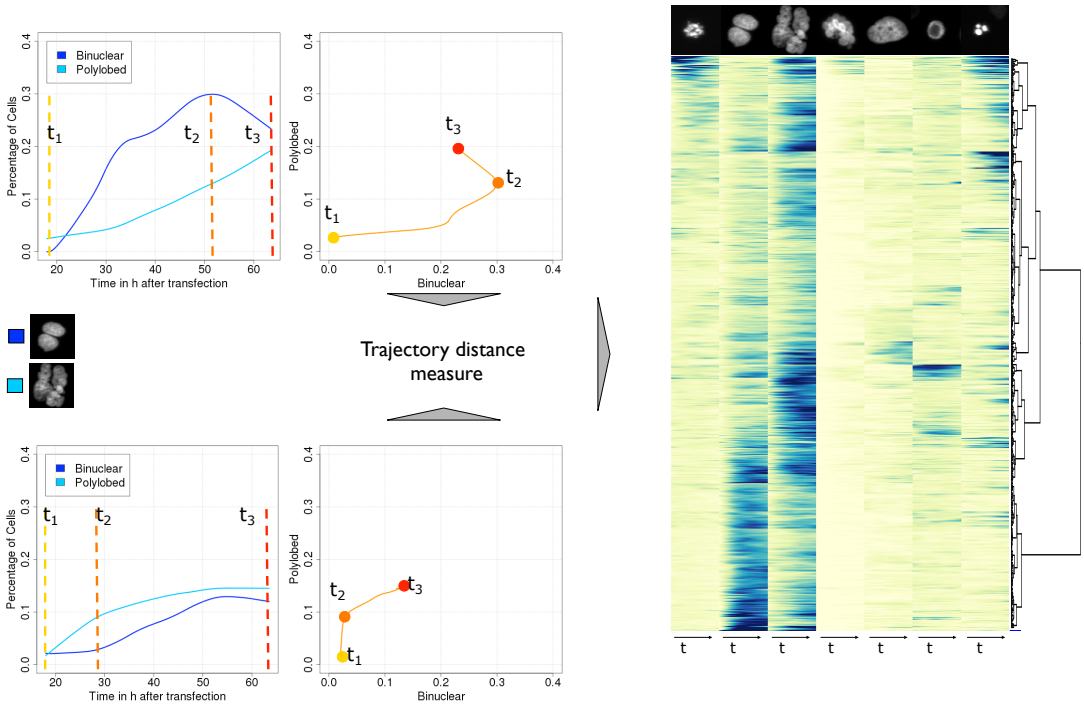

Figure 3.1: Principle of trajectory clustering for multi-dimensional time series: each set of time-series is represented by a trajectory in the phenotypic space, where time is an implicit parameter. These trajectories are represented by a set of vectors (here a pair of vectors). The dissimilarity is calculated on these trajectories. The result is a time-resolved heatmap (on the right), where every gene corresponds to one line, for every morphology, we display its percentage over time.

phenotypic onset does not occur prior to the start of the imaging). The idea is to represent the K-dimensional time series by a trajectory in a K-dimensional space, where time is an implicit parameter of the phenotypic trajectory: trajectories represent the joint evolution of the different time curves, independent from their phenotypic onset and speed of change. We argue that RNAi experiments with overall similar trajectories are likely to show a similar overall phenotype (Walter et al. 2010a). The method is illustrated in figure 3.1.

In a collaboration with Gregoire Pau and Wolfgang Huber, we investigated an alternative way of modeling the time series (Pau et al. 2013), where we focused on the transitions between different phenotypes at the population level. Based on a model of possible transitions between different phenotypic states, we modeled the time series by a system of Ordinary Differential Equations (ODE). An important aspect here is that the transitions between different morphological classes cannot be considered to be constant. Indeed, during the experiment, we can expect that the cells run out of the down-regulated proteins and consequently, the parameters of the model, and namely the transition rates are bound to change over time. While the drawback of the method is that we need an a priori model regarding existing transitions between morphologies, this method provides nonetheless a powerful tool for the description of dynamic phenotypes. We also showed that the identified model parameters can also be used for clustering (Pau et al. 2013).

3.1.2 Phenotypic trajectories at the single cell level in secondary screening experiments

All of the methods presented in this chapter so far aimed at analyzing population phenotypes over time, i.e. for each time point, the entire population of cells is analyzed and their pheno-

type summarized as absolute or relative count vectors in the respective classes. An alternative strategy is to track individual cells over time and to measure the order of phenotypic events and phenotypic state transitions at the single cell level instead of estimating them at the population level. While this is in principle attractive, we actually followed this strategy first for secondary screens with higher spatial and temporal resolution and a better understanding of the phenotypes we were facing. Indeed, tracking all individual cells in the lower-resolution genome-wide data set is not an easy task, which we endeavored only at a later stage (see section 3.2).

Secondary Screening Primary screens as the one presented in section 2.2.1 aim at providing comprehensive surveys of the molecular basis of fundamental biological processes, here cell division (Neumann et al. 2010). The result of such a screening project consists in lists of genes with putative function in these processes, and — depending on the richness of the phenotypic readout — a more detailed hypothesis on the function of the encoded proteins. In figure 3.1, we see for instance that there are different clusters corresponding to different phenotypes: there are groups of genes whose downregulation leads to an accumulation of cells in an early mitotic state, suggesting problems in the formation of a mitotic spindle, sometimes followed by cell death, sometimes by micro-nucleation (a defect in the reforming of the nuclear envelope), sometimes by segregation defects. Another group of gene knockdowns leads to an accumulation of binucleated cells, thus indicating a defect in cytokinesis. We therefore see that even though all these phenotypes are related to cell division, the nature of the failure can be very different.

In order to elucidate the role of these different groups of genes and to better understand the underlying mechanisms, one usually performs secondary screening experiments at higher spatial and temporal resolution with markers that are tailored to the hypothesis inferred from the primary screening data. Secondary Screening projects concern thus fewer genes (typically tens to hundreds), are less comprehensive and less exploratory (with the consequence that we have a stronger biological prior on the expected phenotype), and usually employ more complex imaging conditions. I have been involved in several secondary screening projects regarding mitosis (Mall et al. 2012; Hériché et al. 2014; Isokane et al. 2016).

A secondary screening example: the regulation of lamin disassembly during early mitosis In this secondary screening approach with Moritz Mall from Ian Mattaj's group at the EMBL, we were interested in elucidating the role of two kinases PKC and CDK1 to orchestrate the disassembly of nuclear lamina during early mitosis. The nuclear lamina form a dense meshwork of intermediary filaments that underpins the nuclear envelope. During envelope breakdown, the lamina disassemble and translocate either to the mitotic endoplasmic reticulum (ER) or to the cytoplasm.

In order to investigate the disassembly of the lamina and to relate it to the different phases of cell division, we used a cell line stably expressing EGFP-LAMINB1 to report on lamin organization and H2B-mCherry to track mitotic phases. By live cell imaging at high temporal resolution (2 minutes), we monitored the dynamics of lamin translocation from the nucleus to the cytoplasm (see figure 3.2.a). Concretely, we took the following steps:

1. We classified mitotic phases from the H2B-mCherry signal.

2. We tracked individual cells over time with a simple nearest neighbor tracker. For each cell n we had thus an individual time series of classification results $\{c_t^{(n)}\}_{t=1\ldots T_n}$.

3. We detected the transition from interphase to prophase and aligned cells undergoing mi-

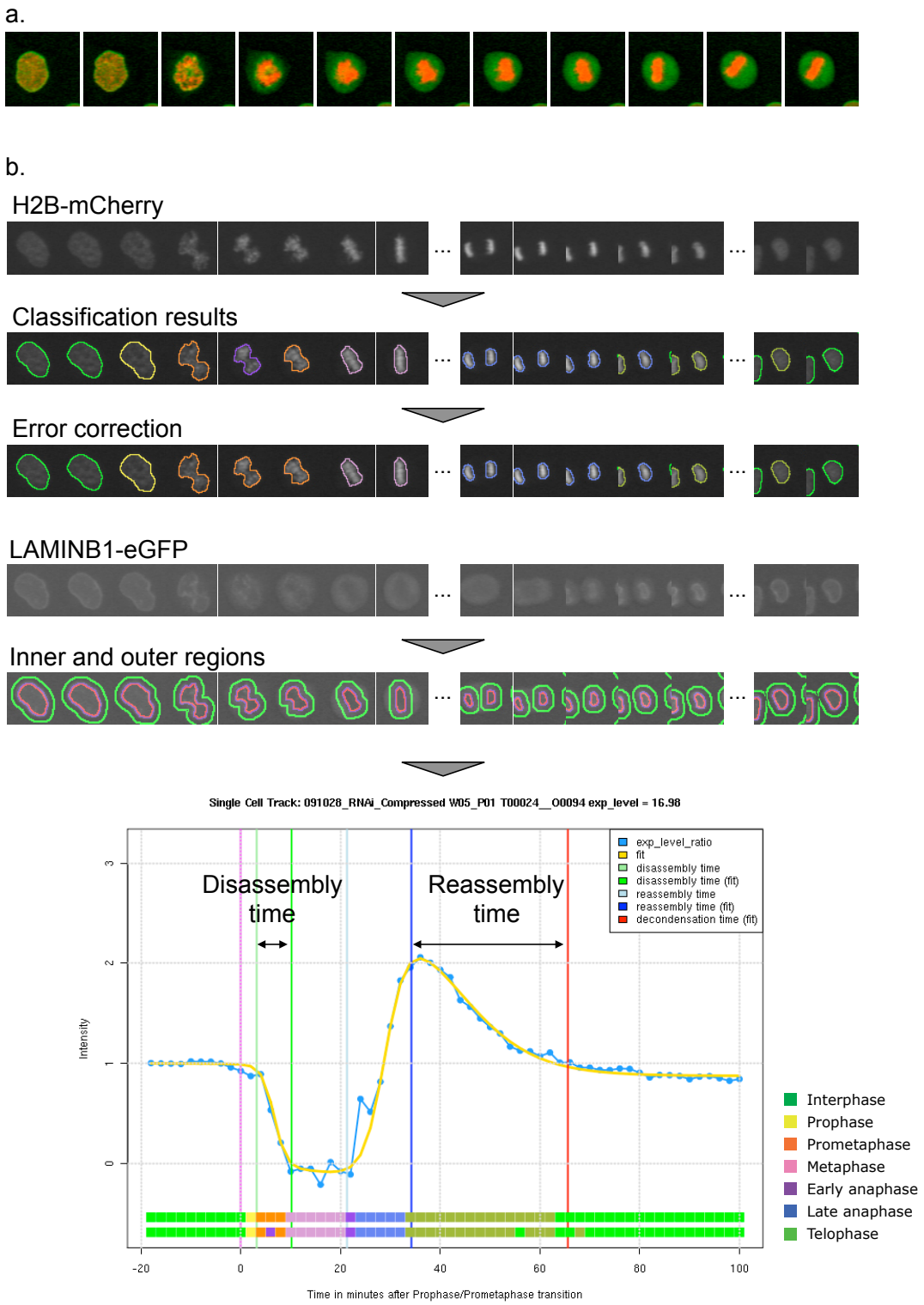

Figure 3.2: Analyzing disassembly of the nuclear lamina. a. Gallery of a single cell undergoing mitosis. Green: LAMINB1-eGFP. Red: H2B-mCherry. b. Workflow: from the H2B channel (first panel), nuclei are segmented and classified (second panel). HMM are used for correction (3rd panel). Intensities inside and outside the nucleus are measured. For each dividing cell we can measure a time series, as shown on the bottom plot.

tosis on this transition (in silico alignment).

4. We used Hidden Markov Models (HMM) in order to correct for classification errors. The confusion matrix defined the emission probabilities, some transition probabilities were set to 0 according to our prior knowledge on cell division, and the rest of the parameters was inferred by the Baum-Welch algorithm. Application of the Viterbi algorithm then provided us with the sequence $(\tilde{c}_t^{(n)})_{t=1...T_n}$ of corrected classification results.

5. We used to the nuclei segmentation in order to measure intensity in the Lamin channel inside and outside the nucleus and normalized about the value in interphase:

$$\frac{x_{in}(t) - x_{out}(t)}{x_{in}(t) - x_{out}(0)} \tag{3.1}$$

6. We inferred disassembly and reassembly time from this time-series.

The approach is illustrated in figure 3.2.b. Application of this model to a number of RNAi and chemical perturbation experiments allowed us to build a model of the control of lamina disassembly during early mitosis. On a more technical note, we see that each project might come with its own specificities and its own quantification problems, but there is still an algorithmic backbone that is similar across projects, and which was therefore used in a number of secondary screening projects I have been involved in.

3.2 Analysis of movement types

Large-scale genomic screens are rarely informative about only one biological process. For instance, the genome-wide screen described in section 2.2.1, initially designed for the identification of human genes required for cell division (Neumann et al. 2010), was in principle also informative about proliferation, survival and nuclear motility. While scoring for other morphological phenotypes was straightforward and solely required new class definitions and annotations the other elements of the workflow remaining identical, analysis of movement required different algorithmic approaches. For this, we first tracked cells over time. The resulting spatial trajectories were then mapped to an original feature space describing various movement properties. From the feature distributions we first identified those experiments that were significantly different from negative controls and then turned to unsupervised analysis in order to identify trajectory classes. We validated this approach on a simulated screen and applied it to the genome-wide screen presented in 2.2.1.

3.2.1 Tracking by learning

Cell tracking faces several challenges in videos from high content screens like Mitocheck. The algorithm has to handle apparitions, disappearances, divisions and fusions (due to occlusions or segmentation errors). In addition, it has to cope with a high phenotypic inter-cell variability, and must not rely on strong a priori assumptions on movement, as we wish to identify abnormal movement types. Therefore, we extended a non-parametric structured learning approach from (Lou et al. 2011). For this, we consider all possible matches between objects at time t and $t+1$ under movement type $e \in \{move, appear, disappear, split, merge\}$. In practice, this is represented by indicator variables $z_{i,j}^e(t) \in \{0,1\}$ which are 1 if object i at time t corresponds to

object j at time $t+1$ under movement type e and 0 otherwise. Appearances and disappearances are modeled by the use of virtual objects ($i = 0$ for appearances and $j = 0$ for disappearances); for splits and fusions we need to consider links between objects and sets of objects (which we omit here for sake of simplicity). Next, we characterize each match by a describing feature vector composed of euclidean distance, orientation distance (angular distance of the principal axes at t and $t+1$, respectively), and the difference in texture and shape features, as defined in section 2.2.2.

The optimal object matching $\widehat{z}(t)$ comes down to bi-partite graph matching: it is solved by maximizing a likelihood function L which depends on the weights w and the match features $f^e_{i,j}$, subject to the constraint that all objects are matched in both frames (cf equation 3.2).

$$\widehat{z}(t) = \underset{z(t)}{\arg\max}\ \mathbf{L}(z(t); w) \tag{3.2}$$

where

$$\begin{aligned}\mathbf{L}(z(t); w) = &\sum_{\substack{e \in \mathbf{E} \\ Obj_{i,t} \\ Obj_{j,t+1}}} < w^e, f^e_{i,j} > z^e_{i,j}(t) \\ &s.t.\ \ \forall i \sum_{\substack{e \\ Obj_{j,t+1}}} z^e_{i,j}(t) = 1 \\ &and\ \ \forall j \sum_{\substack{e \\ Obj_{i,t}}} z^e_{i,j}(t) = 1\end{aligned}$$

The weights w are learned by a support vector machine using annotated trajectories, following the formulation of (Lou et al. 2011). The likelihood maximization is an integer linear programming (ILP) problem that can be solved by IBM Cplex.

As we showed in (Schoenauer Sebag et al. 2015b), this algorithm compared favorably to constrained nearest neighbor tracking and a widely used method where tracking is formulated as a linear assignment problem (Jaqaman et al. 2008). It must be noted however, that for simple movements at low to medium cell density, there is barely a difference between the different tracking methods (Schoenauer Sebag et al. 2015b). Indeed, the main advantage of this method is that complex object features are used to establish correspondences, and this is only relevant in cases where the Euclidean distance is not informative enough (e.g. in the case of splits, fast movements and dense populations).

3.2.2 Trajectory features

Tracking allows us to represent each experiment by a set of N spatial trajectories $\{\Gamma^{(n)} \mid n = 1..N\}$, each trajectory of length T corresponding to the sequence of cell center positions $\Gamma^{(n)} = \{u^{(n)}_t\}_{t=1...T}$. In analogy to the approach described in section 2.2.2, we characterize each trajectory by a set of features $x_n \in \mathbb{R}^P$, describing different movement properties. We have used a set of 14 features, consisting in basic movement features, features based on moments of displacement (Sbalzarini et al. 2005) and features we have designed according to different aspects we wanted to be represented. These features are described in detail in annex B.

3.2.3 Identification of movement types by unsupervised learning

Given the trajectories $\Gamma^{(n)}$ and their feature descriptions x_n, the next task is to classify individual trajectories into one out of several movement categories and to summarize the RNAi experiment by the percentages of trajectories falling in each category. However, unlike the approaches for morphological phenotyping, we do not know the different types of movement we can expect in this large-scale screen. We therefore went for unsupervised learning, as explained in section 2.1.

However, with roughly 20 million trajectories and an important biological variability, unsupervised learning did not prove to be successful when applied to the full set of trajectories, for a wide range of clustering techniques (k-means, Gaussian mixtures models, spectral clustering, fuzzy c-means, kernel k-means). There are several reasons for this: the number of trajectories is very large and some more powerful methods for clustering cannot cope with such a large number of samples. In addition, the data is highly imbalanced and the vast majority of trajectories are just variants of normal behavior. In particular, the imbalance makes randomly downsampling a questionable approach. Our approach was therefore a mixture of outlier detection and clustering: we first identified experiments that significantly deviated from negative controls with respect to at least one of the features and then used the pooled trajectories from this reduced set of experiments for further clustering. We validated this approach on a virtual screen, where we simulated trajectories according to a priori defined movement types and tested whether we could identify the movement categories and assign the trajectories to the correct classes (accuracy of $\sim 90\%$). Application to the genome-wide screen allowed us to identify different movement types in a screen that was not initially designed for the analysis of nuclear motility (Schoenauer Sebag et al. 2015b; Schoenauer Sebag et al. 2015a).

Regarding the methodological developments, we learned two lessons from this project:

1. For very large datasets and a serious imbalance, it is advantageous to enrich for rare cases in order to give the minority classes a chance to be represented by their own cluster. This is actually also what we do when we manually annotate datasets in supervised learning: we do not reproduce the prior probabilities in the data sets, but we tend to enrich for rare classes.

2. Validation on simulated data is an interesting option for unsupervised problems, where no ground truth is available. Ideally complemented with other types of validation, such as enrichment analysis of the results, cluster stability and validation on annotated data (if available), it is in a truly unsupervised case often the only option we have.

3.3 Conclusion: Supervision and Publications

Section 3.1 was dedicated to the analysis of morphological phenotypes over time. This work was done at the EMBL in close collaboration with experimental scientists at the EMBL. In (Walter et al. 2008; Walter et al. 2010a), I published the entire workflow including the recognition of nuclear phenotypes (described in section 2.2) and the clustering of population time-series. In (Neumann et al. 2010) we applied the method to a genome-wide screen on cell division. In (Tegha-Dunghu et al. 2008; Tegha-Dunghu et al. 2014), we applied the methods for computational phenotyping on small-scale studies, in (Mall et al. 2012; Hériché et al. 2014; Isokane et al. 2016) to secondary screening data. In a collaboration with the Gerlich group, we implemented these and other

methods in the open-source software CellCognition (Held et al. 2010). In a collaboration with the Huber group, we also modeled the time-series by Ordinary Differential Equations (Pau et al. 2013). In addition to this, I also contributed to a protocol paper on experimental HCS techniques (Erfle et al. 2007) and the Micropilot project where the morphological classification was used to guide image acquisition (Conrad et al. 2011). In (Walter et al. 2010b), colleagues and I reviewed visualization tools for bioimaging data, including data from High Content Screening; (Terjung et al. 2010) was a review dedicated to technical aspects of High Content Screening.

Section 3.2 was dedicated to the analysis of spatial trajectories. The work was performed by Alice Schoenauer Sebag, a PhD student in our lab and led to two publications (Schoenauer Sebag et al. 2015b; Schoenauer Sebag et al. 2015a).

4 Localization phenotyping: spatial transcriptomics

In chapter 2, we have discussed how to computationally analyze morphological data, chapter 3 was dedicated to the analysis of time-resolved microscopy data. In this chapter, I will show our developments in view of analyzing spatial distributions of RNA molecules inside cells.

4.1 Biological context: RNA localization

Gene expression is the process by which genetic information is transformed into functional products. For this, genetic information is copied to RNAs, a process named transcription. RNA molecules are then further processed before they eventually leave the nucleus in order to either fulfill a function of their own (such as ribosomal RNA) or to serve as a blueprint for protein generation, a process named translation. In any case, the number of RNA molecules present in a cell or a cellular population can be seen as a proxy of the "activity" of that gene. Tight regulation of gene expression is essential for a gene to fulfill its basic functions, and disregulation of gene expression can lead to serious failures at the cellular, tissular and organism level. For all these reasons, the study of gene expression has been one of the major fields in genome biology for many years and has triggered the development of both experimental techniques, such as microarrays or RNA sequencing, and computational methods in Bioinformatics.

Traditionally, these studies focused on the expression level, i.e. the number of RNA molecules or a proxy thereof. More recently however, it has become apparent that it is not only the number of RNA molecules that matters, but also their localization inside cells. Indeed, RNA molecules may localize in specific regions of the cytoplasm, i.e. they distribute according to a specific localization pattern. Subcellular localization of mRNAs is thought of as playing an important role for the spatial control of gene expression; its misregulation is linked to an increasing number of diseases (Buxbaum et al. 2014; Chin et al. 2017). However, the function and mechanism of RNA localization are not yet well understood. In addition, it is likely that not all localization patterns are known so far and it is still unclear which mRNAs distribute according to which localization pattern.

These questions can be addressed by large-scale image-based assays, where individual mRNA molecules are visualized by single molecule Fluorescence in situ hybridization (smFISH). smFISH allows for the visualization of individual mRNA molecules in their native cellular environment (Raj et al. 2008; Tsanov et al. 2016). The principle of smFISH is to target mRNA with several fluorescently labeled oligonucleotides (see figure 4.1.a). Many variants of this method exist, with optimizations regarding signal to noise ratio (SNR), experimental protocol, specificity of the targeting, scalability, automatization and cost. In our project, we use single molecule inexpensive FISH (smiFISH, see figure 4.1.b), a technique that is particularly inexpensive and therefore scalable at the level of High Content Screening (Tsanov et al. 2016). Individual RNA molecules appear as small diffraction limited spots under the microscope (see figure 4.1.c).

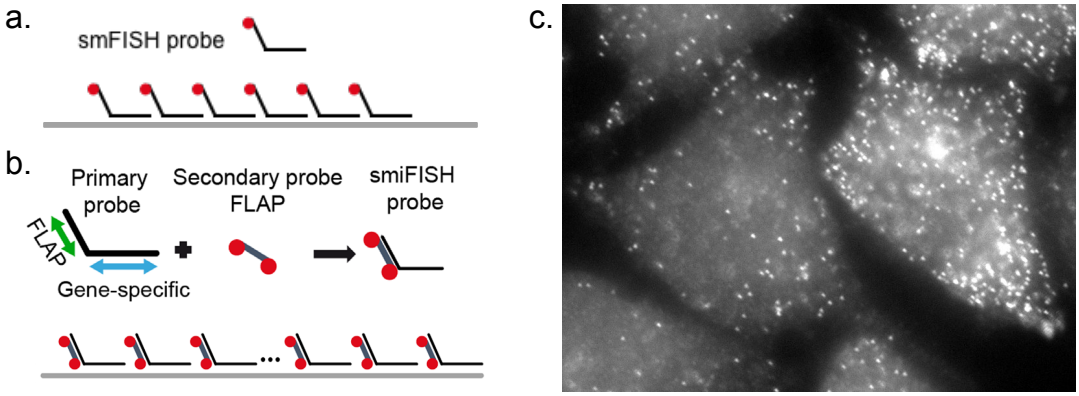

Figure 4.1: single molecule FISH - a. Standard smFISH procedure: fluorescently labeled probes are hybridized on the target RNA. b. smiFISH: gene specific unlabeled probes are hybridized to the target RNA. They contain a FLAP sequence, onto which the fluorescently labeled probes are hybridized in a second step. c. Example smFISH image.

In order to identify the RNAs[1] with non-random localization and to obtain the landscape of RNA localization patterns, we can thus perform a large-scale screen, where we visualize the mRNA molecules for one gene at a time by smFISH for a predefined set of genes (typically tens to hundreds of genes for one screen). This raises the computational challenge of identifying the different mRNA localization patterns and to assign each tested mRNA to one of them.

4.2 Computational analysis of subcellular localization patterns of mRNAs

As described in section 4.1, computational analysis can in principle be casted as a problem of unsupervised learning: given the full set of screening images, we wish to identify all localization patterns present in the data set. Once the set of localization patterns is known, we wish to assign each cell to one of them, and the computational task can then be formulated as a supervised learning problem.

Overview In figure 4.2, we see the basic workflow of the computational analysis, consisting in cell segmentation, detection of individual mRNAs, description of their spatial distribution by features, and a machine learning step, consisting in either supervised or unsupervised learning. This step might be replaced by projection to a low-dimensional space and visual inspection. Compared to the workflow shown in section 2.1.4, the main difference is (1) we detect individual RNAs prior to classification, (2) that consequently, we aim at classifying point clouds rather than general textures and (3) that in most cases, we do not have annotated data sets, as little is known about existing localization patterns.

Segmentation and detection tasks In (Samacoits et al. 2018), we have presented methods for the segmentation of the cytoplasm from a dedicated marker (CellMask) or — with less accuracy — from the background FISH signal. For this, we have developed novel techniques for

[1] In the following, "mRNA molecules" denote all instances of the same mRNA. In contrast, with "mRNAs" we refer to different mRNAs (i.e. the transcripts of different genes). Furthermore, this study is concerned with mRNA only (so whenever we talk of RNA, we actually mean mRNA).

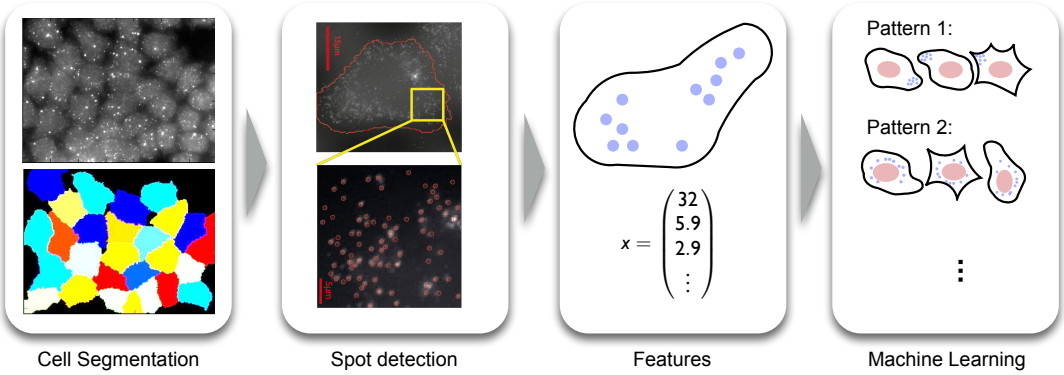

Figure 4.2: Workflow for the analysis of smFISH images. Cells are segmented, individual RNAs detected. The spatial distribution of points inside each cell can be mapped to a feature space. In the last step, machine learning allows to assign a localization pattern to each cell.

focus projection and explored both traditional techniques based on Mathematical Morphology and Deep Learning strategies for multiple instance segmentation. In addition, we also detect the nuclei from the DAPI signal with traditional methods. For the detection of spots we used traditional techniques for spot detection based on the Laplacian of Gaussian and local maxima (Mueller et al. 2013). In order to resolve agglomerations of spots, we decomposed them using mixture models. At the end of this step, we have thus for each cell the cytoplasmic region R_c, the nuclear region R_n and the locations u_i of the individual RNA molecules.

Localization features To describe the spatial distribution of molecules in the cell, we calculate a set of features, some of which coming from the literature (Battich et al. 2013), others were introduced by us (Samacoits et al. 2018). There are features that represent the distance distributions from landmarks in the cell (i.e. distances from the cytoplasmic membrane or distances from the nucleus) and others relate to the inter-point distances (mostly features based on Ripley's L-function). It is important to normalize the features such that they do not (or only mildly) depend on the point density (expression level) and the cellular morphologies, as both morphology and expression level might act as potential confounders. For more details, we refer to (Samacoits et al. 2018).

Simulation of smFISH images One of the main difficulties in the unsupervised approach illustrated in figure 4.2, is that in the absence of an annotated ground truth data set for RNA localization in smFISH data, it is virtually impossible to assess usefulness of features and performance of the clustering workflow. As mentioned in section 3.2.3, simulated ground truth data can be used to validate clustering workflows. In the case of smFISH images, this is particularly interesting, as a human might find it difficult to make annotations on 3D data, without getting influenced by confounding factors, such as point density and morphology. The impact of these covariates is amplified by the fact that we might know only few example genes for a given localization pattern. In the extreme case, we would not even know a single example gene for a hypothetical localization pattern, but we might nevertheless be interested in knowing whether this (postulated) pattern is present in the data set. All of these considerations encouraged us to build a simulation framework, where we generated point patterns from known localization rules to create large amounts of ground-truth data.

For this, we first developed an experimental workflow allowing us to generate a library of cell reference volumes with precise information on the plasma membrane and the nuclear region in

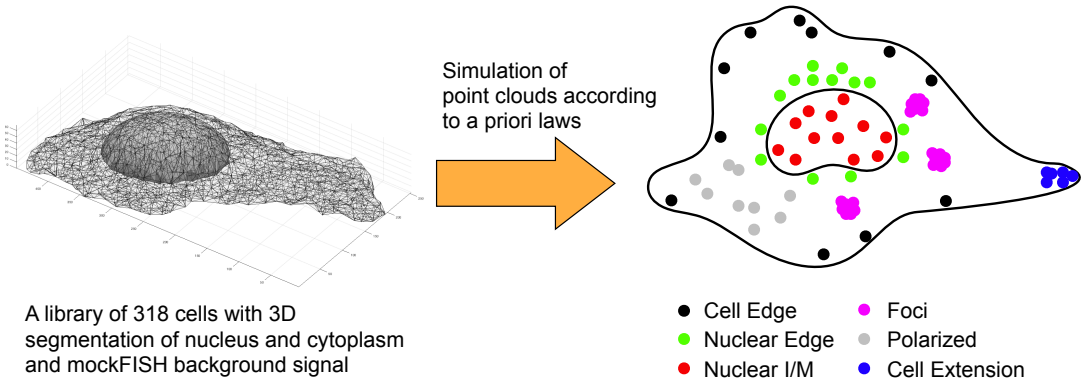

Figure 4.3: Simulation of smFISH images: a library of cell reference volumes is acquired. In these volumes, we place points according to a priori laws of localization. 6 patterns are shown here. Cell edge and nuclear edge are localizations close to the cellular and nuclear boundary respectively (in 2D). Nuclear I/M refers to localizations inside the nucleus or on the nuclear membrane in 3D. Foci correspond to clustering of several RNA molecules in relatively small aggregates.

3D, as well as the background signal obtained by mockFISH. We then placed points into these reference volumes according to a priori laws of localization. In order to control the strength of the pattern, positions were drawn from a mixture distribution between random localization and pattern localization: $u_i \sim \pi\mathcal{S} + (1-\pi)\mathcal{R}$, where $\mathcal{S}$ is the pattern distribution and $\mathcal{R}$ the random distribution. In order to benchmark our method and other existing methods, we simulated 8 different patterns, 6 of which are shown in figure 4.3. In order to test the performance of the algorithms independently from confounders such as cell shape and expression level, we systematically varied cell shape and expression levels for all simulated localization patterns.

With this validation framework, we benchmarked different existing and newly developed methods, in particular different hand-crafted feature sets. In an unsupervised setting, we were able to identify the correct classes with an accuracy of 88% (Samacoits et al. 2018).

Application to real data We applied the workflow to experimental smFISH data for 10 genes (a total of $\sim$ 2000 cells). We then applied several unsupervised machine learning methods in order to analyze these data. Figure 4.4 shows a t-SNE projection (Maaten et al. 2008) of the localization features for the data. We observe that genes with similar localization patterns tend to live in close proximity in the feature space, whereas genes with different localization patterns have the tendency to be located in different regions. This being said, there are no clearly distinguishable clusters: not surprisingly, localization seems to cover a continuum in the feature space. Moreover, we also observe an important heterogeneity in terms of localization, which — given the stochasticity of the entire expression process — was also expected.

Our analysis workflow thus allows us in principle to analyze RNA localization data, to identify localization patterns and to classify cells according to the localization of the mRNA molecules. In addition to the pilot screen shown in figure 4.4, we are currently working on a medium-scale study. In the frame of this project, we extend our existing workflow to include more features. A preliminary analysis along with a series of follow-up experiments allowed us already to identify mRNAs that are translated in specialized translation factories, potentially enabling new gene regulatory mechanisms (Chouaib et al. 2018). While this is already an exciting finding, we suspect that there are many more discoveries to be made that will complement our current understanding of gene expression at the cellular level.

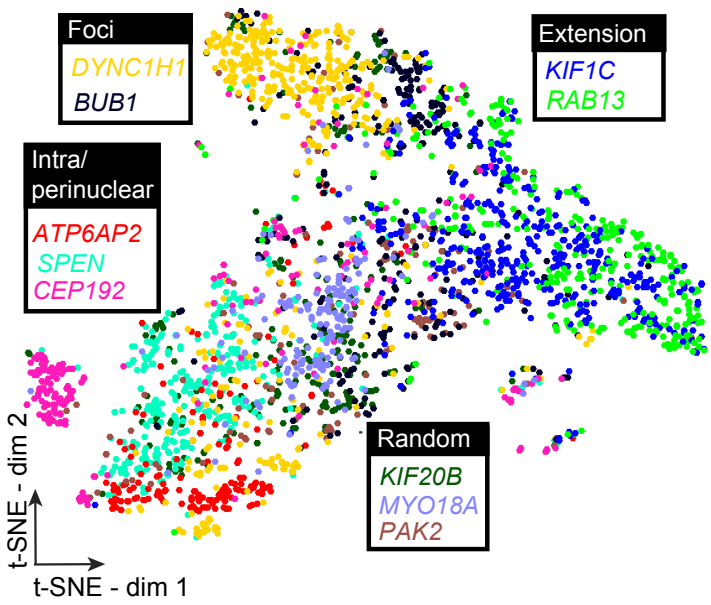

Figure 4.4: t-SNE projection of the localization features for 10 genes. Each point is one cell, genes are color coded according to the legend.

4.3 Conclusion and perspectives

4.3.1 Supervision and publications

The work presented in this chapter is the result of a very fruitful collaboration between my team, Florian Müller from the Imaging and Modeling group (director: Christoph Zimmer, Institut Pasteur) and Edouard Bertrand and his team from the Institut Génétique Moléculaire de Montpellier (IGMM). Our first PhD student was Aubin Samacoïts, cosupervised by Florian Müller and myself, who developed the simulation framework and the analysis with traditional features and applied these methods to the pilot screen (Tsanov et al. 2016; Samacoits et al. 2018). Since then, we have started to work on several extensions of the workflow, namely the use of convolutional neural networks trained on simulated data for the recognition of localization patterns. This approach was investigated by Rémy Dubois, a master student at the CBIO in 2018 and Arthur Imbert, a PhD student who arrived in 2018 (Dubois et al. 2019). Aubin Imbert is also deeply involved in the analysis of the screen on RNA and protein localization, currently under review (Chouaib et al. 2018). The supervision continues to be shared between Florian Müller and myself. We have received funding from the GDR ImaBio (master thesis by Rémy Dubois) and from the ANR (2015-2018).

In a parallel project, I also developed tools to analyze image based cytometry data, where we could analyze 20 protein channels in the human tonsil in order to identify cell types and analyze their spatial distribution (M. Durand et al. 2019).

4.3.2 Perspectives

With our encouraging first biological results, the availability of experimental protocols allowing for large-scale screening and the first generation of computational tools we have developed, there are many exciting perspectives opening up in this field today.

Screens for RNA localization In order to identify more RNAs with non-random localization, our collaborators are currently performing large-scale screens on different gene families, where hundreds of RNAs are probed individually. Our short-term objective is to identify more RNAs translated in translation factories, detectable by organization of mRNA in foci. Moreover, we are generally interested in exploring the localization landscape of gene expression, i.e. identifying all localization patterns in these large-scale screening data. One interesting aspect is to include other markers into this workflow, allowing us not only to evaluate the spatial distributions of RNA molecules with respect to nucleus and cytoplasmic membrane, but also to other organelles in the cell. As an example, we will work on a screen with an additional marker for centrosomes, allowing us to elucidate the role spatial control of gene expression might play for cell division. We also envision the use of *in silico* labeling 2.4 in order to predict major cellular compartments from phase contrast images. This would allow us to monitor the intra-cellular localization of RNA with unprecedented precision. Finally, we will aim at predicting the localization pattern from sequence motifs. This will generate hypotheses of the regulatory mechanisms controlling RNA localization. These data and their computational analysis are bound to provide us with new insights into the local control of gene expression. This project has been accepted for funding by the ANR (project TRANSFACT, ANR-19-CE12-0007-03, 2019-2023).

Learning from simulations With our simulation framework, we can generate large data sets with known ground truth. Rather than using them for validation, we can also use them to train classifiers in a supervised setting. Recently, we have shown that this is possible for a simple Random Forest classifier trained on the handcrafted features proposed in (Samacoits et al. 2018) and with Convolutional Neural Networks, thereby omitting the feature engineering step (Dubois et al. 2019). In principle, such an approach is interesting as it allows one to classify patterns according to biophysical laws rather than annotated examples of known patterns. In addition, we can control the covariates, such as expression level and cell morphology, and obtain classifiers that are much more robust in practice. However, simulated data - albeit visually similar - follow a different distribution than real data. We will therefore apply domain adaptation by adversarial training (Ganin et al. 2016; Shrivastava et al. 2017) in order to overcome the distributional differences between simulated and real data. We will also make use of Generative Adversary Networks (GANs) in order to improve the simulated data.

Spatial Transcriptomics in tissues The screens we have introduced so far use smFISH techniques where the transcripts of a single gene are visualized. On a longer term perspective, we will use techniques that have been recently proposed to visualize the transcripts of hundreds of genes in the same cells (Chen et al. 2015; Moffitt et al. 2016; Achim et al. 2015; Fazal et al. 2019; Eng et al. 2019). These methods are in many cases applicable both to cell culture and tissues, and the newest versions give also access to the subcellular localizations. Measuring several transcripts in the same cells allows us to relate the expression levels and transcript localization of hundreds of genes to each other. Co-expression and transcript co-localization can therefore be measured at an unprecedented level in individual cells with intact morphology.

In particular for tissues, these technical advances will allow us to study challenging and exciting questions. Tissues are composed of cells of diverse cell types interacting physically and chemically to fulfill the tissue specific functions. Cell types are determined by their transcriptional programs, i.e. by the set of genes that are expressed. For this reason, it is usually sufficient to measure the expression levels of a set of genes (typically tens of genes) in order to determine the type of each cell in a given tissue. Understanding the spatial composition of tissues at the cellular level is an extremely interesting objective in itself that has been addressed recently for several tissue types, such as the liver (Halpern et al. 2017). These studies ideally complement the cell type surveys

based on RNAseq recently published for a number of tissue types. A highly interesting question is how this composition changes upon disease. In a collaboration with experimental groups at the Institut Curie and the Institut Pasteur, we will analyze the differences in the abundance and the spatial distribution of the diverse cell types in pulmonary fibrosis, induced by irradiation in mice. The experimental setup allows us also to correlate the composition changes to the time after irradiation, and thereby to understand the order in which these changes occur. This project has obtained funding by the ANR (project LUSTRA, 2020-2024). On a longer-term perspective, such approaches can be also extremely valuable to study the composition of the tumor micro-environment, which is thought to play a major role for cancer progression and outcome.

5 Tissue phenotyping

Most of the work I have presented in this manuscript deals with computational phenotyping at the cellular level. In section 4.3, I have presented an interesting perspective of analyzing molecularly defined cell types in tissue in order to study tissular characteristics in disease conditions. While the proposed study relies on major experimental advances over the last few years (Shah et al. 2016; Eng et al. 2019), the analysis of tissue slides for diagnostic purpose has a long tradition in medicine. In this chapter I will present the contributions of my team in the field of computational pathology. After an introduction to the biomedical context in section 5.1, I will explain two complementary approaches to the analysis of tissue phenotypes: the analysis of single cell data in tissues 5.2 and the end-to-end analysis of full slides in section 5.3, the latter being work in progress at the time of writing. Finally, I will detail the perspectives in this field in section 5.4.

5.1 Biomedical context

Histopathology and H&E staining Histopathological examination of stained tissue slides is a cornerstone of cancer diagnosis and prognosis. Haematoxylin and Eosin (H&E) staining - invented more than 100 years ago - is the most widely used staining protocol in histopathology: cell nuclei are stained in blue (haematoxylin), whereas the cytoplasm is colored in pink (eosin). Other structures can take combinations of blue and pink. H&E staining allows the pathologist to inspect nuclear morphologies, to categorize cell types, and to appreciate the general architecture of a tissue sample (see figure 5.1), albeit without molecular information.

Interpretation of H&E stained images requires several years of special training and a tremendous expertise in medicine and cancer biology. Importantly, given the central role of cellular phenotypes for the visual inspection of a tissue slide, histopathology represents a natural link between fundamental research in cancer cell biology and disease relevance.

Digital Pathology Today, slides can be scanned, visualized, analyzed and annotated on the screen; this is usually referred to as Digital Pathology. Digital pathology has paved the way for the development and application of algorithms to automatically or semi-automatically analyze histopathology images. In most cases, the objective of such a method is to directly assist the pathologist in the diagnosis (Computer Aided Diagnosis, CAD) by detecting and quantifying certain features, such as detecting and measuring the tumor region (Qaiser et al. 2019), estimating the number of dividing cells (Veta et al. 2014), quantifying necrosis (Homeyer et al. 2013) or predicting the cancer type (Coudray et al. 2018). Recently, systems have been proposed to automatically classify entire slides, e.g. for metastasis detection in lymph nodes (Bejnordi et al. 2017; Liu et al. 2017; Lin et al. 2019). The idea is to unburden the pathologist's work by pre-screening the large number of slides and removing those that can be classified as being non-cancerous with high confidence. The workload could thus be reduced by $65 - 75\%$ (Campanella et al. 2019). Another more research-oriented objective is to identify image-based

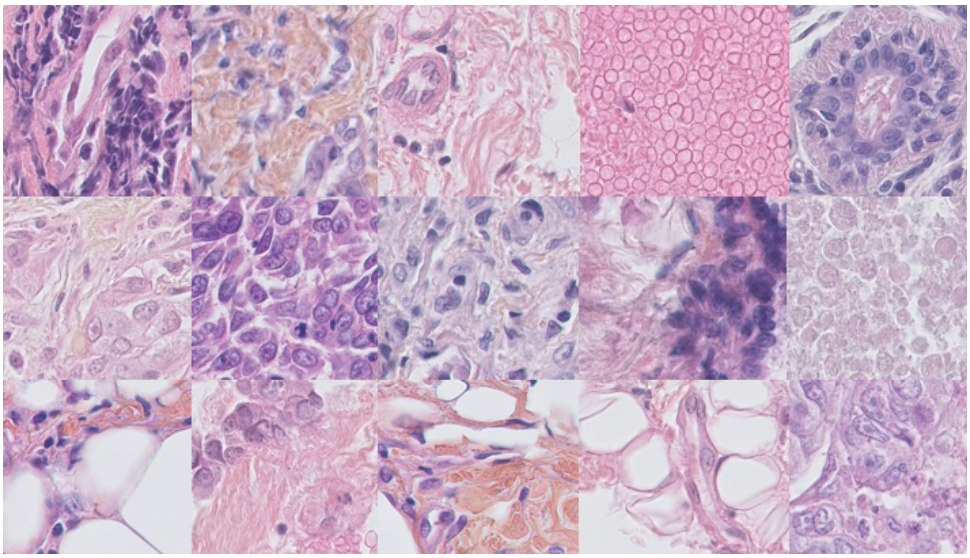

Figure 5.1: Examples of H&E stained tissue sections typically used in histopathology

biomarkers that are predictive for certain outcome variables and that can ideally complement gene-expression signatures (Bera et al. 2019). Another interesting research field is to combine image and molecular data. This can be done in different ways: images can be used in order to deconvolve bulk measurements (Yuan et al. 2012), or to predict the mutational status of genes (Coudray et al. 2018). In summary, while Computer Aided Diagnosis is clearly the most important application of computational pathology, there are more and more articles that go beyond automatically reproducing interpretations by human pathologists.

Computational tasks related to digital pathology Irrespective of the concrete biomedical questions, the main tasks in computational pathology typically fall into one or combinations of the following categories:

1. Color normalization in order to remove the bias induced by different staining protocols used in different centers.

2. Detection, segmentation and classification of cell nuclei.

3. Segmentation of regions (metastatic region, necrotic region, ...)

4. Prediction of output variables from the entire slide.

While points 1, 2 and 3 are clearly defined - albeit challenging - problems, the general strategy to address point 4 is debatable. Indeed, one of the main difficulties in histopathology, as compared to other problems in computer vision, is the size of the images. They are typically in the Gigapixel range (100000×100000 pixels). This makes handling of the images difficult and their processing time-consuming. More importantly, it is unclear how to encode the information of an entire slide: like in genomics, we have a massive amount of data for each patient most of which is probably irrelevant for the disease, and it is unclear which pieces of information are important for the final prediction in a very general setting. In general, there are two main avenues to tackle these problems: either we aim at quantifying a slide with respect to biologically meaningful variables and build predictive models based on these variables or we try to solve the prediction problem directly and eventually try to understand a posteriori which elements of the slide were responsible for the predictions. Finding a suitable encoding is also a prerequisite of integrating

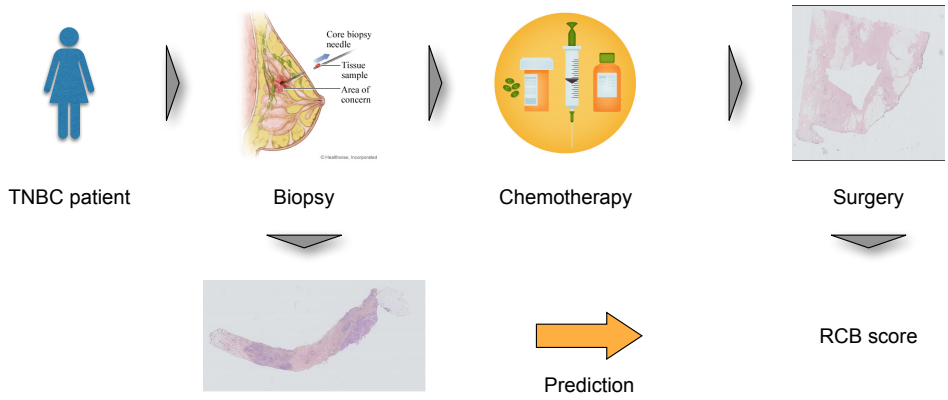

Figure 5.2: Prediction of the response to neoadjuvant therapy

image data with genomic and transcriptomic data in order to use both complementary sources of information for predictions. The methodological developments in this field are paralleled by a considerable increase in the size of generated data sets over the last years. Given all these considerations, we can expect major developments and discoveries in the field of computational pathology.

Prediction of treatment response from biopsy data The concrete medical question my team started to address is the prediction of treatment response from biopsies in Triple Negative Breast Cancer (TNBC). Among women in France breast cancer is the most common cancer and leading cause of cancer deaths with 18.2% of deaths among female cancer patients(Cancer 2017). TNBC is a subtype of breast cancer with poor prognosis and limited treatment options. In TNBC, the malignant invasive cells do not contain receptors for estrogen (ER), progesterone (PR) or HER2 and can therefore not be treated with hormone therapies or medications that work by blocking HER2. The treatment used is neoadjuvant chemotherapy, i.e. chemotherapy prior to surgery. Response to neoadjuvant chemotherapy varies among patients and can be quantified after surgery via a Residual Cancer Burden (RCB) score (score between 0 and 3, where 0 corresponds to a complete response). The objective of our project is to predict this score from biopsy data (see figure 5.2). Upon success, patients could benefit from such a prediction, as they could be spared an invasive treatment that is likely to fail in their case. On a research perspective, we hope that such a system would point us to the cellular and tissular features that are informative about the responsiveness. Of note, pathologists are currently not able to predict resistance to neoadjuvant chemotherapy from biopsies.

5.2 Cellular Phenotyping in cancer tissues

In view of extracting biologically interpretable features from a Whole Slide Image (WSI), the cellular level plays a pivotal role. Cancer is a genetic disease, where cells acquire a set capabilities that moves them to a neoplastic state (Hanahan et al. 2011). Moreover, we have a detailed, yet incomplete, understanding of how changes in the genome or the transcriptional program effect cellular phenotypes, and we can relate morphological phenotypes to affected biological processes. It seems therefore logical to include a cellular level in the analysis of diseased tissue. In the context of analysis of H & E stained tissue sections, it makes sense to focus on nuclei, because they are indicative of many cellular phenotypes (Chow et al. 2012), their morphology is currently used by pathologists in order to identify the mitotic index and the level of nuclear pleomorphism

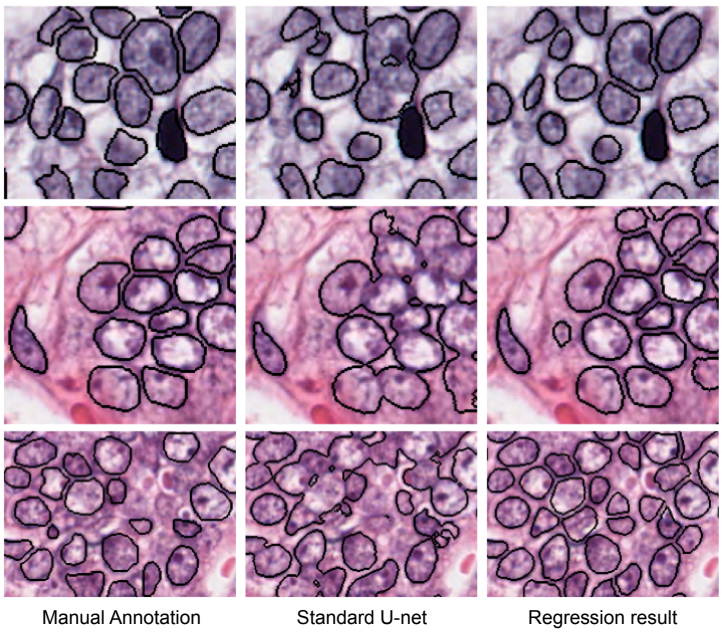

Figure 5.3: Results for nuclei segmentation, as presented in (Naylor et al. 2018)

(Elston et al. 1991) and — unlike other cellular structures — they appear reasonably well contrasted under standard staining procedures.

Segmentation of nuclei in stained tissue sections Segmentation of nuclei in WSI is the first essential step for cellular and tissular phenotyping. As can be seen from figure 5.1, it is actually a rather challenging problem and many traditional image analysis methods have been proposed to address this problem, including mathematical morphology, level sets and graph-based segmentation (Irshad et al. 2014; Xing et al. 2016). Today, fully convolutional neural networks are considered to be among the most powerful methods for image segmentation, the U-net (Ronneberger et al. 2015) being particularly popular for biological applications (Falk et al. 2019).

While fully convolutional neural networks without any kind of postprocessing often give excellent results at the pixel level, they typically fail to segment touching objects and therefore tend to give bad results at the object level. In order to address this issue, there are several strategies, such as giving larger weights to pixels in close proximity of the object contours (Ronneberger et al. 2015), predicting both the objects and their contours (Van Valen et al. 2016; Kumar et al. 2017) or learning a notion of the object by combining object region prediction with pixel-level segmentation (He et al. 2017).

In (Naylor et al. 2017; Naylor et al. 2018), we published two methods to address this issue. In (Naylor et al. 2017), we observed that the posterior probabilitiy obtained by the fully convolutional network typically decreases towards the borders of the nucleus. We therefore argued that objects should be split if on any path linking two local maxima (nuclei centers) the decrease in posterior probability is sufficiently large, which sums up to application of morphological dynamics (Michel Grimaud 1992) and the watershed transformation (Beucher et al. 1979). While appealing at first sight, we observed also that this strategy can lead to severe oversegmentation in difficult cases, i.e. in cases where the hypothesis that the posterior probability gradually decreases between center and border of the nucleus is violated. We therefore proposed to formulate the instance segmentation as a regression problem, where we predict the distance map instead of predicting hard pixel classes (Naylor et al. 2018), i.e. instead of predicting for each pixel u_i

the variable y_i:

$$y_i = \begin{cases} 1, & \text{if } u_i \in \bigcup_j S_j \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

we predict a continuous variable y_i:

$$y_i = \begin{cases} \min_{v \notin S_j} d(u_i, v), & \text{if } u_i \in S_j \\ 0, & \text{if} u_i \notin \bigcup_j S_j \end{cases} \tag{5.2}$$

Here, S_j denotes the connected components of the groundtruth data. The formulation in equation 5.2 shows already that the variable we are trying to predict combines pixel with object level. Consequently, the network which is optimized to predict y_i is bound to learn the notion of an object. More generally, it makes sense to predict some object feature along the pixel label in order to segment instances, which is finally also the rationale in the famous mask-R-CNN (He et al. 2017). Examples are shown in figure 5.3, for a detailed quantitative analysis and benchmarking, I refer to (Naylor et al. 2018).

With these results, we have now the challenging opportunity to segment all nuclei in WSI in a variety of project and to study the morphological landscape of nuclei as well as their spatial distribution in different types and subtypes of cancer, and in particular to relate this phenotypic information to a variety of clinical variables. We are going to detail these approaches in section 5.4.

5.3 Analysis of Whole Slide Images

Complementary to the approach of cellular phenotyping, we also started to investigate the problem of direct prediction of the output variable from the WSI, without prior segmentation of cells or tissue areas. As mentioned in section 5.1, one of the major problems in digital pathology is the size of the images which makes it impossible to process entire WSI with Neural Networks. In addition, there are many features at the slide level that are not informative about the output variables, such as the shape and the size of the biopsy itself. For these reasons, the common approach is to partition the WSI into a large number of smaller images, usually referred to as tiles. In contrast, the most important annotations are made at the slide level (such as disease state or prognosis or — in our case — the response to treatment). This problem therefore falls into the category of weak supervision (inaccurate supervision) and can be addressed with techniques usually referred to as "multiple instance learning" (MIL). Importantly, it is not known a priori how many of the tiles may be informative: even a small region in the entire slide might contain important and even decisive information. Taken together, we are still in need of systematic studies of the existing methods and new algorithms to reach a meaningful encoding of entire slides.

In analogy to the HCS approach shown in section 2.1, we argued, that if there were different tile classes, such as tumor, stroma, etc. we could simply represent each slide i by the vector of percentages $z^{(i)}$ of tiles falling into each of the categories. As we do not have annotations at the tile level, we can find tile classes by unsupervised learning from the pooled (and down-sampled) set of tiles from all slides, where each tile j in slide i is represented by a feature vector $x_j^{(i)}$ from a pretrained network. This approach is illustrated in figure 5.4.

Using this approach as a baseline, we argued that we can replace the cumbersome off-line clustering step by a bottleneck layer in an end-to-end neural network approach, illustrated

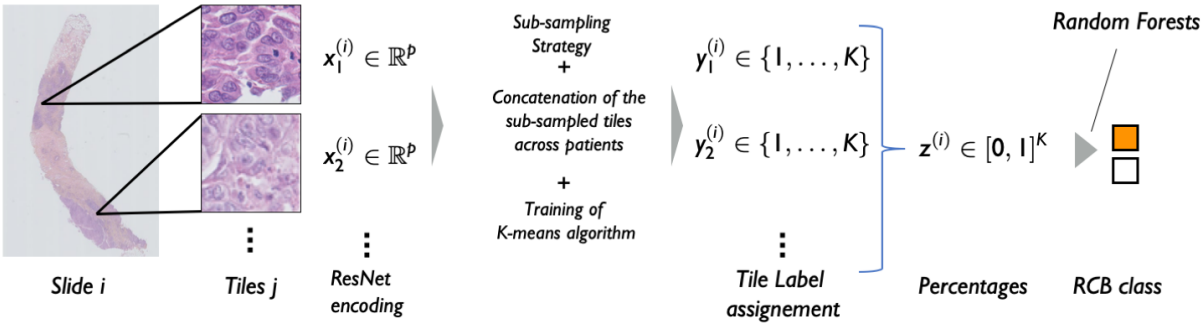

Figure 5.4: Two step method: ResNet features $x_j \in \mathbb{R}^P$ are extracted and used for clustering. Each tile j is therefore represented by the cluster label y_j and the slide is represented by the percentages of tiles in each of the clusters.

in figure 5.5. Instead of assigning to each feature vector $x_j^{(i)}$ one hard cluster label we map each $x_j^{(i)}$ to a low dimensional representation $y_j^{(i)} \in \mathbb{R}^K$. This can be seen as a generalization of a cluster assignment. We use a 1-dimensional convolution to learn the mapping. We then pool these representations $y_j^{(i)}$ for all j to reach a description $z^{(i)}$ of the entire slide. If we set $K = 1$ and use the Weldon pooling (T. Durand et al. 2016) this model is essentially the one proposed by (Courtiol et al. 2017). Prediction of RCB as a proxy of chemotherapy efficiency reached 60.6%, and therefore gave similar performance as prediction with Random Forests from manual measurements obtained by a pathologist (59.8%). Training on some additional images allowed us to further increase the score up to 70%. While this is far from being usable in clinical practice, we see that there is at least some signal. We are currently working on extensions and modifications of this method, also integrating more data.

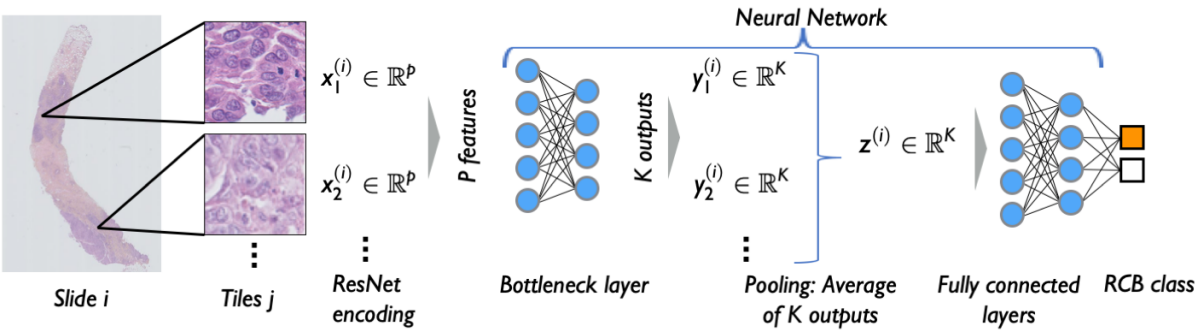

Figure 5.5: The clustering approach from figure 5.4 is replaced by a bottleneck layer of a Neural Network.

5.4 Conclusion and perspectives

5.4.1 Supervision and publications

My team is working in the field of histopathology since 2014. We started working on the detection of mitoses (Veta et al. 2014) and turned then first to the segmentation of nuclei (Naylor et al. 2017; Naylor et al. 2018). More recently, we have started to work on the prediction of treatment response (Naylor et al. 2019). The main body of work was performed by Peter Naylor who defended his PhD thesis in 2019. The PhD project was based on a collaboration between my team and Fabien Reyal (Institut Curie). Peter Naylor received a PhD fellowship by the Ligue contre le cancer.

In fall 2019, Tristan Lazard started his PhD thesis in the field of digital pathology, funded by QLife. Guillaume Bataillon, a trained pathologist, joined the team in 2020 for 6 months, and later in the year we expect the arrival on a new PhD student who will also work in the field of digital pathology, but on a different cancer type.

5.4.2 Perspectives

As described in section 5.1, there are numerous perspectives, both short-term and long-term. Of note, we are regularly confronted with new biomedical questions as more and more large-scale data sets are acquired. While the last data sets we were working on contained less than 200 WSI, newer datasets contain 1000-2000, and this number is ever increasing.

The data we will focus on for the next 2-3 years comes from the pathology department of the Institut Curie thanks to a collaboration with Anne Vincent-Salomon. These data sets are specific, in a sense that they focus on one particular type of cancer.

- Ovary cancer (1400 slides). In addition to the slides, we have data on survival, risk of relapse and the mutational status of BRCA1 and BRCA2, both genes whose mutation is known to be an important risk factor.

- Breast cancer (800 slides). In addition to the slides, we have data on grade, survival, risk of relapse, mutational status of BRCA1 and BRCA2 and information on the homologous recombination deficiency (HRD), a defect in the DNA repair pathway.

Cellular Phenotyping As discussed in section 5.1, cellular phenotyping is one of the main approaches in order to make interpretable predictions from WSI slides, because cellular phenotypes directly point to deregulated cellular processes whose molecular basis is often at least partially understood. Cellular phenotyping therefore provides an interesting link between basic research in cell biology and the clinical impact of these findings. The availability of segmentation and detection methods we and others have developed (Naylor et al. 2018; Hollandi et al. 2019; Stringer et al. 2020) provides us with the challenging opportunity to study the morphological landscape of cell nuclei in cancer. Technically, there are a number of difficulties:

- There is a large number of different cell types, not all of which are distinguishable in H&E images. In addition, the cell type is not only defined by shape and texture, but also by the spatial relation to other cells.

- Annotated data sets are scarce as cell type annotation does not belong to the standard diagnostic procedure.

- Different tissue types contain different cell types; annotated data sets for one tissue type are therefore not easily transferable to other tissue types.

- Differences in staining procedure and the scanning system make classifiers difficult to use for different or heterogeneous data sets.

- Finally, the most difficult problem is to aggregate information at the cellular level in order to make predictions at the slide level.

5 Tissue phenotyping

My team is currently working on methods for cellular phenotyping. We use both hand-crafted and Deep Neural network features in order to explore the morphological landscape of different cancers and classify cells according to type (e.g. epithelial, endothelial, lymphocytes, ...) or phenotype (e.g. mitotic cells, dead cells, ...).

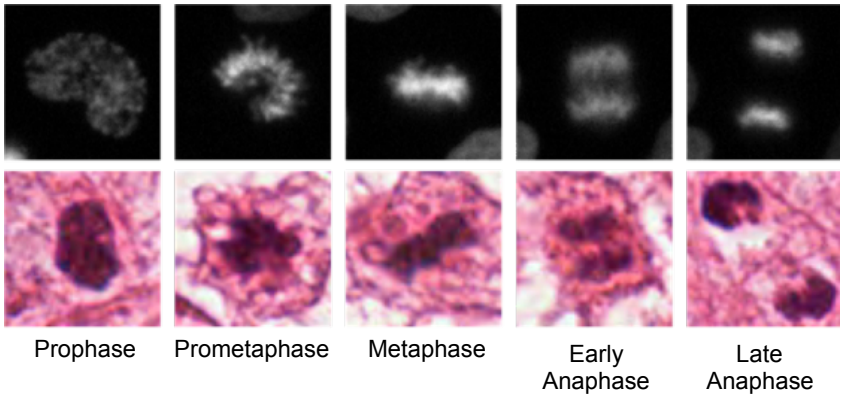

Figure 5.6: Mitotic phases as observed by fluorescence microscopy in cell culture and in H&E stained tissue sections

In order to tackle the problem of scarce annotated data sets, we propose to use fluorescence microscopy data in combination with domain adaptation in order to boost the performance of trained classifiers (see figure 5.6). Being able to train classifiers on fluorescence microscopy data and to apply them to histopathology data would not only boost phenotypic recognition, but could also help in identifying similar morphological phenotypes between experiments in cultured cells and in diseased tissue.

However, this approach can not be extended to the recognition of cell types. For this, we still need to make use of manual annotations in a more traditional setting. In contrast to phenotype recognition and whole slide image annotations, there are few attempts to publish annotated data sets for cell type classification so far. On the long run however, there will be more publicly available annotated data sets for cellular phenotyping. Another interesting avenue is experimental ground truth generation as we have discussed in 2.4.

WSI encoding In order to aggregate cellular information, we propose to build a low-resolution image of multiple channels, where each tile corresponds to one pixel. The value of each pixel in one channel is the percentage of cells in the tile falling into one of the different cell type categories. Then, we train a neural network on these low-resolution images to predict the output variables. The neural network therefore integrates information from the cell level into the decision taken at the slide level and thereby propagates information from higher resolutions to lower resolutions. Importantly, the network integrates both phenotypic information and their spatial distribution. First results show promising albeit not yet conclusive results. Alternative approaches include measurements of cell type abundance inside important regions and analyzing the spatial distribution by graph based approaches.

We will also continue to work on our previous approach described in section 5.3. Indeed, there are many design choices that remain to be explored. For instance, we can try to not only predict the RCB score, but the full set of measurements reported by the pathologist. In this multi-task setting, we could probably learn more meaningful representations.

Integration of molecular and image data The most important challenge we are facing is the integration of molecular and image data. In the projects described above, we aim at predicting the mutational status of genes known to play a mechanistic role in cancer progression. In a second step, we can extend these approaches to predict the expression levels of more genes, potentially with unknown role for cancer progression. This can be done both at the cellular and the slide level. First attempts seem to be rather promising (Bera et al. 2019; Coudray et al. 2018). Another interesting approach is to integrate genomic and image data in order to make predictions of clinical variables. This data integration can be achieved elegantly by kernel methods.

6 Conclusion and Perspectives

In this chapter, I will briefly summarize my developments over the last years and summarize the perspectives I have detailed in the chapters 2 to 5.

6.1 Conclusion

This manuscript describes my developments in the field of computational phenotyping between 2006 and 2020 with applications in High Content Screening and histopathology. High Content Screening allows us to explore the phenotypic space, i.e. to study the cellular phenotypes that arise from changes in gene expression or from perturbation by drugs. This gives us precious insights into the regulation of important cellular functions, including disease relevant processes, as well as the mechanism of action of drugs. HCS can also be used to better understand properties of living systems without perturbation, such as localization of biomolecules (RNA or proteins). Outside the context of experiments, new insights can also be gained in analyzing diseased tissue data, where we can study the effect of natural perturbations caused by disease. Stained tissue sections provide an excellent link between clinical relevance and fundamental science, and this link is provided by the cellular phenotypes.

I subdivided my work over the last years into 4 main domains, according to the biological property we are exploring: morphological phenotyping of cells (chapter 2), analysis of phenotypic dynamics (chapter 3), analysis of spatial patterns (chapter 4) and analysis of tissue architecture (chapter 5). All of these projects are biology driven; the objective of my work is to answer biological questions with computational methods, applied to large scale image data. From a methodological point of view, most of the problems I am confronted with in this context can be framed as supervised or unsupervised learning problems, such as the classification of nuclear morphologies, clustering of phenotypic time-series, characterization and clustering of spatial trajectories, classification of localization patterns, as well challenging segmentation tasks.

Given the methodological (r)evolution in computer vision, traditional pattern recognition approaches (segmentation, feature extraction, shallow classification) are more and more replaced by deep learning approaches. Indeed, deep learning does not simply improve existing methods in terms of accuracy, but provides a new arsenal of methods that have the potential to revolutionize the field of Bioimage Informatics.

6.2 Perspectives

My work is situated at the interface between biology and computer vision. For this reason there are always two aspects for my research projects: the biological question along with the expected discoveries and the methodological developments that are required to answer these questions.

6 Conclusion and Perspectives

From a biological perspective, my research can be divided into three axes : computational phenotyping for High Content Screening (HCS), understanding the spatial aspects of gene expression and digital pathology. These three application domains are good examples for the sources of information that can be efficiently explored with imaging assays: morphological phenotypes, spatial configurations and multi-scale organization of living systems.

From a methodological perspective, my current research interests focus on Deep Learning for biomedical image analysis, and namely for computational phenotyping. Indeed, Deep Learning has revolutionized Computer Vision, and many problems that were considered several years ago to be too difficult for automatic image analysis can now be successfully addressed. Yet, the major hurdle in using deep neural networks for bioimaging are the large annotated data sets that are typically required and which — in contrast to medical applications, where imaging protocols and readouts are standardized — seem unrealistic for many biological applications, where imaging conditions and the analysis task vary dramatically between different projects. For each of the biological applications, I will therefore detail the strategy to overcome this obstacle.

Learning without manual annotations: High Content Screening by label-free microscopy

Computer Vision has always been the method of choice for the analysis of High Content Screening data. In most cases, traditional methods for supervised and unsupervised learning achieved excellent results in the recognition of protein localization patterns (Glory et al. 2007) or morphological phenotypes (Neumann et al. 2010). I argue that Deep Learning has not only the potential to improve results obtained by traditional learning methods, but also to address new problems we did not consider to be solvable before.

Usually, HCS employs fluorescent markers tailored to the biological question studied. However, fluorescent markers also come with a number of drawbacks, in particular for live cell imaging. For instance, the number of channels is limited (often to 2 or 3), live dyes often fade out and stable expression requires genetic modification of the cell lines, which might not be wanted. It turns out that much of the information that is highlighted by fluorescent markers is in principle present in traditional label-free microscopy. In order to reveal this information, we will first acquire a dataset with both label-free and fluorescence microscopy. From there, we have two options:

1. We can predict the fluorescence image from the label-free microscopy image — a strategy referred to as "in silico labeling" (Christiansen et al. 2018). This strategy can be used to detect important elements inside the cell (e.g. nucleus, cytoplasm, Golgi), which can then either serve as reference markers to enrich the phenotypic description of each cell.

2. We can use the fluorescence images to assign a hard label to the entire cell according to its phenotype (e.g. the cell cycle phase or alive/dead) and then train a classifier to predict this label from the label-free microscopy. In this case, the fluorescence microscopy serves as an experimental ground truth generator.

While seemingly similar, the two options imply different analysis workflows and architectures (image generation versus object detection).

From a methodological point of view, this approach can very elegantly overcome the need for massive annotations, simply by replacing manual annotation by experimental data. First works on cross-modality image reconstruction show indeed very promising results (Christiansen et al. 2018; Weigert et al. 2018; Ounkomol et al. 2018), but they have not yet been demonstrated to perform well in screening mode, where we typically face a large phenotypic heterogeneity.

Furthermore, the training of Neural Networks for object detection and cell classification from fluorescence microscopy images as ground truth is not yet mainstream, but clearly is a very promising strategy to overcome the need for massive image annotation. These approaches also raise the interesting question of how to statistically treat measurements that are made on predicted images, rather than on acquired images (Whitehill et al. 2018).

From a biological point of view, I will apply these methods in three different projects: (1) a large series of live-cell imaging experiments for the study of Chimeric antigen receptor (CAR) T-cells, an important anti-cancer treatment, (2) a cytokinesis screen by live cell imaging aiming at identifying key regulators for this last step of cell division and (3) a number of medium-scale RNA localization screens, where we wish to increase the number of reference structures in the cell and a way to stratify the cells with respect to their phenotypes (such as cell cycle).

Learning from simulations: RNA localization and spatial transcriptomics A good example for the power of imaging approaches and their complementarity to *omics* data is the study of spatial aspects of gene expression, both at the cellular and the tissular level. Preferential localization of transcripts inside cells was long considered to be of minor importance. Recent studies show that many RNAs are non-uniformly distributed inside cells, but we are still lacking a comprehensive picture on the subcellular localization preferences of RNAs. Single Molecule Fluorescence in situ hybridization (smFISH) allows to visualize individual transcripts inside cells. Importantly, the technique is scalable and can be applied in screening mode, probing hundreds of different RNAs. These screens will ultimately allow us to identify the patterns according to which RNA localize in cells and to understand which RNA localizes according to which pattern. Thorough analysis of these exciting new data have the potential to point to new mechanisms for the spatial control of gene expression (Chouaib et al. 2018). While in principle the analysis of these data can be addressed by a pattern recognition workflow, where we segment the cells, detect the individual RNA molecules, represent their spatial distribution by features and finally apply supervised or unsupervised learning on the resulting feature vector, we are facing the problem of generating representative training sets (Samacoits et al. 2018). This is difficult, because the data is complicated to analyze manually, and because we cannot control for known co-variates, such as the number of RNA molecules and the shape of the cell. In order to overcome these problems related to image annotation, we propose training classifiers on simulated data and to apply domain adaptation (Ganin et al. 2016) in order to make them applicable to real data. Another interesting extension to the current state-of-the art is the use of in silico labeled cells, as described above. This would help us to identify and interpret new localization patterns.

At the tissular level, smFISH can be used in multiplex and/or sequential mode in order to visualize RNA molecules transcribed from different genes in the same cells. There is a number of techniques to perform this multiplexed smFISH; ultimately they all result in large multi-channel images of a tissue. By single cell segmentation and analysis of the smFISH in the different channels, we can derive for each cell a gene expression profile for a potentially large number of genes. While the approach is usually not comprehensive, the number of RNAs simultaneously monitored is largely sufficient to make the cell type identifiable. The cell types present in the tissue type can be identified independently by single cell RNA sequencing (scRNAseq), and a sparse representation that recapitulates the cell types can be found. smFISH against these markers then allows to identify these cell types and to map them spatially in the tissue (Zhu et al. 2018). With my team, I will work on computational methods to analyze the spatial configurations of the mapped cells in their native context and infer physical interactions. In particular, I am interested in the change of these configurations under pathological conditions. In our recently funded project LUSTRA, we follow this strategy by analyzing lung tissue of

mice with induced pulmonary fibrosis. The overall strategy will be also applicable to other tissue types and diseases.

Phenotyping diseased tissue: applications to digital pathology New imaging technologies, such as spatial transcriptomics, will provide us with unprecedented views on tissue architecture, but — as all new technologies — they do not allow us to relate these insights to longer-term observations, such as patient outcome or relapse risk. For this, we need to turn to large cohorts imaged with traditional techniques, such as H&E staining.

Indeed, analysis of H&E stained tissue is both timely and important in several ways: first, in the context of computer aided diagnosis, machine learning can help pathologists in their daily routine, e.g. to quantify certain aspects, to pre-screen slides or to point to suspicious regions in the image. Second, histopathology data provides a link between biologically interpretable cellular phenotypes and clinical variables. We therefore aim at analyzing tissue slides both at the level of single cell phenotypes and higher level tissue architecture. Concretely, we wish to predict clinical variables as well as genetic and expression data from stained tissue slides. This will unravel the link between molecular and cellular information and thereby help to bridge the gap between the molecular scale and the patient scale.

Technically, the main difficulty is the size of the images, and the imbalance between the scale of the image and the scale of the potentially relevant information: small tissue regions and even single cells can contribute to the final classification of the entire slide, while many global features are meaningless (such as the shape and size of the tissue sample). For this reason as well as memory limitations on GPU cards, a slide is usually subdivided into tiles that are then processed individually. The tile results are then aggregated in a second step. However, most clinical annotations are made at the slide or patient level. The challenge is therefore to learn useful representations at the tile level from a global annotation concerning a set of tiles. The relevant technique is called "Multiple instance learning" (MIL) and represents the most promising strategy for digital pathology today. However, there are many design choices and open questions within this framework that are far from being definitely solved, e.g. how to represent tiles, how to aggregate tile scores and how to normalize the color between different imaging centers. Moreover, most publications still focus on the prediction of tumor presence and tumor segmentation (Campanella et al. 2019) or subtype prediction. With my team, I will work on the prediction of response to different treatments, such as neoadjuvant chemotherapy or immunotherapy, as well as the prediction of molecularly defined data, such as homologous recombination deficiency (HRD) or genetic mutations, which would allow these algorithms to be part of a precision medicine approach and unravel the link between genetic information and tissular phenotypes.

Another interesting approach is to use cellular phenotypes for whole slide prediction. Based on our recent work on nuclei segmentation in tissues (Naylor et al. 2019), we can morphologically profile all individual cells in a tissue and thereby build an atlas of morphological phenotypes at the single cell level for cancer tissues. In a second step, my team will make use of innovative deep learning approaches to reveal the impact of cellular phenotypes with respect to clinical variables, such as subtype, treatment response or risk of relapse. Preliminary analyses (Naylor 2019) show that these analyses are on par with the aforementioned MIL approaches, but provide a much higher level of interpretability. Finally, the morphological cancer atlas can also serve as reference to query for the presence of particular cellular phenotypes in cancer tissues and therefore provides a natural link between basic research in cell biology, where cellular phenotypes are often well understood and their implication for cancer.

Altogether, my research project aims at developing new methods to analyze cellular and tissular phenotypes. Bioinformatics traditionally describes living systems by the analysis of genetic and transcriptional information. As stated by major scientists in the field, such as Trey Ideker (personal communication) or Sydney Brenner (Brenner 2010), one of the major obstacles of large scale approaches in genome and systems biology is the enormous gap between information at the single nucleotide level and the level of an entire biological system. Images play a pivotal role in bridging this gap, as they have the potential to link multiple scales and to explore dimensions that are complementary to molecular information, such as space, time and morphology. As computational scientists, we are facing today the challenging opportunity to design methods that will allow us to link these types of information to the molecular level, and thereby to bring an important piece to our understanding of the mechanisms of life.

A Microscopy

Here, I very briefly review the most current microscopy techniques used in life sciences today. This presentation cannot be exhaustive of course, but it should provide the reader with an idea of the variety and diversity of existing methods to acquire the data the methods described in this thesis aim at dealing with. The presentation is taken from (Walter et al. 2010b). Figure A.1 illustrates examples for the different techniques.

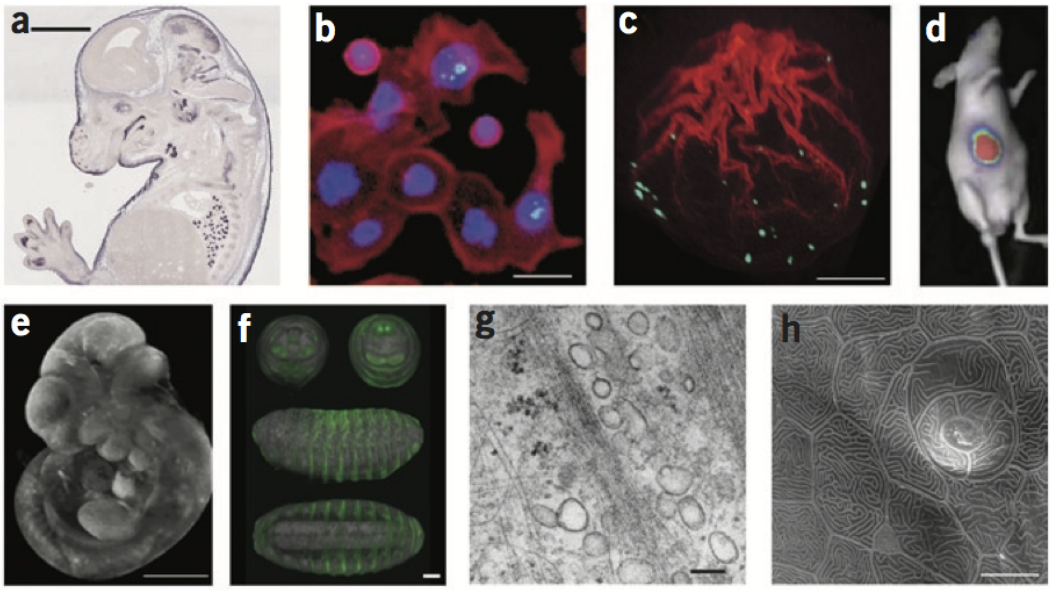

Figure A.1: Microscopy techniques. (a) Brightfield microscopy: mouse embryo, in situ expression pattern of Irx1, Eurexpress; scale bar: $2mm$. (b) Fluorescence microscopy: HT29 cells stained for DNA (blue), actin (red) and phospho-histone H3 (green)75; scale bar: $20\mu m$. (c) Confocal microscopy: actin polymerization along the breaking nuclear envelope during meiotic maturation of a starfish oocyte. Actin filaments, red (rhodamine-phalloidin stain); chromosomes, cyan (Hoechst 33342 stain). Projection of confocal sections, (image courtesy P. Lénárt); scale bar: $20\mu m$. (d) Bioluminescence imaging: in vivo bioluminescence imaging of mice after implantation of Gli36-Gluc cells76, (figure courtesy B.A. Tannous). (e) Optical projection tomography: mouse embryo, EMAP33,66; scale bar: $1mm$. (f) Single/selective plane illumination microscopy: late-stage Drosophila embryo probed with anti-GFP antibody and DRAQ5 nuclear marker: frontal, caudal, lateral and ventral views of the same embryo77; scale bar: $50\mu m$. (g) Transmission electron microscopy: human fibroblast, glancing section close to surface (image courtesy R. Parton and M. Floetenmeyer); scale bar: $100nm$. (h) Scanning electron microscopy: zebrafish peridermal skin cells (courtesy R. Parton and M. Floetenmeyer); scale bar: $10\mu m$

Brightfield microscopy with colorimetric stains is the primary technique for capturing tissue and whole organism morphology (Fig. A.1.a). For high-throughput capture of in situ expression patterns, automated bright-field microscopy has been used for whole-genome projects such as the Allen Brain Atlas.

Widefield fluorescence microscopy is the most widely used imaging technique in biology

(Fig. A.1.b). Fluorescent markers make it possible to see particular structures with high contrast, either in fixed samples using immunostaining or in living cells with expressed GFP-tagged proteins83. The resolution is limited by diffraction to about 200 nm.

Confocal scanning microscopy generates optical sections through a specimen by pointwise scanning of different focal planes and thereby reduces both scattered light from the focal plane and out-of-focus light84. The image quality of two- dimensional images is therefore improved, and 3D images can be taken (axial resolution is typically 2 to 3 times lower than lateral resolution; see Fig. A.1.c). The method is also applicable to live cell imaging. There are variants of this method increasing axial resolution (for example, 4Pi microscopy)85.

Computational optical sectioning microscopy (COSM) achieves optical sectioning by taking a series of two-dimensional images with a widefield microscope focusing in different planes of the specimen84. Out-of-focus light is then removed computationally.

Structured illumination microscopy acquires several widefield images at different focal planes using spatial illumination patterns84. As the out-of-focus light is less dependent on the spatial illumination pattern than the in-focus light, combinations of different images at the same focal plane under laterally shifted illumination patterns allow computational attenuation of out-of-focus light.

Two-photon microscopy is similar to confocal scanning microscopy but uses nonlinear excitation involving two-photon (or multiphoton) absorption86. This allows the use of longer excitation wavelengths, permitting deeper penetration into the tissue and — owing to the nonlinearity — confines emission to the perifocal region, leading to substantial reduction of scattering.

Super-resolution fluorescence microscopy groups several recently developed methods in light microscopy capable of significantly increasing resolution and visualizing details at the nanometer scale. In stimulated emission depletion (STED) microscopy (Hell 2003), the focal spot is ńarrowedb́y overlapping it with a doughnut-shaped spot that prevents the surrounding fluorophores from fluorescing and thereby contributing to the collected light. In PALM (photoactivated localization microscopy) (Betzig et al. 2006) and STORM (stochastic optical reconstruction microscopy) (Rust et al. 2006), subsets of the fluorophores present are activated and localized. Iterating this process and combining the acquired raw images yields a high-resolution image.

Bioluminescence imaging (Fig. A.1.d) is based on the detection of light produced by luciferase-mediated oxidation of a substrate in living organisms. Transfected cells expressing luciferase can be injected into animals, or transgenic animals can be created that express luciferase as a reporter gene. When such animals are injected with a luciferase substrate, light is produced by the luciferase-expressing cells in the presence of oxygen. The bioluminescence image is often superimposed on a white-light image to show localization of the light-producing cells.

Optical projection tomography captures object projections in different directions as line integrals of the transmitted light89 (Fig. A.1.e). From these projections (corresponding to the śhadow' of the object), a volumetric model can be calculated by means of back- projection algorithms.

Light sheet-based fluorescence microscopy uses a thin sheet of laser light for optical sectioning and a perpendicularly oriented objective with a CCD camera for detection of the fluorescent signal. Single- or selective plane illumination microscopy (SPIM)90 (Fig. A.1.f) adds sample rotation that enables acquisition of large samples from multiple angles. Low phototoxicity, high

acquisition speed and ability to cover large samples make it particularly suitable for in toto time-lapse imaging of developing biological specimens, such as model organism embryos, with cellular resolution.

Transmission electron microscopy (TEM) (Fig. A.1.g) uses accelerated electrons instead of visible light for imaging. As a result, the achievable resolution (typically 2 nm) is much higher than in light microscopy. The method is not applicable to live cell imaging, and the specimen preparation is technically very complex. In electron tomography, the specimen is physically sectioned and 3D images are obtained by imaging each section at progressive angles of rotation, followed by computational reassembly to yield a tomogram. Resolution ranges from 20-30 nm to 5 nm or less.

Scanning electron microscopy (SEM) (Fig. A.1.h) produces an image of the 3D structure of the surface of the specimen by collecting the scattered electrons (rather than the transmitted electrons as in TEM). The resolution is typically lower than for TEM.

B Trajectory features

A single cell trajectory corresponds to a sequence of cell center positions :

$$u_{t=1...T} \qquad \text{with} \qquad u_t \in \mathbb{R}^2 \tag{B.1}$$

T denotes the length of the trajectory, i.e. the number of consecutive frames the objects has been tracked on.

Basic trajectory features Some basic trajectory features can be defined:

- Effective path length L $= \|u_T - u_1\|_2$
- Effective speed $\frac{L}{T}$
- Largest Move $\max_t \|u_t - u_{t+1}\|_2$
- Straightness index $\sqrt{T}L/P$, where P is the total path length $\sum_t^{T-1} \|u_{t+1} - u_t\|_2$.

Trajectory features based on moments of displacement These features are inspired by biophysical approaches describing diffusive properties of particle movement, as described in (Sbalzarini et al. 2005). While diffusive modeling of an active process such as cell migration or nuclear motility might be questionable, these features can still be calculated and describe meaningful properties of the trajectories, irrespective of the underlying physical process.

The moments of displacement for a single trajectory $(u)_{t=1...T}$ are defined as:

$$\mu_\nu(\Delta_t) = \frac{1}{T - \Delta_t} \sum_{t=1}^{T-\Delta_t} \|u_{t+\Delta_t} - u_t\| \tag{B.2}$$

The original definition from statistical physics involves averaging over N particle trajectories, rather than time, i.e. $\mu_\nu(t) =< (u_n(t) - u_n(0) >_{n=1...N}$. Equation B.2 therefore requires ergodicity and long trajectories which are questionable in our case. Nevertheless, we can represent important features of our trajectories based on equation B.2.

Following the presentation in (Sbalzarini et al. 2005), we can assume that $\mu_\nu \propto t^{\gamma_\nu}$. Furthermore, for all self-similar processes, we have $\gamma_\nu \propto \nu$. The proportional factor between γ_ν is called *movement type* γ and quantifies how directed the particle motion is. If γ is equal to 1, the movement is perfectly directed, whereas if γ is equal to 0.5, it is perfectly diffusive. Between 0.5 and 1, the movement is super-diffusive, whereas below 0.5 it is called sub-diffusive.

A special case of equation B.2 occurs for $\gamma = 0.5$ and $\nu = 2$, which corresponds to mean-squared displacement (MSD) for a perfectly diffusive process. In this case, $\gamma_2 = 1$, and we see that the MSD is proportional to Δ_t:

$$\mu_2(\Delta_t) = D\Delta_t \tag{B.3}$$

D is called the *diffusion coefficient* and can also be used as feature. In order to include some measure of how adequately the process can be described as a diffusive process, we propose to include the correlation between $\mu_2(\Delta_t)$ and Δ_t as a feature (*diffusion adequation*).

In summary, we have 4 features based on moments of displacement: Motion type, Diffusion coefficient, Diffusion adequation, Mean squared displacement for $\Delta_t = 1$.

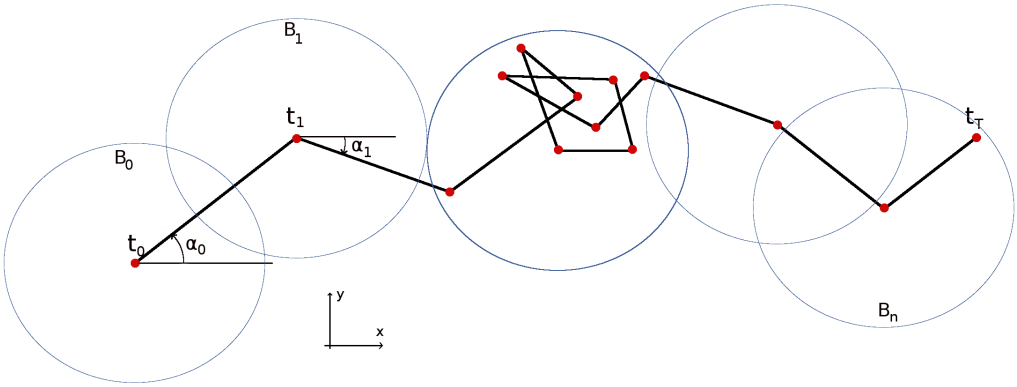

Figure B.1: Annotated cell trajectory

New features We also designed some novel features in order to describe the spatial trajectories:

- Area of the convex hull of the points of the trajectory. This feature mimics a traditional feature used in the cell migration field, where the temporal trajectory is measured without live-cell-imaging using phagokinetic tracks (traces that are left by migrating cells) (Naffar-Abu-Amara et al. 2008).

- Mean curvature: for each position t, an orthogonal regression is performed on $\{u_t, \ldots, u_{t+\Delta_t}\}$ using orthogonal distance regression ($\Delta_t = 10$). The mean curvature of the trajectory is the average of all regression sums of squares.

- Mean turning angle $\arctan(\frac{\Sigma \sin(\|\alpha_{t+1} - \alpha_t)\|}{\Sigma \cos(\|\alpha_{t+1} - \alpha_t)\|})$, where α_t are the angles as shown in figure B.1.

- Entropy features: we place a series of balls of fixed radius r_i, such that they cover the maximum number of consecutive data points. We repeat this until all points have been assigned to exactly one ball (see figure B.1 for illustration, details can be found in (Schoenauer Sebag et al. 2015a; Schoenauer Sebag et al. 2015b)). We then calculate the entropy of points inside the ball:

$$E_r = -\frac{1}{T} \sum_{B_r} \frac{card(B_r)}{T} \log(\frac{card(B_r)}{T}) \tag{B.4}$$

  with B_r the sets of points falling in the balls with radius r. The final features are E_r and the number of balls.

C Curriculum Vitae

Thomas Walter

Curriculum Vitae

27 rue des Acacias
75017 Paris, France
+33 (0)6.49.98.72.08
Thomas.Walter@mines-paristech.fr
members.cbio.mines-paristech.fr/~twalter/index.html
h-index (Google Scholar): 27

Education

2003 **PhD in Mathematical Morphology**, (with greatest honors).
Centre for Mathematical Morphology, Applied Mathematics, Mines ParisTech, France

1999 **Diploma in Electrical Engineering**, (with greatest honors).
Department of Telecommunications, Saarland University, Germany

Current Position

Since 2018 **Director of the Centre for Computational Biology**, *Mines ParisTech / ARMINES*, Paris, France.

Codirector of the department "Cancer and Genome: Bioinformatics, Biostatistics, Epidemiology of Complex Systems", *Institut Curie / Mines ParisTech / ARMINES / INSERM*, Paris, France.

Previous Positions

2015-2018 **Scientist, tenured (CR1)**, *Centre for Computational Biology, Mines ParisTech / ARMINES, associated with Curie Institute, INSERM (U900 joint laboratory on Bioinformatics and Computational Systems Biology of Cancer*, Paris, France.

2012-2015 **Scientist, Tenure Track (CR2)**, *Centre for Computational Biology, Mines ParisTech / ARMINES, associated with Curie Institute, INSERM (U900 joint laboratory on Bioinformatics and Computational Systems Biology of Cancer*, Paris, France.

2006-2012 **Scientist**, *European Molecular Biology Laboratory (EMBL)*, Heidelberg, Germany.

2003-2006 **Postdoc**, *Centre for Mathematical Morphology, Mines ParisTech / ARMINES*, Fontainebleau, France.

Membership of scientific societies and networks

Since 2019 **Advanced interdisciplinary Chair at PRAIRIE**, *Paris Artificial Intelligence Research Institute*.

Since 2017 **Executive Board of GDR ImaBio**, *national CNRS research group for light microscopy*.

2015-2019 **Work group leader France Bioimaging (FBI)**.

2013-2017 **IEEE Bio Imaging and Signal Processing Technical Committee**.

Supervision

since 2019 PhD student: Tristan Lazard, 50%

since 2018 PhD student: Arthur Imbert, 50%

since 2016 PhD student: Joseph Boyd, 50%

2015-2019	PhD student: Peter Naylor, 80%
2015-2018	PhD student: Aubin Samacoits, 50%
2013-2016	PhD student: Vaïa Machairas, 50%
2012-2015	PhD student: Alice Schoenauer Sebag, 80%
2014-2015	PostDoc: Xiwei Zhang, 100%
2020	Master Student: Guillaume Bataillon, 100%
2018	Master Student: Rémy Dubois, 50%
2015	Master student: Peter Naylor, 100%
2013	Master student: Denis Samuylov, 50%
Before 2012	3 master students

Teaching

Since 2018	Master Course on Deep Learning for Image Analysis (organizer, 40h Mines ParisTech)
Since 2014	Master Course on Genome Biology and Bioinformatics at Mines ParisTech (organizer, 40h, Mines ParisTech)
Since 2013	Master Course on Bioimage Informatics (lecturer, 8h, University of Strasbourg)
Since 2017	Master Course on Machine Learning for Biology and Chemistry (lecturer, 2h, PSL University)
Since 2018	MifoBio – summer school on functional microscopy (co-organizer, 40h, GDR ImaBio)
2008-2016	EMBO Course on High Throughput Microscopy for Systems Biology (lecturer, EMBL Heidelberg)
2015-2017	Molecular Biology of the Cell (lecturer, Institut Pasteur / Institut Curie)
2015	Image Analysis for Biologists (lecturer, University Cambridge)
2012	FBI-AT Course on Bioimage Informatics (lecturer, FBI)

Reviewing

Journals	Nature Methods, eLife, Nature Communications, Bioinformatics, BMC Bioinformatics, Molecular Systems Biology, PLOS Computational Biology, PLOS One, IEEE Transactions on Medical Imaging, Medical Image Analysis
Conferences	International Conference on Intelligent Systems for Molecular Biology (ISMB), International Symposium on Biomedical Imaging (ISBI), International Conference on Image Processing (ICIP), International Conference on Accoustic, Speech and Image Processing (ICASSP), International Conference on Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML)

Current and past collaborations

Bertrand group, IGMM, France

Zimmer group, Institut Pasteur, France

Pathology department, Institut Curie, France

Radiology department, Institut Curie, France

Reyal group, Institut Curie, France

Carazo Salas group, University of Bristol, UK

Centre for Mathematical Morphology, Mines ParisTech, France

Ellenberg group, EMBL Heidelberg, Germany

Barouki group, Université Descartes, France
Amblard group, Institute for Basic Science, South Corea
Almouzni group, Institut Curie, France
Gerlich group, Institute of Molecular Biotechnology (IMBA), Vienna, Austria

Languages

German	Native speaker
French	Near native
English	Near native
Spanish	Good command

Prizes and awards

2013 ranked 2nd in the DREAM challenge on toxicogenetics
`http://www.niehs.nih.gov/news/newsletter/2013/12/science-challenge/`

2014 ranked 2nd in the 2014 mitosis detection challenge MITOS-ATYPIA-14
`http://mitos-atypia-14.grand-challenge.org/results2`

2011 ranked 1st in the national recruitment competition of INRIA

Software

CellCognition (`http://www.cellcognition.org`): *core developer*
Scikit-image(`http://scikit-image.org`) : *core developer*

Grants

2019-2023 PRAIRIE (ANR), 450 k€
2020-2024 LUSTRA (ANR), 150 k€
2019-2023 TRANSFACT (ANR), ANR-19-CE12-0007-03, 150 k€
2013-2015 Systems Microscopy (NoE, EC), FP7/2007-2013, 258068, 200 k€
2014-2019 HI-FISH (ANR), ANR-14-CE10-0018-04, 60 k€
2019-2022 PhD fellowship Q-life (Tristan Lazard), 100 k€
2016-2019 PhD fellowship PSL (Joseph Boyd), 100 k€
2015-2018 PhD fellowship Ligue contre le cancer (Peter Naylor), 100 k€

Presentations (workshops, conferences, seminars), since 2012

2020/01/20 Journée thématique, Université Paris V, Paris: "Machine Learning for Computational Phenotyping" (invited talk)

2020/01/20 Quantitative Bioimaging, Oxford, UK: "Machine Learning for Computational Phenotyping: how to overcome the need for massive image annotation" (invited talk)

2019/12/07 ASCB, Washington DC, USA: "Machine Learning Methods for Exploring the Spatial Dimensions of Gene Expression" (invited talk)

2019/11/27 Colloque Données de santé et compétitivité : quels défis pour la technique et le droit ? - Le diagnostic par l'intelligence artificielle (invited talk)

2019/10/22 Danish Bioimaging Meeting, Aarhus, Denmark: "Machine Learning for Computational Phenotyping" (keynote lecture)

2019/09/18 Data Science Day, Paris, France: "Computer Vision for Bioimage Analysis" (invited talk)

2019/07/22 LifeTime Unconference, Barcelona, Spain: "Machine Learning approaches to explore the spatial dimension of gene expression" (contributed talk)

2019/07/03 JOBIM, Nantes, France: "Learning strategies to explore the spatial dimension of gene expression" (invited talk)

2019/06/28 Retraite PMS Sein, Paris, France: "Prediction of the response to neo-adjuvant chemotherapy from biopsies by artificial intelligence" (invited talk)

2019/06/18 Mines ParisTech Deep Learning Workshop, Fontainebleau, France: "A deep learning approach to identify mRNA localisation patterns" (invited talk)

2019/05/27 GDR ImaBio – Meeting with industrial partners, Paris, France: "Artificial intelligence for Microscopy – current developments and challenges" (invited talk)

2019/04/10 ISBI, Venice, Italy: "A deep learning approach to identify mRNA localisation patterns" (contributed talk)

2019/03/11 Mini-Symposium Demystifying Machine Learning for Microscopists, Toulouse, France: "Computer Vision for Cell segmentation and classification - Applications to histopathology and High Content Screening" (keynote lecture)

2018/11/30 LifeTime Meeting, Paris, France: "Disruptive Technologies: Machine Learning and Computer Vision for Biology" (invited talk)

2018/10/19 Artificial Intelligence in Biology and Health, Montpellier, France: "Machine Learning for computational phenotyping" (invited talk)

2018/10/11 MifoBio, Seignosse, France: "Machine Learning for Bioimaging in the context of High Content Screening" (invited talk)

2018/06/08 SIAM Conference on Imaging Science, Bologna, Italy: "Spatial patterns in large scale imaging data" (invited talk)

2018/05/24 OptDiag, Paris, France: "Big data approaches for computational phenotyping" (invited talk)

2017/06/27 SCANDEM, Copenhagen, Denmark: "Big data approaches for computational phenotyping" (invited talk)

2017/12/07 Journée intelligence artificielle, Paris, France: « Intelligence artificielle dans le domaine biomédicale : applications et défis » (invited talk)

2017/09/20 SLAS High-Content Screening Conference, Madrid, Spain: "Big data approaches for computational phenotyping" (invited talk)

2017/09/14 12th European Congress for Stereology and Image Analysis, Kaiserslautern, Germany: "Current Challenges in Bioimage Informatics" (keynote lecture)

2017/06/29 Imaging the cell, Rennes, France: "Big data approaches for computational phenotyping" (invited talk)

2016/11/03 Institut Curie Retreat Pict IbISA, Tours, France: "Exploring morphological phenotypes and spatial organization by large-scale screening approaches" (invited talk)

2016/10/21 EMBO Course on High Throughput Microscopy for Systems Biology, Heidelberg, Germany: "Bioimage Informatics for High Content Screening" (lecture)

2016/09/09 European Conference on Computer Vision (ECCV), Amsterdam, Netherlands: "Bioimage Informatics for phenomics" (invited talk)

2015/12/01 Microscopy Conference, London, UK: "Bioimage Informatics for Phenomics" (invited talk)

2015/09/23 Institut Curie Physical Chemistry Seminar, Paris, France: "Bioimage Informatics for Phenomics" (invited talk)

2015/09/17 France Bioimaging annual meeting, Paris, France: "Bioimage Informatics for Phenomics" (invited talk)

2015/07/07 JOBIM (Workshop on Biology, Computer Science and Mathematics), Clermont-Ferrand, France: "Bioimage Informatics for Phenomics" (keynote)

2015/06/09 Course: Image Analysis for Biology (EMBL), Heidelberg, Germany: "Computer Vision for Cellular Phenotyping" (lecture)

2015/04/13 Clore Center Workshop for Biological Physics, Weizman Institute, Rehovot, Israel: "Bioimage Informatics for Systems Microscopy" (invited talk)

2015/02/12 Statistical Methods for Postgenomic Data (SMPGD), Munich, Germany: "Bioimage Informatics for Cellular Phenotyping in High Content Screening and Histopathology" (contributed talk)

2015/01/20 Course: Molecular Biology of the Cell (Institut Pasteur/ Institut Curie), Paris, France: "Bioimage Informatics for High Content Screening" (lecture)

2015/01/09 Quantitative Bioimaging (QBI), Paris, France: "Quantitative phenotypic profiling for live cell imaging data in high content screening" (invited talk)

2014/11/17 2nd High Throughput Cell Biology: from screening to applications, Paris, France: "Quantitative phenotypic profiling for live cell imaging data in high content screening" (invited talk)

2014/10/24 EMBO Course on High Throughput Microscopy for Systems Biology, Heidelberg, Germany: "Bioimage Informatics for High Content Screening" (lecture)

2014/08/24 International Conference on Pattern Recognition (ICPR), Stockholm, Sweden: "Computational analysis of cellular phenotypes: from high-content-screening to histopathology" (contributed talk)

2014/07/01 Workshop on Bioimage Informatics & Modeling at the cellular scale, Toulouse, France: "Quantitative phenotypic profiling for live cell imaging data in High Content Screening" (invited talk)

2014/06/20 European Symposium on Biopathology, Paris, France: "Dissecting cancer relevant cellular processes by phenotypic profiling from live cell imaging data" (invited talk)

2013/11/06 CellCognition User Meeting, EMBL Heidelberg, Germany: "Tracking for CellCognition" (invited talk)

2013/07/05 FBI-AT Course on Bioimage Informatics: "From images to gene clusters: computational approaches for phenotypic profiling" (lecture)

2013/07/04 Colloquium of the French Society for Microscopy, Nantes, France: "Quantitative Phenotypic Profiling for live cell imaging data in High Content Screening" (invited talk)

2013/02/02 Finish Institute of Molecular Medicine, Helsinki, Finland: "Quantitative phenotypic profiling for live cell imaging data in High Content Screening" (invited seminar)

2012/10/05 Bioinformatics and Cancer, Paris, France: "Dissecting cancer relevant cellular processes by phenotypic profiling from live cell imaging data" (invited talk)

2012/10/20 EMBO Course on High Throughput Microscopy for Systems Biology, Heidelberg, Germany: "From images to gene clusters: computational approaches for phenotypic profiling" (lecture)

2012/09/17 Bioimage Informatics Conference, MPI Dresden, Germany: "Quantitative phenotypic profiling for live cell imaging data in the context of High Content Screening" (invited talk)

Publications

[1] Joseph C Boyd, Alice Pinheiro, Elaine Del Nery, Fabien Reyal, et al. "Domain-Invariant Features for Mechanism of Action Prediction in a Multi-Cell-Line Drug Screen". In: *Bioinformatics* 36.5 (Oct. 14, 2019), pp. 1607–1613. DOI: 10.1093/bioinformatics/btz774.

[2] Rémy Dubois, Arthur Imbert, Aubin Samacoits, Marion Peter, et al. "A Deep Learning Approach to Identify mRNA Localization Patterns". In: *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 1386–1390. DOI: 10.1109/ISBI.2019.8759235.

[3] Mélanie Durand, Thomas Walter, Tiphène Pirnay, Thomas Naessens, et al. "Human Lymphoid Organ cDC2 and Macrophages Play Complementary Roles in T Follicular Helper Responses". In: *The Journal of Experimental Medicine* 216.7 (July 1, 2019), 1561 LP –1581. DOI: 10.1084/jem.20181994.

[4] Inna Kuperstein and Emmanuel Barillot, eds. *Computational Systems Biology Approaches in Cancer Research*. Chapman and Hall/CRC, 2019.

[5] P Naylor, J Boyd, M Laé, F Reyal, et al. "Predicting Residual Cancer Burden In A Triple Negative Breast Cancer Cohort". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 933–937. DOI: 10.1109/ISBI.2019.8759205.

[6] Peter Naylor, Marick La, Fabien Reyal, and Thomas Walter. "Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map". In: *IEEE Transactions on Medical Imaging* 38.2 (2019), pp. 448–459. DOI: 10.1109/TMI.2018.2865709.

[7] Joseph Boyd, Alice Pinhiero, Elaine D. Nery, Fabien Reyal, et al. "Analysing Double-Strand Breaks in Cultured Cells for Drug Screening Applications by Causal Inference". In: *International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*. Vol. 2018-April. 2018, pp. 445–448. DOI: 10.1109/ISBI.2018.8363612.

[8] Racha Chouaib, Adham Safieddine, Xavier Pichon, Oh Sung Kwon, et al. *A Localization Screen Reveals Translation Factories and Widespread Co-Translational Protein Targeting*. SSRN Scholarly Paper ID 3300043. Rochester, NY: Social Science Research Network, Dec. 12, 2018.

[9] Aubin Samacoits, Racha Chouaib, Adham Safieddine, Abdel-Meneem Traboulsi, et al. "A Computational Framework to Study Sub-Cellular RNA Localization". In: *Nature Communications* 9.1 (2018), p. 4584. DOI: 10.1038/s41467-018-06868-w. pmid: 30389932.

[10] Elsa Bernard, Yunlong Jiao, Erwan Scornet, Véronique Stoven, et al. "Kernel Multitask Regression for Toxicogenetics". In: *Molecular Informatics* 36 (Jan. 1, 2017).

[11] Peter Naylor, Marick Lae, Fabien Reyal, and Thomas Walter. "Nuclei Segmentation in Histopathology Images Using Deep Neural Networks". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (2017), IEEE, EMB, IEEE Signal Proc Soc. DOI: 10.1109/ISBI.2017.7950669.

[12] Mayumi Isokane, Thomas Walter, Robert Mahen, Bianca Nijmeijer, et al. "ARHGEF17 Is an Essential Spindle Assembly Checkpoint Factor That Targets Mps1 to Kinetochores". In: *The Journal of Cell Biology* 212.6 (Mar. 14, 2016), pp. 647–659. DOI: 10.1083/jcb.201408089.

[13] Vaïa Machairas, Thérèse Baldeweck, Thomas Walter, and Etienne Décencière. "New General Features Based on Superpixels for Image Segmentation Learning". In: *International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*. 2016, (accepted for publication).

[14] Nikolay Tsanov, Aubin Samacoits, Racha Chouaib, Abdel Meneem Traboulsi, et al. "SmiFISH and FISH-Quant - A Flexible Single RNA Detection Approach with Super-Resolution Capability". In: *Nucleic Acids Research* 44.22 (2016). DOI: 10.1093/nar/gkw784. pmid: 27599845.

[15] Federica Eduati, Lara M Mangravite, Tao Wang, Hao Tang, et al. "Prediction of Human Population Responses to Toxic Compounds by a Collaborative Competition". In: *Nature Biotechnology* 33 (Aug. 10, 2015), p. 933.

[16] Vaïa Machairas, Etienne Decencière, and Thomas Walter. "Spatial Repulsion Between Markers Improves Watershed Performance". In: *International Symposium of Mathematical Morphology*. 2015.

[17] Vaïa Machairas, Etienne Decencière, and Thomas Walter. "Spatial Repulsion Between Markers Improves Watershed Performance BT - Mathematical Morphology and Its Applications to Signal and Image Processing". In: ed. by Jón Atli Benediktsson, Jocelyn Chanussot, Laurent Najman, and Hugues Talbot. Cham: Springer International Publishing, 2015, pp. 194–202.

[18] Vaïa Machairas, Matthieu Faessel, David Cárdenas-peña, Théodore Chabardes, et al. "Waterpixels". In: *IEEE Transactions on Image Processing* 24.11 (2015), pp. 3707–3716.

[19] A. Schoenauer Sebag, S. Plancade, C. Raulet-Tomkiewicz, R. Barouki, et al. "A Generic Methodological Framework for Studying Single Cell Motility in High-Throughput Time-Lapse Data". In: *Bioinformatics* 31.12 (2015), pp. i320–i328. DOI: 10.1093/bioinformatics/btv225.

[20] Alice Schoenauer Sebag, Sandra Plancade, Céline Raulet-Tomkiewicz, Robert Barouki, et al. "Infering an Ontology of Single Cell Motions from High-Throughput Microscopy Data". In: *Proceedings of the 12th IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*. New York, New York, USA, 2015, pp. 160–163.

[21] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, et al. "A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms." In: *Nature biotechnology* 32.12 (2014), pp. 1–103. DOI: 10.1038/nbt.2877. pmid: 24880487.

[22] Veronika Graml, Xenia Studera, Jonathan L D Lawson, Anatole Chessel, et al. "A Genomic Multiprocess Survey of Machineries That Control and Link Cell Shape, Microtubule Organization, and Cell-Cycle Progression". In: *Developmental Cell* 31.2 (2014), pp. 227–239. DOI: 10.1016/j.devcel.2014.09.005.

[23] Jean-Karim Hériché, Jon G Lees, Ian Morilla, Thomas Walter, et al. "Integration of Biological Data by Kernels on Graph Nodes Allows Prediction of New Genes Involved in Mitotic Chromosome Condensation". In: *Molecular Biology of the cell* 25 (2014), pp. 2522–2536. DOI: 10.1091/mbc.E13-04-0221.

[24] Vaïa Machairas, Etienne Decencière, and Thomas Walter. "Waterpixels: Superpixels Based on the Watershed Transformation". In: *International Conference on Image Processing (ICIP)*. 2014, pp. 4343–4347.

[25] Justus Tegha-Dunghu, Elena Bausch, Beate Neumann, Annelie Wuensche, et al. "MAP1S Controls Microtubule Stability throughout the Cell Cycle in Human Cells". In: *Journal of Cell Science* 127.23 (Dec. 1, 2014), 5007 LP –5013. DOI: 10.1242/jcs.136457.

[26] Mitko Veta, Paul J Van Diest, Stefan M Willems, Haibo Wang, et al. "Assessment of Algorithms for Mitosis Detection in Breast Cancer Histopathology Images". In: *Medical Image Analysis* (2014), pp. 1–23.

[27] Gregoire Pau, Thomas Walter, Beate Neumann, Jean-karim Hériché, et al. "Dynamical Modelling of Phenotypes in a Genome-Wide RNAi Live-Cell Imaging Assay". In: *BMC bioinformatics* 14.308 (2013), pp. 1–10. DOI: 10.1186/1471-2105-14-308.

[28] Moritz Mall, Thomas Walter, Mátyás Gorjánácz, Iain F Davidson, et al. "Mitotic Lamin Disassembly Is Triggered by Lipid-Mediated Signaling." In: *The Journal of cell biology* 198.6 (Sept. 17, 2012), pp. 981–90. DOI: 10.1083/jcb.201205103. pmid: 22986494.

[29] Christian Conrad, Annelie Wünsche, Tze Heng Tan, Jutta Bulkescher, et al. "Micropilot: Automation of Fluorescence Microscopy-Based Imaging for Systems Biology." In: *Nature methods* 8.3 (Mar. 2011), pp. 246–9. DOI: 10.1038/nmeth.1558. pmid: 21258339.

[30] Seán I O Donoghue, Anne-claude Gavin, Nils Gehlenborg, David S Goodsell, et al. "Visualizing Biological Data — Now and in the Future". In: *Nature Methods* 7 (3s 2010), S2–S4. DOI: 10.1038/nmeth0310-S2.

[31] Bénédicte Dupas, Thomas Walter, Ali Erginay, John-Richard Ordonez, et al. "Evaluation of Automated Fundus Photograph Analysis Algorithms for Detecting Microaneurysms, Haemorrhages and Exudates, and of a Computer-Assisted Diagnostic System for Grading Diabetic Retinopathy." In: *Diabetes & metabolism* 36.3 (June 2010). Place: France, pp. 213–220. DOI: 10.1016/j.diabet.2010.01.002. pmid: 20219404.

[32] Michael Held, Michael Schmitz, Bernd Fischer, Thomas Walter, et al. "CellCognition: Time-Resolved Phenotype Annotation in High-Throughput Live Cell Imaging." In: *Nature methods* 7.9 (Sept. 2010), pp. 747–54. DOI: 10.1038/nmeth.1486. pmid: 20693996.

[33] Stefan Terjung, Thomas Walter, Arne Seitz, Beate Neumann, et al. "High-Throughput Microscopy Using Live Mammalian Cells." In: *Live Cell Imaging: A Laboratory Manual*. Ed. by R.D. Goldman, Jason R. Swedlow, and David L. Spector. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2010.

[34] Thomas Walter, Michael Held, Beate Neumann, Jean-Karim Hériché, et al. "Automatic Identification and Clustering of Chromosome Phenotypes in a Genome Wide RNAi Screen by Time-Lapse Imaging." In: *Journal of structural biology* 170.1 (Apr. 2010), pp. 1–9. DOI: 10.1016/j.jsb.2009.10.004. pmid: 19854275.

[35] Thomas Walter, Beate Neumann, Jean-Karim Hériché, Jutta Bulkescher, et al. "Phenotypic Profiling of the Human Genome by Time-Lapse Microscopy Reveals Cell Division Genes." In: *Nature* 464.7289 (Apr. 1, 2010), pp. 721–7. DOI: 10.1038/nature08869. pmid: 20360735.

[36] Thomas Walter, David W Shattuck, Richard Baldock, Mark E Bastin, et al. "Visualization of Image Data from Cells to Organisms". In: *Nature Methods* 7.3 (2010), S26–S41. DOI: 10.1038/NmEtH.1431.

[37] Justus Tegha-Dunghu, Beate Neumann, Simone Reber, Roland Krause, et al. "EML3 Is a Nuclear Microtubule-Binding Protein Required for the Correct Alignment of Chromosomes in Metaphase". In: *Journal of Cell Science* 121.10 (May 15, 2008), 1718 LP –1726. DOI: 10.1242/jcs.019174.

[38] Thomas Walter, Michael Held, Beate Neumann, Jean Karim Hériché, et al. "A Genome Wide RNAi Screen by Time Lapse Microscopy in Order to Identify Mitotic Genes - Computational Aspects and Challenges". In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI* (2008), pp. 328–331. DOI: 10.1109/ISBI.2008.4540999.

[39] Eric Denion, John-Richard Ordonez, Jean-Claude Klein, Agnes Glacet-Bernard, et al. "Redistribution of the Neurosensory Retina in Inferior Limited Macular Translocation: An Evaluation Using Image Registration." In: *Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie* 245.3 (Mar. 2007). Place: Germany, pp. 437–442. DOI: 10.1007/s00417-006-0408-1. pmid: 16944187.

[40] Holger Erfle, Beate Neumann, Urban Liebel, Phill Rogers, et al. "Reverse Transfection on Cell Arrays for High Content Screening Microscopy". In: *Nature Protocols* 2.2 (Feb. 2007), pp. 392–399. DOI: 10.1038/nprot.2006.483.

[41] Gabin Sihn, Thomas Walter, Jean-Claude Klein, Isabelle Queguiner, et al. "Anti-Angiogenic Properties of Myo-Inositol Trispyrophosphate in Ovo and Growth Reduction of Implanted Glioma." In: *FEBS letters* 581.5 (Mar. 6, 2007), pp. 962–6. DOI: 10.1016/j.febslet.2007.01.079. pmid: 17316624.

[42] Thomas Walter, Pascale Massin, Ali Erginay, Richard Ordonez, et al. "Automatic Detection of Microaneurysms in Color Fundus Images." In: *Medical Image Analysis* 11.6 (2007), pp. 555–66.

[43] Thomas Walter and Jean-Claude Klein. "Automatic Analysis of Color Fundus Photographs and Its Application to the Diagnosis of Diabetic Retinopathy BT - Handbook of Biomedical Image Analysis: Volume II: Segmentation Models Part B". In: ed. by Jasjit S Suri, David L Wilson, and Swamy Laxminarayan. Boston, MA: Springer US, 2005, pp. 315–368. DOI: 10.1007/0-306-48606-7_7.

[44] Thomas Walter, Richard Ordonez, and Jean-Claude Klein. "A Morphological Approach for Skeleton Filtering with Reconstruction of the Relevant Branches". In: 10th Computer Vision Winter Workshop. Zell, Austria, Feb. 2005.

[45] Thomas Walter. "Application de la Morphologie Mathématique au diagnostic de la Rétinopathie Diabétique à partir d'images couleur". PhD thesis. Fontainebleau, France: Mines ParisTech, Sept. 2003. 228 pp.

[46] Thomas Walter and Jean-Claude Klein. "A Computational Approach to Diagnosis of Diabetic Retinopathy". In: *6th Conference on Systemics, Cybernetics and Informatics (SCI)*. 2002, pp. 521–526.

[47] Thomas Walter and Jean-Claude Klein. "Automatic Detection of Microaneurysms in Color Fundus Images of the Human Retina by Means of the Bounding Box Closing". In: International Symposium on Medical Data Analysis (ISMDA). Ed. by Alfredo Colosimo, Paolo Sirabella, and Alessandro Giuliani. Vol. 2526. Lecture Notes in Computer Science (LNCS). Rome, Italy: Springer Berlin Heidelberg, 2002, pp. 210–220.

[48] Thomas Walter, Jean-claude Klein, Pascale Massin, and Ali Erginay. "A Contribution of Image Processing to the Diagnosis of Diabetic Retinopathy — Detection of Exudates in Color Fundus Images of the Human Retina". In: *IEEE Transactions on Medical Imaging* 21.10 (2002), pp. 1236–1243. DOI: 10.1109/TMI.2002.806290.

[49] Thomas Walter and Jean-Claude Klein. "Segmentation of Color Fundus Images of the Human Retina : Detection of the Optic Disc and the Vascular Tree Using Morphological Techniques". In: International Symposium on Medical Data Analysis (ISMDA). Ed. by Jose Crespo, Maojo, Victor, and Fernando Martin. Vol. 2199. Lecture Notes in Computer Science (LNCS). Madrid, Spain: Springer Berlin Heidelberg, 2001, pp. 282–287. DOI: 10.1007/3-540-45497-7_43.

[50] Thomas Walter, Jean-Claude Klein, Pascale Massin, and Frédéric Zana. "Automatic Segmentation and Registration of Retinal Fluorescein Angiographies - Application to Diabetic Retinopathy". In: First International Workshop on Computer Assisted Fundus Image Analysis (CAFIA). Copenhagen, Denmark, May 2000.

[51] Thomas Walter. "Gewinnung von Merkmalen transitorisch evozierter otoakustischer Emissionen". Diploma Thesis. Saarbrücken, Germany: Saarland university, Oct. 1999. 156 pp.

D Selected articles

I have selected the following articles from my publication list in order to represent the different aspects of my work:

- Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes (2010): https://cloud.mines-paristech.fr/index.php/s/kBBY0MCc8v0kLu4
- A generic methodological framework for studying single cell motility in high-throughput time-lapse data (2015): https://cloud.mines-paristech.fr/index.php/s/tqvr3MHEF43PjNT
- A computational framework to study sub-cellular RNA localization (2018): https://cloud.mines-paristech.fr/index.php/s/D9QsQ1OIRnnnf61
- Segmentation of Nuclei in Histopathology Images by deep regression of the distance map (2018): https://cloud.mines-paristech.fr/index.php/s/6Wkf6CNIhbHcqEi

Bibliography

1000 Genomes Project Consortium (Oct. 2010). "A map of human genome variation from population-scale sequencing." *Nature* 467.7319, pp. 1061–73. DOI: 10.1038/nature09534.

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., et al. (2018). "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices". *npj Digital Medicine* 1.1, p. 39. DOI: 10.1038/s41746-018-0040-6.

Achim, K., Pettit, J.-b., Saraiva, L. R., Gavriouchkina, D., et al. (2015). "High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin". *Nature biotechnology* 33.5, pp. 503–509. DOI: 10.1038/nbt.3209.

Adams, C. L., Kutsyy, V., Coleman, D. A., Cong, G., et al. (2006). "Compound Classification Using Image-Based Cellular Phenotypes". *Methods in Enzymology*. Vol. 414. 06, pp. 440–468. DOI: 10.1016/S0076-6879(06)14024-0.

Alessandro, B., Giulio, I., Onofri, L., and Peng, H. (2016). "TeraFly : real-time three-dimensional visualization and annotation of terabytes of multidimensional volumetric images". *Nature Methods* 13.3, pp. 192–194.

Antal, B., Chessel, A., and Carazo Salas, R. E. (2015). "Mineotaur: A tool for high-content microscopy screen sharing and visual analytics". *Genome Biology* 16.1, pp. 1–5. DOI: 10.1186/s13059-015-0836-5.

Battich, N., Stoeger, T., and Pelkmans, L. (2013). "Image-based transcriptomics in thousands of single human cells at single-molecule resolution". *Nature Methods* 10.11. DOI: 10.1038/nmeth.2657.

Bejnordi, B. E., Veta, M., Diest, P. J. v., Ginneken, B. v., et al. (Dec. 12, 2017). "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer". *JAMA* 318.22, pp. 2199–2210. DOI: 10.1001/jama.2017.14585.

Belthangady, C. and Royer, L. A. (July 8, 2019). "Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction". *Nature Methods*, pp. 1–11. DOI: 10.1038/s41592-019-0458-z.

Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., et al. (Aug. 9, 2019). "Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology". *Nature Reviews Clinical Oncology*, pp. 1–13. DOI: 10.1038/s41571-019-0252-y.

Bernard, E., Jiao, Y., Scornet, E., Stoven, V., et al. (Jan. 1, 2017). "Kernel multitask regression for toxicogenetics". *Molecular Informatics* 36.

Betzig, E., Patterson, G. H., Sougrat, R., Lindwasser, O. W., et al. (Sept. 2006). "Imaging intracellular fluorescent proteins at nanometer resolution." *Science (New York, N.Y.)* 313.5793, pp. 1642–5. DOI: 10.1126/science.1127344.

Beucher, S. and Lantuéjoul, C. (1979). "Use of watershed in contour detection". *International Workshop on image processing: real-time Edge and Motion detection/estimation*.

Boland, M. V., Markey, M. K., and Murphy, R. F. (1998). "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images". *Cytometry* 33, pp. 366–375. DOI: 10.1002/(SICI)1097-0320(19981101)33:3<366::AID-CYTO12>3.0.CO;2-R.

Boyd, J., Pinhiero, A., Nery, E., Reyal, F., et al. (2018). "Analysing double-strand breaks in cultured cells for drug screening applications by causal inference". *Proceedings - International Symposium on Biomedical Imaging* 2018-Apr.Isbi, pp. 445–448. DOI: 10.1109/ISBI.2018.8363612.

Boyd, J. C., Pinheiro, A., Del Nery, E., Reyal, F., et al. (Oct. 14, 2019). "Domain-Invariant Features for Mechanism of Action Prediction in a Multi-Cell-Line Drug Screen". *Bioinformatics* 36.5, pp. 1607–1613. DOI: 10.1093/bioinformatics/btz774.

Bray, M.-A., Carpenter, A., Imaging Platform, B. I. o. M., and Harvard (2012). "Advanced Assay Development Guidelines for Image-Based High Content Screening and Analysis". *Assay Guidance Manual*. Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences. Chap. Advanced A, pp. 619–652.

Brenner, S. (Jan. 12, 2010). "Sequences and Consequences." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365.1537, pp. 207–12. DOI: 10.1098/rstb.2009.0221. pmid: 20008397.

Buxbaum, A. R., Haimovich, G., and Singer, R. H. (2014). "In the right place at the right time: visualizing and understanding mRNA localization". *Nature reviews. Molecular cell biology* 16.2, pp. 95–109. DOI: 10.1038/nrm3918.

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., et al. (2017). "Data-analysis strategies for image-based cell profiling". *Nature Methods* 14.9, pp. 849–863. DOI: 10.1038/nmeth.4397.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., et al. (Aug. 2019). "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". *Nature Medicine* 25.8, pp. 1301–1309. DOI: 10.1038/s41591-019-0508-1.

Cancer, I. N. du (2017). *Les cancers en France*. 2017th ed. Springer-Verlag New York, Inc.

Cardona, A. and Tomancak, P. (July 2012). "Current challenges in open-source bioimage informatics." *Nature methods* 9.7, pp. 661–5. DOI: 10.1038/nmeth.2082.

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., et al. (Jan. 2006). "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." *Genome biology* 7.10, R100. DOI: 10.1186/gb-2006-7-10-r100.

Carpenter, A. E. and Sabatini, D. M. (Jan. 2004). "Systematic genome-wide screens of gene function". *Nature Reviews Genetics* 5.1, pp. 11–22. DOI: 10.1038/nrg1248.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., et al. (1994). "Green Fluorescent Protein as a Marker for Gene Expression". *Science* 263.2, pp. 802–805.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., et al. (2015). "Spatially resolved, highly multiplexed RNA profiling in single cells". *Science* 348.6233, pp. 412–425. DOI: 10.1126/science.aaa6090.

Chenouard, N., Smal, I., Chaumont, F. de, Maska, M., et al. (Mar. 2014). "Objective comparison of particle tracking methods". *Nat Meth* 11.3, pp. 281–289.

Chin, A. and Lécuyer, E. (Nov. 1, 2017). "RNA localization: Making its way to the center stage". *Biochimica et Biophysica Acta (BBA) - General Subjects*. Biochemistry of Synthetic Biology - Recent Developments 1861.11, pp. 2956–2970. DOI: 10.1016/j.bbagen.2017.06.011.

Chouaib, R., Safieddine, A., Pichon, X., Kwon, O. S., et al. (Dec. 12, 2018). *A Localization Screen Reveals Translation Factories and Widespread Co-Translational Protein Targeting*. SSRN Scholarly Paper ID 3300043. Rochester, NY: Social Science Research Network.

Chow, K.-H., Factor, R. E., and Ullman, K. S. (Mar. 2012). "The nuclear envelope environment and its cancer connections." *Nature reviews. Cancer* 12.3, pp. 196–209. DOI: 10.1038/nrc3219.

Christiansen, E. M., Yang, S. J., Ando, D. M., Javaherian, A., et al. (2018). "In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images". *Cell* 173.3. Publisher: Elsevier Inc. ISBN: 8415683111, 792–803.e19. DOI: 10.1016/j.cell.2018.03.040. arXiv: cond-mat/0411586.

Coelho, L. P., Peng, T., and Murphy, R. F. (2010). "Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing". *Bioinformatics* 26, pp. 7–12. DOI: 10.1093/bioinformatics/btq220.

Collinet, C., Stöter, M., Bradshaw, C. R., Samusik, N., et al. (Mar. 2010). "Systems survey of endocytosis by multiparametric image analysis." *Nature* 464.7286, pp. 243–9. DOI: 10.1038/nature08779.

Conrad, C., Wunsche, A., Tan, T. H., Bulkescher, J., et al. (2011). "Micropilot: automation of fluorescence miscroscopy-based imaging for systems biology". *Nature Methods* 8.3, pp. 246–249. DOI: 10.1038/nmeth.1558.

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., et al. (2014). "A community effort to assess and improve drug sensitivity prediction algorithms." *Nature biotechnology* 32.12, pp. 1–103. DOI: 10.1038/nbt.2877.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., et al. (Oct. 2018). "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning". *Nature Medicine* 24.10, pp. 1559–1567. DOI: 10.1038/s41591-018-0177-5.

Courtiol, P., Tramel, E. W., Sanselme, M., and Wainrib, G. (2017). "Classification and disease localization in histopathology using only global labels: a weakly supervised approach". *CoRR*, pp. 1–13. arXiv: 1802.02212v1.

Danuser, G. (2011). "Computer Vision in Cell Biology". *Cell* 147.5, pp. 973–978. DOI: 10.1016/j.cell.2011.11.001.

Dao, D., Fraser, A. N., Hung, J., Ljosa, V., et al. (2016). "CellProfiler Analyst : interactive data exploration , analysis and classification of large biological image sets". 32 (June), pp. 3210–3212. DOI: 10.1093/bioinformatics/btw390.

Donà, E., Barry, J. D., Valentin, G., Quirin, C., et al. (Sept. 2013). "Directional tissue migration through a self-generated chemokine gradient". *Nature* 503, p. 285. DOI: 10.1038/nature12635.

Dubois, R., Imbert, A., Samacoits, A., Peter, M., et al. (2019). "A deep learning approach to identify mRNA localization patterns". *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1386–1390. DOI: 10.1109/ISBI.2019.8759235.

Durand, M., Walter, T., Pirnay, T., Naessens, T., et al. (July 2019). "Human Lymphoid Organ cDC2 and Macrophages Play Complementary Roles in T Follicular Helper Responses". en. *Journal of Experimental Medicine* 216.7, pp. 1561–1581. DOI: 10.1084/jem.20181994.

Durand, T., Thome, N., and Cord, M. (June 2016). "WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, pp. 4743–4752. DOI: 10.1109/CVPR.2016.513.

Dürr, O. and Sick, B. (2016). "Single-cell phenotype classification using deep convolutional neural networks". *Journal of Biomolecular Screening* 21.9, pp. 998–1003. DOI: 10.1177/1087057116631284.

Echeverri, C. J. and Perrimon, N. (May 2006). "High-throughput RNAi screening in cultured cells: a user's guide." *Nature reviews. Genetics* 7.5, pp. 373–84. DOI: 10.1038/nrg1836.

Eduati, F., Mangravite, L. M., Wang, T., Tang, H., et al. (Aug. 10, 2015). "Prediction of human population responses to toxic compounds by a collaborative competition". *Nature Biotechnology* 33. Publisher: The Author(s), p. 933.

Elston, C. W. and Ellis, O. (1991). "Pathological prognostic factors in breast cancer . I . The value of histological grade in breast cancer : experience from a large study with long-term follow-up". *Histopathology* 19, pp. 403–410.

ENCODE Project Consortium (June 2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." *Nature* 447.7146, pp. 799–816. DOI: 10.1038/nature05874.

ENCODE Project Consortium (2012). "An Integrated Encyclopedia of DNA Elements in the Human Genome". *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247.An.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., et al. (2019). "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". *Nature* 568.7751, pp. 235–239. DOI: 10.1038/s41586-019-1049-y.

Erfle, H., Neumann, B., Liebel, U., Rogers, P., et al. (Feb. 2007). "Reverse transfection on cell arrays for high content screening microscopy". *Nature Protocols* 2.2, pp. 392–399. DOI: 10.1038/nprot.2006.483.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., et al. (2019). "U-Net: deep learning for cell counting, detection, and morphometry". *Nature Methods* 16.1, pp. 67–70. DOI: 10.1038/s41592-018-0261-2.

Fazal, F. M., Han, S., Parker, K. R., Kaewsapsak, P., et al. (July 11, 2019). "Atlas of Subcellular RNA Localization Revealed by APEX-Seq". *Cell* 178.2, 473–490.e26. DOI: 10.1016/j.cell.2019.05.027.

Fowlkes, C. C., Hendriks, C. L. L., Keränen, S. V. E., Weber, G. H., et al. (Apr. 2008). "A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm." *Cell* 133.2, pp. 364–74. DOI: 10.1016/j.cell.2008.01.053.

Fuchs, F., Pau, G., Kranz, D., Sklyar, O., et al. (June 2010). "Clustering phenotype populations by genome-wide RNAi and multiparametric imaging." *Molecular systems biology* 6.370, p. 370. DOI: 10.1038/msb.2010.25.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., et al. (2016). "Domain-Adversarial Training of Neural Networks". *Journal of Machine Learning Research* 17.17. ISBN: 15324435, pp. 1–35. DOI: 10.1088/1475-7516/2015/08/013. arXiv: 1505.07818.

Garcia-Canton, C., Anadon, A., and Meredith, C. (2013). "Assessment of the in vitro γH2AX assay by High Content Screening as a novel genotoxicity test". *Mutation Research - Genetic Toxicology and Environmental Mutagenesis* 757.2, pp. 158–166. DOI: 10.1016/j.mrgentox.2013.08.002.

Glory, E. and Murphy, R. F. (Jan. 2007). "Automated subcellular location determination and high-throughput microscopy." *Developmental cell* 12.1, pp. 7–16. DOI: 10.1016/j.devcel.2006.12.007.

Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W., et al. (Feb. 15, 2017). "A multi-scale convolutional neural network for phenotyping high-content cellular images". *Bioinformatics* 33.13, pp. 2010–2019. DOI: 10.1093/bioinformatics/btx069.

Goldberg, I. G., Allan, C., Burel, J.-M., Creager, D., et al. (Jan. 2005). "The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging." *Genome biology* 6.5, R47. DOI: 10.1186/gb-2005-6-5-r47.

Graml, V., Studera, X., Lawson, J. L. D., Chessel, A., et al. (2014). "A Genomic Multiprocess Survey of Machineries that control and Link Cell Shape, Microtubule Organization, and Cell-Cycle Progression". *Developmental Cell* 31.2, pp. 227–239. DOI: 10.1016/j.devcel.2014.09.005.

Guirao, B., Rigaud, S. U., Bosveld, F., Bailles, A., et al. (2015). "Unified quantitative characterization of epithelial tissue development". *eLife* 4.e08519, pp. 1–52. DOI: 10.7554/eLife.08519.

Guyon, I. and Elisseeff, A. (Mar. 2003). "An Introduction to Variable and Feature Selection". *Journal of Machine Learning Research* 3, pp. 1157–1182.

Halpern, K. B., Shenhav, R., Matcovitch-Natan, O., Tóth, B., et al. (Feb. 2017). "Single-Cell Spatial Reconstruction Reveals Global Division of Labour in the Mammalian Liver". *Nature* 542.7641, pp. 352–356. DOI: 10.1038/nature21065.

Hanahan, D. and Weinberg, R. A. (2011). "Hallmarks of cancer: the next generation." *Cell* 144.5. Publisher: The Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, EPFL, Lausanne CH-1015, Switzerland. dh@epfl.ch, pp. 646–674.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). "Textural features for image classification". *IEEE Transactions on Systems, Man and Cybernetics* SMC-3.6, pp. 610–621.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (Mar. 20, 2017). "Mask R-CNN". *arXiv:1703.06870 [cs]*. arXiv: 1703.06870.

Held, M., Schmitz, M. H. a., Fischer, B., Walter, T., et al. (Sept. 2010). "CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging." *Nature methods* 7.9, pp. 747–54. DOI: 10.1038/nmeth.1486.

Hell, S. W. (Nov. 2003). "Toward fluorescence nanoscopy." *Nature biotechnology* 21.11, pp. 1347–55. DOI: 10.1038/nbt895.

Hériché, J.-k., Lees, J. G., Morilla, I., Walter, T., et al. (2014). "Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation". *Molecular Biology of the cell* 25, pp. 2522–2536. DOI: 10.1091/mbc.E13-04-0221.

Hoehndorf, R., Harris, M. a., Herre, H., Rustici, G., et al. (July 2012). "Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology." *Bioinformatics (Oxford, England)* 28.13, pp. 1783–9. DOI: 10.1093/bioinformatics/bts250.

Hollandi, R., Szkalisity, A., Toth, T., Tasnadi, E., et al. (Jan. 1, 2019). "A Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer". *bioRxiv*, p. 580605. DOI: 10.1101/580605.

Homeyer, A., Schenk, A., Arlt, J., Dahmen, U., et al. (June 1, 2013). "Practical quantification of necrosis in histological whole-slide images". *Computerized Medical Imaging and Graphics* 37.4, pp. 313–322. DOI: 10.1016/j.compmedimag.2013.05.002.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). "Phenomics : the next challenge". *Nature Reviews Genetics* 11.12, pp. 855–866. DOI: 10.1038/nrg2897.

Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J., et al. (Aug. 2004). "Optical sectioning deep inside live embryos by selective plane illumination microscopy." *Science (New York, N.Y.)* 305.5686, pp. 1007–9. DOI: 10.1126/science.1100035.

International Cancer Genome Consortium (Apr. 2010). "International network of cancer genome projects." *Nature* 464.7291, pp. 993–8. DOI: 10.1038/nature08987.

International Human Genome Sequencing Consortium (Feb. 2001). "Initial sequencing and analysis of the human genome". *Nature* 409, pp. 860–921.

International Human Genome Sequencing Consortium (Oct. 2004). "Finishing the euchromatic sequence of the human genome." *Nature* 431.7011, pp. 931–945.

Irshad, H., Veillard, A., Roux, L., and Racoceanu, D. (2014). "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential". *IEEE Reviews in Biomedical Engineering* 7, pp. 97–114. DOI: 10.1109/RBME.2013.2295804.

Isokane, M., Walter, T., Mahen, R., Nijmeijer, B., et al. (Mar. 2016). "ARHGEF17 is an essential spindle assembly checkpoint factor that targets Mps1 to kinetochores". *The Journal of Cell Biology* 212.6, pp. 647–659. DOI: 10.1083/jcb.201408089.

Ivashkevich, A., Redon, C. E., Nakamura, A. J., Martin, R. F., et al. (2012). "Use of the γ-H2AX assay to monitor DNA damage and repair in translational cancer research". *Cancer Letters* 327.1-2, pp. 123–133. DOI: 10.1016/j.canlet.2011.12.025. arXiv: NIHMS150003.

Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., et al. (2008). "Robust single-particle tracking in live-cell time-lapse sequences". *Nature methods* 5.8, pp. 695–702. DOI: 10.1038/NMETH.1237.

Jones, T. R., Kang, I. H., Wheeler, D. B., Lindquist, R. a., et al. (Jan. 2008). "CellProfiler Analyst: data exploration and analysis software for complex image-based screens." *BMC bioinformatics* 9, p. 482. DOI: 10.1186/1471-2105-9-482.

Jug, F., Pietzsch, T., Preibisch, S., and Tomancak, P. (2014). "Bioimage Informatics in the context of Drosophila research". *Methods* 68.1, pp. 60–73. DOI: 10.1016/j.ymeth.2014.04.004.

Kervrann, C., Óscar, C., Sorzano, S., Acton, S. T., et al. (2016). "A Guided Tour of Selected Image Processing and Analysis Methods for Fluorescence and Electron Microscopy". *IEEE Journal of Selected Topics in Signal Processing* 10.1, pp. 6–30. DOI: 10.1109/JSTSP.2015.2505402.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., et al. (July 2017). "A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology". *IEEE Transactions on Medical Imaging* 36.7, pp. 1550–1560. DOI: 10.1109/TMI.2017.2677499.

Kümmel, A., Selzer, P., Beibel, M., Gubler, H., et al. (2011). "Comparison of multivariate data analysis strategies for high-content screening". *Journal of Biomolecular Screening* 16.3, pp. 338–347. DOI: 10.1177/1087057110395390.

Lécuyer, E. and Tomancak, P. (2008). "Mapping the gene expression universe". *Current Opinion in Genetics and Development* 18.6, pp. 506–512. DOI: 10.1016/j.gde.2008.08.003.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., et al. (Oct. 2007). "Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function." *Cell* 131.1, pp. 174–87. DOI: 10.1016/j.cell.2007.08.003.

Lin, H., Chen, H., Graham, S., Dou, Q., et al. (Aug. 2019). "Fast ScanNet: Fast and Dense Analysis of Multi-Gigapixel Whole-Slide Images for Cancer Metastasis Detection". *IEEE Transactions on Medical Imaging* 38.8, pp. 1948–1958. DOI: 10.1109/TMI.2019.2891305.

Linkert, M., Rueden, C. T., Allan, C., Burel, J. M., et al. (2010). "Metadata matters: Access to image data in the real world". *Journal of Cell Biology* 189.5, pp. 777–782. DOI: 10.1083/jcb.201004104.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., et al. (2017). "Detecting Cancer Metastases on Gigapixel Pathology Images", pp. 1–13. arXiv: 1703.02442.

Ljosa, V., Caie, P. D., Ter Horst, R., Sokolnicki, K. L., et al. (2013). "Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment". *Journal of Biomolecular Screening* 18.10, pp. 1321–1329. DOI: 10.1177/1087057113503553.

Loo, L.-h., Wu, L. F., and Altschuler, S. J. (2007). "Image-based multivariate profiling of drug responses from single cells". *Nature Methods* 4.5, pp. 445–453. DOI: 10.1038/NMETH1032.

Lou, X. and Hamprecht, F. A. (2011). "Structured Learning for Cell Tracking". *Advances in Neural Information Processing Systems (NIPS)*, pp. 1–9.

Maaten, L. V. D. and Hinton, G. (2008). "Visualizing Data using t-SNE". *Journal of Machine Learning Research* 9, pp. 2579–2605.

Mall, M., Walter, T., Gorjánácz, M., Davidson, I. F., et al. (Sept. 2012). "Mitotic lamin disassembly is triggered by lipid-mediated signaling." *The Journal of cell biology* 198.6, pp. 981–90. DOI: 10.1083/jcb.201205103.

Meyer, F. and Serra, J. (Apr. 1989). "Contrasts and Activity Lattice". *Signal Processing*. Special Issue on Advances in Mathematical Morphology 16.4, pp. 303–317. DOI: 10.1016/0165-1684(89)90028-5.

Meyer, F. (1986). "Automatic screening of cytological specimens". *Comput. Vision Graph. Image Process.* 35.3, pp. 356–369. DOI: http://dx.doi.org/10.1016/0734-189X(86)90005-8.

Michel Grimaud (June 1, 1992). "New measure of contrast: the dynamics". Proc.SPIE. Vol. 1769.

Moffitt, J. R., Hao, J., Bambah-Mukku, D., Lu, T., et al. (Dec. 13, 2016). "High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing". *Proceedings of the National Academy of Sciences* 113.50, pp. 14456–14461. DOI: 10.1073/pnas.1617699113.

Mueller, F., Senecal, A., Tantale, K., Marie-Nelly, H., et al. (Apr. 2013). "FISH-quant: automatic counting of transcripts in 3D FISH images". *Nature Methods* 10.4, pp. 277–278. DOI: 10.1038/nmeth.2406.

Myers, G. (July 2012). "Why bioimage informatics matters." *Nature methods* 9.7, pp. 659–60. DOI: 10.1038/nmeth.2024.

Naffar-Abu-Amara, S., Shay, T., Galun, M., Cohen, N., et al. (Jan. 23, 2008). "Identification of Novel Pro-Migratory, Cancer-Associated Genes Using Quantitative, Microscopy-Based Screening". *PLOS ONE* 3.1, e1457. DOI: 10.1371/journal.pone.0001457.

Naylor, P., Boyd, J., Laé, M., Reyal, F., et al. (2019). "Predicting Residual Cancer Burden In A Triple Negative Breast Cancer Cohort". *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 933–937. DOI: 10.1109/ISBI.2019.8759205.

Naylor, P. (Dec. 18, 2019). "From Cellular Phenotypes to the Analysis of Whole Slide Images : Application to Treatment Response in Triple-Negative Breast Cancer". Theses. Paris, France: PSL Research University.

Naylor, P., La, M., Reyal, F., and Walter, T. (2018). "Segmentation of Nuclei in Histopathology Images by deep regression of the distance map". *IEEE Transactions on Medical Imaging* 0062 (c), pp. 1–12. DOI: 10.1109/TMI.2018.2865709.

Naylor, P., Lae, M., Reyal, F., and Walter, T. (2017). "Nuclei Segmentation in Histopathology Images Using Deep Neural Networks". *2017 Ieee 14th International Symposium on Biomedical Imaging (isbi 2017)*. ISBN: 978-1-5090-1172-8, IEEE, EMB, IEEE Signal Proc Soc. DOI: 10.1109/ISBI.2017.7950669.

Neumann, B., Walter, T., Hériché, J.-K., Bulkescher, J., et al. (Apr. 2010). "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes." *Nature* 464.7289, pp. 721–7. DOI: 10.1038/nature08869.

Orlov, N., Shamir, L., Macura, T., Johnston, J., et al. (Jan. 2008). "WND-CHARM: Multi-purpose image classification using compound image transforms". *Pattern recognition letters* 29.11, pp. 1684–1693.

Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F., et al. (2018). "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy". *Nature Methods* 15.11, pp. 917–920. DOI: 10.1038/s41592-018-0111-2.

Pau, G., Walter, T., Neumann, B., Hériché, J.-k., et al. (2013). "Dynamical modelling of phenotypes in a genome-wide RNAi live-cell imaging assay". *BMC bioinformatics* 14.308, pp. 1–10. DOI: 10.1186/1471-2105-14-308.

Paul-Gilloteaux, P., Heiligenstein, X., Belle, M., Domart, M. C., et al. (2017). "EC-CLEM: Flexible multidimensional registration software for correlative microscopies". *Nature Methods* 14.2, pp. 102–103. DOI: 10.1038/nmeth.4170.

Pawlowski, N., Caicedo, J. C., Singh, S., Carpenter, A. E., et al. (Nov. 2, 2016). "Automating Morphological Profiling with Generic Deep Convolutional Networks". *bioRxiv*, p. 085118. DOI: 10.1101/085118.

Peng, H. (Sept. 2008). "Bioimage informatics: a new area of engineering biology." *Bioinformatics (Oxford, England)* 24.17, pp. 1827–36. DOI: 10.1093/bioinformatics/btn346.

Peng, H., Bateman, A., Valencia, A., and Wren, J. D. (Apr. 2012). "Bioimage informatics: a new category in Bioinformatics." *Bioinformatics (Oxford, England)* 28.8, p. 1057. DOI: 10.1093/bioinformatics/bts111.

Peng, H., Chung, P., Long, F., Qu, L., et al. (2011). "BrainAligner : 3D registration atlases of Drosophila brains". *Nature Methods* 8.6, pp. 493–500. DOI: 10.1038/nmeth.1602.

Pepperkok, R. and Ellenberg, J. (2006). "High-throughput fluorescence microscopy for systems biology". *Nature reviews. Molecular cell biology* 7.9, pp. 690–696.

Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., et al. (Nov. 2004). "Multidimensional drug profiling by automated microscopy." *Science (New York, N.Y.)* 306.5699, pp. 1194–8. DOI: 10.1126/science.1100709.

Qaiser, T., Tsang, Y.-W., Taniyama, D., Sakamoto, N., et al. (July 1, 2019). "Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features". *Medical Image Analysis* 55, pp. 1–14. DOI: 10.1016/j.media.2019.03.014.

Raj, A., Bogaard, P. van den, Rifkin, S. A., Oudenaarden, A. van, et al. (Oct. 2008). "Imaging individual mRNA molecules using multiple singly labeled probes". *Nature Methods* 5.10, pp. 877–879. DOI: 10.1038/nmeth.1253.

Rajaram, S., Pavie, B., Wu, L. F., and Altschuler, S. J. (2012). "PhenoRipper : software for rapidly profiling microscopy images". *Nature Methods* 9.7, pp. 635–637.

Reeve, R. J. and Prokop, A. P. (1992). "A Survey of Moment-Based Techniques For Unoccluded Object Representation and Recognitio n". *CVGIP: Graphical Models and Image Processing*, pp. 438–460.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., et al. (2017). "The human cell atlas". *eLife* 6, pp. 1–30. DOI: 10.7554/eLife.27041. arXiv: 121202 [10.1101].

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". *Medical Image Computing and Computer-Assisted Intervention*

– *MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Springer International Publishing, pp. 234–241. arXiv: 1505.04597v1.

Rust, M. J., Bates, M., and Zhuang, X. (2006). "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)". *Nature Methods* 3.10, pp. 793–795. DOI: 10.1038/NMETH929.

Samacoits, A., Chouaib, R., Safieddine, A., Traboulsi, A.-M., et al. (2018). "A computational framework to study sub-cellular RNA localization". *Nature Communications* 9.1, p. 4584. DOI: 10.1038/s41467-018-06868-w.

Sarder, P. and Nehorai, A. (2006). "Deconvolution methods for 3-D fluorescence microscopy images". *IEEE Signal Processing Magazine* 23.3, pp. 32–45. DOI: 10.1109/MSP.2006.1628876.

Sbalzarini, I. F. and Koumoutsakos, P. (Aug. 2005). "Feature point tracking and trajectory analysis for video imaging in cell biology." *Journal of structural biology* 151.2, pp. 182–95. DOI: 10.1016/j.jsb.2005.06.002.

Schoenauer Sebag, A., Plancade, S., Raulet-Tomkiewicz, C., Barouki, R., et al. (2015a). "A generic methodological framework for studying single cell motility in high-throughput time-lapse data". *Bioinformatics* 31.12, pp. i320–i328. DOI: 10.1093/bioinformatics/btv225.

Schoenauer Sebag, A., Plancade, S., Raulet-Tomkiewicz, C., Barouki, R., et al. (2015b). "Infering an ontology of single cell motions from high-throughput microscopy data". *Proceedings of the 12th IEEE International Symposium on Biomedical Imaging (ISBI): From nano to macro*. New York, New York, USA, pp. 160–163.

Schubert, W., Bonnekoh, B., Pommer, A. J., Philipsen, L., et al. (Oct. 2006). "Analyzing proteome topology and function by automated multidimensional fluorescence microscopy." *Nature biotechnology* 24.10, pp. 1270–8. DOI: 10.1038/nbt1250.

Serra, J. (1983). *Image Analysis and Mathematical Morphology*. Orlando, FL, USA: Academic Press, Inc.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). "In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus". *Neuron* 92.2. Publisher: Elsevier Inc. ISBN: 0896-6273, pp. 342–357. DOI: 10.1016/j.neuron.2016.10.001.

Sharpe, J., Ahlgren, U., Perry, P., Hill, B., et al. (2002). "Optical Projection Tomography as a Tool for 3D Microscopy and Gene Expression Studies". *Science* 296, pp. 541–545. DOI: 10.1126/science.1068206.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., et al. (2017). "Learning from Simulated and Unsupervised Images through Adversarial Training". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, pp. 2242–2251. DOI: 10.1109/CVPR.2017.241.

Sihn, G., Walter, T., Klein, J.-C., Queguiner, I., et al. (Mar. 2007). "Anti-angiogenic properties of myo-inositol trispyrophosphate in ovo and growth reduction of implanted glioma." *FEBS letters* 581.5, pp. 962–6. DOI: 10.1016/j.febslet.2007.01.079.

Simpson, J. C., Joggerst, B., Laketa, V., Verissimo, F., et al. (July 2012). "Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway." *Nature cell biology* 14.7, pp. 764–74. DOI: 10.1038/ncb2510.

Simpson, K. J., Selfors, L. M., Bui, J., Reynolds, A., et al. (Sept. 2008). "Identification of genes that regulate epithelial cell migration using an siRNA screening approach." *Nature cell biology* 10.9, pp. 1027–38. DOI: 10.1038/ncb1762.

Singh, D. K., Ku, C.-J., Wichaidit, C., Steininger, R. J., et al. (May 2010). "Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities." *Molecular systems biology* 6.369, p. 369. DOI: 10.1038/msb.2010.22.

Singh, S., Carpenter, A. E., and Genovesio, A. (2014). "Increasing the content of high-content screening: An overview". *Journal of Biomolecular Screening* 19.5, pp. 640–650. DOI: 10.1177/1087057114528537.

Slack, M. D., Martinez, E. D., Wu, L. F., and Altschuler, S. J. (Dec. 2008). "Characterizing heterogeneous cellular responses to perturbations." *Proceedings of the National Academy*

of Sciences of the United States of America 105.49, pp. 19306–11. DOI: 10.1073/pnas.0807038105.

Snijder, B., Sacher, R., Rämö, P., Damm, E. M., et al. (2009). "Population context determines cell-to-cell variability in endocytosis and virus infection". *Nature* 461.7263, pp. 520–523. DOI: 10.1038/nature08282.

Soille, P. (2003). *Morphological Image Analysis: Principles and Applications*. 2nd ed. Springer-Verlag New York, Inc.

Sommer, C., Hoefler, R., Samwer, M., and Gerlich, D. W. (Sept. 27, 2017). "A deep learning and novelty detection framework for rapid phenotyping in high-content screening". *Molecular Biology of the Cell* 28.23, pp. 3428–3436. DOI: 10.1091/mbc.e17-05-0333.

Stringer, C., Michaelos, M., and Pachitariu, M. (Jan. 1, 2020). "Cellpose: A Generalist Algorithm for Cellular Segmentation". *bioRxiv*, p. 2020.02.02.931238. DOI: 10.1101/2020.02.02.931238.

Suratanee, A., Schaefer, M. H., Betts, M. J., Soons, Z., et al. (Sept. 2014). "Characterizing protein interactions employing a genome-wide siRNA cellular phenotyping screen." *PLoS computational biology* 10.9, e1003814. DOI: 10.1371/journal.pcbi.1003814.

Swedlow, J., Goldberg, I., Brauner, E., and Sorger, P. (2003). "Informatics and quantitative analysis in biological imaging". *Science* 300.5616, pp. 100–102. DOI: 10.1126/science.1082602.Informatics.

Tegha-Dunghu, J., Bausch, E., Neumann, B., Wuensche, A., et al. (Dec. 1, 2014). "MAP1S controls microtubule stability throughout the cell cycle in human cells". *Journal of Cell Science* 127.23, 5007 LP –5013. DOI: 10.1242/jcs.136457.

Tegha-Dunghu, J., Neumann, B., Reber, S., Krause, R., et al. (May 15, 2008). "EML3 is a nuclear microtubule-binding protein required for the correct alignment of chromosomes in metaphase". *Journal of Cell Science* 121.10, 1718 LP –1726. DOI: 10.1242/jcs.019174.

Terjung, S., Walter, T., Seitz, A., Neumann, B., et al. (2010). "High-throughput microscopy using live mammalian cells." *Live Cell Imaging: A Laboratory Manual*. Ed. by R. Goldman, J. Swedlow, and D. Spector. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., et al. (2017). "A subcellular map of the human proteome". *Science* 356.6340, eaal3321. DOI: 10.1126/science.aal3321.

Tsanov, N., Samacoits, A., Chouaib, R., Traboulsi, A. M., et al. (2016). "SmiFISH and FISH-quant - A flexible single RNA detection approach with super-resolution capability". *Nucleic Acids Research* 44.22. DOI: 10.1093/nar/gkw784.

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., et al. (2015). "Tissue-based map of the human proteome". *Science* 347.6220, pp. 1260419–1260419. DOI: 10.1126/science.1260419. arXiv: 0208024 [gr-qc].

Uhlmann, V., Singh, S., and Carpenter, A. E. (2016). "CP-CHARM: Segmentation-free image classification made accessible". *BMC Bioinformatics* 17.1. DOI: 10.1186/s12859-016-0895-y.

Van Valen, D. A., Kudo, T., Lane, K. M., Macklin, D. N., et al. (2016). "Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments". *PLoS Computational Biology* 12.11. ISBN: 17444292 (Electronic), pp. 1–24. DOI: 10.1371/journal.pcbi.1005177.

Veta, M., Diest, P. J. V., Willems, S. M., Wang, H., et al. (2014). "Assessment of algorithms for mitosis detection in breast cancer histopathology images". *Medical Image Analysis*, pp. 1–23.

Vladimir, U. (2017). "An objective comparison of cell-tracking algorithms". *Nature Methods* 14.12, pp. 1141–1152. DOI: 10.1038/nmeth.4473.

Walker, R. F. and Jackway, P. T. (1996). "Statistical Geometric Features - Extensions for Cytological Texture Analysis". *ICPR - International Conference on Pattern Regognition*.

Walter, T., Klein, J.-C., Massin, P., and Erginay, A. (Oct. 2002a). "A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus

images of the human retina". *IEEE Transactions on Medical Imaging* 21.10, pp. 1236–1243. DOI: 10.1109/TMI.2002.806290.

Walter, T., Held, M., Neumann, B., Hériché, J. K., et al. (2008). "A genome wide RNAi screen by time lapse microscopy in order to identify mitotic genes - Computational aspects and challenges". *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI*, pp. 328–331. DOI: 10.1109/ISBI.2008.4540999.

Walter, T., Held, M., Neumann, B., Hériché, J.-K., et al. (Apr. 2010a). "Automatic identification and clustering of chromosome phenotypes in a genome wide RNAi screen by time-lapse imaging." *Journal of structural biology* 170.1, pp. 1–9. DOI: 10.1016/j.jsb.2009.10.004.

Walter, T. and Klein, J.-C. (Oct. 2001). *Segmentation of Color Fundus Images of the Human Retina: Detection of the Optic Disc and the Vascular Tree Using Morphological Techniques*, pp. 282–287. DOI: 10.1007/3-540-45497-7_43.

Walter, T. and Klein, J.-C. (2002b). "Automatic Detection of Microaneurysms in Color Fundus Images of the Human Retina by Means of the Bounding Box Closing". *Medical Data Analysis*. Ed. by A. Colosimo, P. Sirabella, and A. Giuliani. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 210–220.

Walter, T. and Klein, J.-C. (2005). "Automatic Analysis of Color Fundus Photographs and Its Application to the Diagnosis of Diabetic Retinopathy BT - Handbook of Biomedical Image Analysis: Volume II: Segmentation Models Part B". Ed. by J. S. Suri, D. L. Wilson, and S. Laxminarayan. Boston, MA: Springer US, pp. 315–368. DOI: 10.1007/0-306-48606-7_7.

Walter, T., Massin, P., Erginay, A., Ordonez, R., et al. (2007). "Automatic detection of microaneurysms in color fundus images". *Medical Image Analysis* 11.6, pp. 555–66. DOI: 10.1016/j.media.2007.05.001.

Walter, T., Shattuck, D. W., Baldock, R., Bastin, M. E., et al. (2010b). "Visualization of image data from cells to organisms". *Nature Methods* 7.3, S26–S41. DOI: 10.1038/NmEtH.1431.

Weigert, M., Schmidt, U., Boothe, T., Müller, A., et al. (2018). "Content-aware image restoration: pushing the limits of fluorescence microscopy". *Nature Methods* 15.12, pp. 1090–1097. DOI: 10.1038/s41592-018-0216-7.

Whitehill, J. and Ramakrishnan, A. (Dec. 19, 2018). "Automatic Classifiers as Scientific Instruments: One Step Further Away from Ground-Truth". *arXiv:1812.08255 [cs, stat]*. arXiv: 1812.08255.

Williams, E., Moore, J., Li, S. W., Rustici, G., et al. (2017). "Image Data Resource: A bioimage data integration and publication platform". *Nature Methods* 14.8, pp. 775–781. DOI: 10.1038/nmeth.4326.

Xing, F., Member, S., and Yang, L. (2016). "Robust Nucleus / Cell Detection and Segmentation in Digital Pathology and Microscopy Images : A Comprehensive Review". *IEEE Reviews in Biomedical Engineering* 9. ISBN: 0324141122, pp. 234–263. DOI: 10.1109/RBME.2016.2515127. arXiv: 15334406.

Yin, Z., Sadok, A., Sailem, H., McCarthy, A., et al. (2013). "A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes." *Nature cell biology* 15.7, pp. 860–71. DOI: 10.1038/ncb2764.

Yin, Z., Zhou, X., Bakal, C., Li, F., et al. (Dec. 2008). "Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens". *BMC Bioinformatics* 9.1, pp. 1–20. DOI: 10.1186/1471-2105-9-264.

Young, I. T. (July 1972). "The Classification of White Blood Cells". *IEEE Transactions on Biomedical Engineering* BME-19.4, pp. 291–298. DOI: 10.1109/TBME.1972.324072.

Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., et al. (Oct. 2012). "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling." *Science translational medicine* 4.157, 157ra143. DOI: 10.1126/scitranslmed.3004330.

Zhao, T. and Murphy, R. F. (2007). "Automated Learning of Generative Models for Subcellular Location : Building Blocks for Systems Biology". *Cytometry* 71A, pp. 978–990. DOI: 10.1002/cyto.a.20487.

Zhu, Q., Shah, S., Dries, R., Cai, L., et al. (Dec. 2018). "Identification of Spatially Associated Subpopulations by Combining scRNAseq and Sequential Fluorescence in Situ Hybridization Data". *Nature Biotechnology* 36.12, pp. 1183–1190. DOI: 10.1038/nbt.4260.