Nanoparticle Induced Cell Magneto-Rotation for the Multiplexed Monitoring of Morphology, Stress and Drug Sensitivity of Suspended Single Cancer Cells.

By

Remy A. Elbez

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied Physics)
in the University of Michigan
2015

Doctoral Committee:

       Professor Raoul Kopelman, Chair
       Professor Roy Clarke
       Professor Charles R. Doering
       Professor Alan J. Hunt (deceased)
       Professor Shuichi Takayama
       Professor Robert M. Ziff

Acknowledgements

First and foremost, I would like to express my profound gratitude to my advisor, Prof. Raoul Kopelman. Prof. Kopelman offered me the invaluable chance to first intern in his lab at the end of my master's degree at University of Michigan. Even though it was supposed to be a short stay, he always accepted to keep me working in the lab when I asked, and without this opportunity I would have never discovered my passion for research, a passion that led me to try to get my PhD. Naturally, Prof. Kopelman did not hesitate and gave me all the support he could to help me get into the Applied Physics program, even though the deadlines were a distant past. This incredible support has never faded since then, and I cannot stress enough the great latitude and freedom I have come to enjoy in my research. I feel extremely privileged to have had access to his immense knowledge and mentorship. Over the (many) years of my PhD, I have come across many challenges and moments of self-doubts, and without Prof. Kopelman's support, I would not have been able to find my way through. For all of this and many more things, thank you.

I would like to thank my committee members, Prof. Roy Clarke, Prof. Charles Doering, Prof. Shuichi Takayama and Prof. Robert Ziff for their help and wonderful guidance. I would like to thank Prof. Clarke for his availability and advice while I was in the first years of my PhD. Even though I am far from being supportive of Manchester City, his favorite football club, he has warmly welcomed me and greatly helped me find my way and my place in the Applied Physics family. I also send all my appreciation to Prof. Shuichi Takayama, whose expertise in the microfluidic world has changed the course of my PhD. Especially, he has shown a lot of patience while my project was developing, and this gave me a sense of confidence that pushed me to do my best to return the favor. I would also want to have a thought for Prof. Alan Hunt, who served in my committee and abruptly passed away in 2012. I will always remember his enthusiasm and

the *Classes Preparatoires* system. No matter how hard things appear to be at first, I gained confidence that with patience and conscientious work, I could excel in the scientific subjects. I would also have special thanks to Patrick Teller, Michel Wirth, Gilles Alozy and Marie-Jo Galabru, my Math, Physics and Philosophy teachers in *classes prepas*. I would never thank you enough for all the solid foundations you gave me for the rest of my education and life.

To my closest friends, Alexandre Dubreucq, Federico Gottardo, Johan Atlani, Erik Loualiche, Valentin Haddad, Ruth Beer, German Martinez, Jeff Weil, Johan Drylewicz, Jeremy Kazzaz, Nicolas Lamorte, Judith Beer-Gabel, Fred Kadouch, Nick DeSanctis and Francesco Sguera, thank you for always being here. I promise to call more often.

Finally, to the most special of all, my family. I love you all and thank so much you for everything, this thesis is for you.

Table of Contents

List of figures

List of Tables

List of Appendices

Abstract

The metastatic process of a cancer relies on the transformation of some of the primary tumor cells into cells capable of migrating through the Extra-Cellular Matrix (ECM), surrounding the tumor, into the bloodstream and the lymph nodes, and then settle in distant tissue, growing new secondary tumors. By identifying, characterizing and quantifying these cells, the progression of cancer in a patient during therapy can be more accurately assessed. Here we describe the development of a new method for quantitative real time monitoring of cell size and morphology, on single live suspended cancer cells, unconfined in three dimensions. The enabling *cell magnetorotation* (CM) method is made possible by *nanoparticle induced cell magnetization*. Using a rotating magnetic field, the magnetically labeled cells are actively rotated, then imaged, using a high definition CCD camera. Under proper conditions, the rotation period of a magnetic object is proportional to its shape factor. We demonstrate first that the rotational period, when measured in real-time, can serve to track cellular response to drugs, cytotoxic agents and other chemical stimuli. In addition, while cells are rotated, they exhibit very specific morphological activities, even without a chemical stimulus. Described also is how to multiplex the CM method, to image several dozens to several thousands of cells simultaneously, and using morphology to classify cells into different phenotypic categories, with each phenotype being correlated with malignancy level. The intrinsic tumor heterogeneity, at the cellular level, can be visualized with relationship graphs. Shown is the ability to monitor cell morphological changes over long periods

of time, in real time, in order to detect the metastatic potential for heterogeneous populations of cancer cells, using tools from statistical analysis methods. The method relies on unsupervised Machine Learning algorithms which do not require human inputs. Overall it is demonstrated that the CM method can be used as a diagnostic tool to evaluate the phenotypical heterogeneity in a cell population in general, and in a cancer cell population in particular. This fast and high throughput method promises to efficiently assess the efficacy of personalized therapeutic strategies.

Chapter 1 :
Introduction

The Challenge of Studying Metastasis

The overwhelming majority of cancer related mortalities is the consequence of cancer metastasis, a process characterized by the colonization of distant tissues by malignant cells from the primary tumor. Even though the body can resist an isolated tumor, it becomes gradually more difficult, especially if new tumors have colonized distant organs, forcing them into failure. Cancer metastasis is one of the hallmarks of malignant tumors. Indeed most forms of cancer can metastasize, although to varying degrees [1]. During metastasis, some aggressive cancer cells acquire the ability to move through the extracellular matrix (ECM) of the tissue of origin, penetrate the walls of lymphatic and/or blood vessels, after which they are able to circulate through the bloodstream, and reach various sites and tissues in the body and there proliferate, creating new masses. Traditional therapies relying on the reduction of the cancer mass or volume, by removing or killing cancer cells, have often seen their efficacy to reduce and control cancer reach a plateau in time, followed by a reversal. Even though the response to treatment might look favorable in the early stages, it appears that traditional therapies still lack the ability to effectively control metastasis, leading to many recurring cancers.

The first step of metastasis is the collective or individual migration of cells from the primary tumor, usually after the Epithelial to Mesenchymal Transition (EMT). Either by mutation or by recruitment from other cells, cancer cells aquire characteristics that favor their migration. In particular, cells going through EMT lose their cell-cell adhesion enabling molecule (E-cadherin), easing their detachment from the rest of the tumor. The switch to a mesenchymal phenotype in cells is characterized by a sharp increase in migratory abilities, due to a more flexible cytoskeleton and nucleus, since oftentimes the cell will squeeze through holes smaller than the size of the nucleus, and to do so the nucleus loses part of its rigidity to be deformed and help the cell go through. Cells also aquire a higher resistance to apoptosis, and an increase in expression levels of enzymes capable of digesting the collagen fibers in the ECM, such as members of the Matrix Metalloproteinases (MMP) family, typically over-expressed after EMT. Simply put, the EMT provides cancer cells with all the necessary tools to go past and survive through numerous obstacles before settling into a new tissue. As mentionned earlier, the migrating cell leaving the tumor has to go through the ECM (either by squeezing in between the collagen fibers or by reducing them with enzymes), then enter the blood stream via intravasation, travel, attach itself to the endothelium and squeeze into a new tissue by extravasation. The extreme harshness of travel in the bloodstream shall not be underestimated, owing to very high mechanical constrain, as well as the difficulty for cells to anchor in the endothelium. The cancer cells that travel via the blood are also known as Circulating Tumor Cells (CTCs), and even though EMT strongly promotes migration, a large part of the CTCs are epithelial ones, and even when epithelial cells do leave the tumor, they are much less succesful at establishing new tumor sites than the mesenchymal cells, for the very reasons described above.

When in a new tissue, the migrating cell reverts to an epithelial phenotype, via the Mesenchymal to Epithelial transition (MET)[2], and grows a secondary tumor. However, even though EMT is pivotal to the metastatic process, it has to be noted that it is not necessarily involved in it. For instance, it has been reported that during collective migration, where a cluster of cells clustered together travels to a distant tissue, not all the cells are in a mesenchymal phenotype [3].

Tackling metastasis can thus rely on two stratagems: either by suppressing the cells that already left the primary tumor and are yet to extravasate into new tissues, or by suppressing the ones inside the primary tumor that have the potential to migrate, settle and populate new tissues.

Contrary to previous beliefs, tumors are extremly heterogeneous in their composition, be it genotypic or phenotypic, and this composition is made dynamic by alterations and modifications over time, as exemplified by the EMT process. It is known that tumor cells are heterogeneous in their composition and thus it has been hypothesized, and there is evidence to suggest it, that subpopulations of cells in the original primary tumor have, from inception or acquire *de novo*, the ability to initiate and/or complete the abbreviated metastatic cascade delineated above. Among these *aggressive cells* there may be varying degrees of plasticity, i.e. for example, varying intra- and extra-vasation ability, which are prerequisites for seeding a secondary tumor. After many years of intense controversy in the scientific community, the concept of Cancer Stem Cells (CSCs) is now widely accepted [4–9]. At the origin of the heterogeneity in a cancer cell population, we can find a minority of cells that have a higher resistance to drugs, a capacity to remain dormant and the ability to repopulate a tumor. In particular, CSCs explain the failure of therapies only oriented towards a reduction of the cancerous mass. If they are not specifically targetted, CSCs grow immune to the applied therapy, and, coupled with their repopulating capacities, regrow into a resistant tumor. Just like leaving intact the queen bees while decimating the large majority of worker bees does not affect much the survival of the whole colony, since the latter cannot reproduce anyway. This also explains why, in the case of many patients thought to be cured, the cancer can come back in a more aggressive form than the initial one, and be resistant to the therapy that used to work once. For this reasson, the development of technologies to understand the potential *in vivo* metastatic behavior of different cell populations is needed to develop therapeutic strategies that address those physical properties enabling successful metastatic steps.

As we can see, deciphering the metastatic process requires special technologies capable of extracting information from numerous cells, at the single cell level. Otherwise, the averaging process accompanying the measurements of given properties of a cancer cell population ends up

3

masking the presence of the minority of cells having the biggest impact on the evolution of cancer. We note that single cell analysis has led to critical discoveries in drug testing, immunobiology and stem cell research [10–13]. Also, a change from 2- to 3-dimensional growth conditions radically affects cell behavior, their phenotype and their protein expression [14–16]. Monitoring cells should thus give more accurate information if the cells in question are observed in an environment as close as possible to the one they are supposed to evolve in when in the human body. This approach already resulted in new observations on gene expression and communication networks and in better predictions of cell responses to their environment [17,18]. However, while morphological changes have been shown to be highly significant [19], it is still difficult to study the size and shape of single cells that are freely suspended. It is especially important that just a small minority of cells, whose behavior could be considered to be statistically insignificant, compared to the large majority of the population, such as stem-like cells, can have a critical biological and medical impact [20–22].

The process of metastasis is highly inefficient however. For instance, a tumor sheds around a million cancer cells every day in the bloodstream but the amount of metastatic sites obviously does not grow accordingly. In animal models, it has been shown that fewer than one out of 4000 circulating tumor cells successfully reached a new tissue [23]. It is very likely that most of the cancer cells that end up in the bloodstream are already dead or in the process of cell death. While the early stages of metastasis have been widely studied (migration from tumor after loss of cell-cell junctions), migration through the extra-cellular matrix, intravasation into the vasculature), very little is known about extravasation. It is generally agreed that for extravasation to occur, tumor cells must go through dynamic morphology changes, for example by forming *protrusions* or through an *amoeboid-like* stage [24,25]. Some cells have the ability to switch modes of invasion via what is termed a Mesenchymal to Amoeboid Transition (MAT) [26], and extravasate into new tissues without the need of specific surface anchoring [27,28]. This process does not appear to depend on expression of surface markers. either in the tumor cells or in the endothelium of the destination tissues. Moreover, cell motility in 3 dimensions is considerably different from motion in 2 dimensions, where cells rely primarily on contact guidance [27]. In 3

4

dimensions, shape influences motion, and shape is affected by multiple environmental factors. For instance, morphological changes, such as *blebs*, have been associated with anti-tumor effects and also with drug resistance [29,30]. Focusing on the cell cytoskeleton and limiting its ability to change shape has recently led to powerful drug discoveries [31]. However, in regards to medical diagnostics, there is no existing automated method to study the shapes of cells in suspension, *in vitro*, so as to achieve a high throughput characterization of extravasation; in fact, existing methods that rely on cell image analyses and processing are exceedingly cumbersome and difficult to generalize across biological platforms.

Specific methods used to track morphological changes of single biological cells include Atomic Force Microscopy (AFM) [32], Optical Tweezers [33] (OT), micropipette aspiration [34], microplate manipulation [35], optical stretching [36,37] and microfluidic deformation through passage in a microchannel [38–40].

AFM uses the sharp tip of a cantilever acting like a spring, to scan the surface of a specimen. When a force is applied to the tip (by reaction to the surface), it bends the cantilever. The tip is very sharp, with only a few atoms sitting at the very end and a radius of curvature in the nanometers. Using the deflection of a laser beam aimed at the cantilever, right above the tip, it is possible to measure the displacement of the tip. The AFM is extremely precise, with sub-nanometers precisions being reported. The very high resolution of the AFM and its specific design to measure force makes it an ideal choice measurement in cell biology, and, notably, of single cell elasticity. AFM has been used over a very broad range of measurements: detection of biomolecules, mechanical response of cells, nano-injection into single cells, and cell selection, among others[41–44]. However, using AFM to monitor live cells has several drawbacks. Imaging a sample necessitates the displacement of the tip, slowing down the process. For instance, even with a fast AFM, it takes several minutes to image an area of 20 by 20 micrometers, roughly the order of magnitude of the surface covered by a cell when plated. Unfortunately, this pace is too slow to be able to capture cellular events that occur in a few seconds, or even in a few minutes. For instance, the changes in cytosckeletal organisation, at the membrane of the cell, such as filopodia or blebbs, only take a few seconds to grow several micrometers and retract or change direction. By the time the AFM has finished covering the area of the cell, the cell might be in a

completely different state. Also, the contact with the AFM probe may perturb the cell. In addition, measuring specimens with an AFM probe requires the specimens to be attached to a surface, and this renders impossible its applicability to cells in suspension.

Optical tweezers (OT) are more adapted to single cell micromanipulation. The basic principle of OTs is the radiation pressure exerted by a focused laser beam on an object (i.e. a particle, a cell etc…). When a pair of laser beams, emitting toward each other, are used, objects can actually be trapped and manipulated by adjusting the intensities of each of the beams. Focusing the beam also produces lateral gradients of forces that act as a spring and help retain the object in the axis of the beams, thus effectively forming an optical trap. OT have been used to manipulate cells for mechanical measurements, cell sorting and isolation, quantification of various biological forces, or for the identification of tumor cells[33,45–47]. However, the use of OTs is limited by the power of the laser beam, which can photodamage the cells, or its force may perturb the cell, and by the difficulty to control the position of the trap regarding the position of the targetted cells. In addition, experiments using an OT require a complex optical and control system, and the scalability of OTs is not ideal when it comes to multiplexing.

Flow cytometry based FACS (Fluorescence Activated Cell Sorting) is also widely used for capturing the heterogeneity in cell populations and for sorting single cells. By flowing cells in a very thin tube at a high flow rate, it is possible to form a stream of single cells. With the help of one or several lasers, fluorescent dyes in the cells can be detected, along with the intensity in each fluorescent channel, at the single cell level. Therefore, FACS is a trully high-throughput single cell analysis and sorting technique [48]. Though multiple cellular events can be tracked in a population (owing to the use of multiple fluorescence channels), the main disadvantage of FACS is the impossibility to monitor the same cells, repeatably, at different time points and by not being able to resolve the localization of the fluorescent labels[48].

Finally, though being the oldest technique presented here, tissue sectioning is still highly regarded by professionals who utilize histological samples on a regular basis. One of the strenghs of tissue sections is the reproduction of a slice of tissue as it is found in the body. Adding stains to identify different cellular or sub-cellular components allows fast and accurate diagnosis. However, the staining process requires fixation of the cells, and it is thus impossible to isolate

them for further analysis or culture. Furtheremore, traditional histology looks at a static array of cells, in 2 dimensions, with no ability to monitor cell dynamics in 3 dimensions. We thus call the CMR-based method presented here *dynamic histology*.

Some of these methods may offer high resolution (AFM and OT), but are limited to a low throughput, and also by the attachment of cells to a surface (AFM), or (for OT) by the irreversible photodamage caused by laser trapping (OT), amongst other technical drawbacks that limit their general applicability. Some of these methods measure cell deformability at a high resolution, but in essence overlook more dynamically relevant cell morphology changes such as *amoeboid* behavior. Therefore, so far, only observation by the human eye (via microscope), or real-time movies that make a longitudinal record of millions of observations of a single cell, have been used to recognize morphological changes in cells that lead to quantitative data.

Since cell morphology mirrors the migratory behavior of the cell [19], studying the morphology of cells dissociated from primary tumors can provide us with a valuable link between intracellular molecular regulation and cellular-scale events, namely motility, invasiveness and especially extravasation. This is the rationale behind our development of the Cell-MagnetoRotaion based Dynamic Histology method described below, where magnetic nanoparticles are used for the *cell magnetic staining*.

Magnetism

Magnetic properties of the Super-paramagnetic Nano-particles used for Cell Magnetic Staining

Magnetic nanoparticles (MNPs) are a widely used type of nanoparticles, with applications as diverse as magnetic fluids[49], magnetic separation of target cells with technologies such as CellSearch®, magnetic resonance imaging[50,51], molecular imaging[52], biosensors[53] and data storage[54]. The ability to heat up magnetic materials using a rapidly changing external field has also been applied, so as to create new forms of cancer treatments, using hypothermia of the targeted cancer cells[55].

Bulk materials can be divided into three different categories, based on their magnetic properties: ferromagnetic, paramagnetic and diamagnetic. In general, what we call magnets are made of ferromagnetic materials. They are permanent magnets because they do have a permanent magnetic moment that creates a strong magnetic field around the material.

Diamagnetic materials are characterized by atoms that only have paired electrons orbiting around them. The angular moments and the magnetic moments thus cancel out, and there is no magnetization. However, in the presence of an external magnetic field, a magnetic moment is induced, and it tends to oppose the causes that induced it, and the material is repelled if it approaches a magnet.

Paramagnetic materials and ferromagnetic materials both have a net magnetic moment. However, in the case of paramagnetic materials, it is extremely small. In the absence of an external field, each atom composing the bulk of the material has a magnetic moment, the orientation of which is independent from the other ones and fluctuates under thermal motion, inducing random orientations of the magnetic moments. Summed together, all these magnetic moments add up to zero. However, in the presence of an external field, all these magnetic moments align with the field, so as to minimize the potential magnetic energy $U_{mag} = -\vec{\mu}.\vec{B}$. The magnetic moment induced by turning on an external field is proportional to the intensity of the field applied.

While paramagnetic materials have magnetic moments whose mutual interaction is of the order of thermal excitation (and thus negligible), this is not the case for ferromagnetic materials, in which the magnetic moments interact and tend to align with each other. Below a certain temperature, called the Curie temperature, ferromagnetic materials exhibit a strong net magnetic moment. The interaction that pushes the moments to align with each other is a relatively short distance one. Between two moments, or spins, if they are too far away, the magnetic interaction prevails, and the spins will try to minimize the potential energy of the interaction. This explains why ferromagnetic materials are made of separate domains, and within each of them, all the magnetic moments are aligned. Depending on the material, the size of a domain is below 100-150nm. Generally, because of the orientation of the domains, a ferromagnetic material is unmagnetized.

Now, if we turn on an external magnetic field and gradually increase its intensity from zero, all the subdomains align with the magnetic field, and for a strong enough field, we might be able to lock the moments in a specific orientation that remains even after the field is turned off, creating a permanent magnet.

As we can see, if we break down ferromagnetic materials into smaller pieces, smaller than the dimension of a magnetic domain, we get a crystal structure in which all the magnetic moments are aligned, and therefore we can talk of the magnetic moment of the (nano)crystal instead. However, because there is no external constrain on the orientation of the moment (since there is no other domain with which to interact), the direction of the magnetic moment of the nanocrystal is random (it responds to thermal fluctuations). However, in the presence of a magnetic field, the magnetic moment of the nanocrystal will align with the external field. On the other hand, they do not retain any magnetization in the absence of a magnetic field. This is the exact same property that is used in the superparamagnetic nanoparticles that we utilize in our experiments. Made of nanocrystals of Maghemite γ-Fe_2O_3, coated with a 2nm monolayer of oleic acid and a 2nm monolayer of amphiphilic acid. Their magnetic moments will align in the presence of an external field, but will not retain any magnetization.

After cellular uptake by endocytosis, the cell actually acts like a super-paramagnetic micro-particle, and we can use it as an actuator and a biosensor.

Theoretical analysis of the equation of motion of the cell:

The magnetic torque of the magnetized cells is given by the following expression:

$$\vec{\tau}_{mag} = \vec{m} \times \vec{B} = (\vec{m}_{perm} + \vec{m}_{ind}) \times \vec{B} \tag{1}$$

Here τ_{mag} is the total magnetic torque of the cell, $\vec{B}$ the external magnetic field and $\vec{m}$ the magnetic moment of the cell. The latter is the sum of the permanent magnetic moment, $\vec{m_{perm}}$ and the induced magnetic moment, $\vec{m_{ind}}$.

With Ω being the frequency of the applied magnetic field and $\chi(\Omega)$ the magnetic susceptibility of the cell, one gets:

$$\vec{\tau}_{mag} = \vec{m}_{perm} \times \vec{B} + Re\left[\chi(\Omega)V_m \frac{\vec{B}}{\mu_0}\right] = [m_{perm}Bsin(\Omega t-\theta) + \chi''(\Omega)\frac{B^2}{\mu_0}V_m]\vec{e_z} \tag{2}$$

Here V_m is the volume of the magnetic content of the cell [56], $\overrightarrow{m_{perm}}$ he norm of the permanent magnetic moment, B the intensity of the applied magnetic field, μ_0 the permeability of the free space, and $\chi''(\Omega)$ the imaginary part of the magnetic susceptibility of the cell. The non-zero imaginary part of the magnetic susceptibility of the superparamagnetic nanoparticles is the element responsible for the presence of a non-zero induced torque along the z-axis.

As can be seen, the magnetic torque is an additive sum of two contributions: a permanent one, and an induced magnetic moment.

However, in cell culture media (DMEM), at room temperature, the critical frequency of the cell is not observed (figure S1), and, as a consequence, for Ω = 15Hz, the permanent magnetic torque is negligible [57]. We are thus left with:

$$\tau_{mag} = \chi''(\Omega)V_m \frac{B^2}{\mu_0} \vec{e}_z \tag{3}$$

Finally, applying Newton's second law of motion to the rotating cell, neglecting the cell's moment of inertia and the Brownian forces, one finds the following equality for the driving magnetic torque and the opposing torque derived from the viscous forces:

$$\vec{\tau}_{mag} = -\vec{\tau}_{drag} = \kappa\eta V \dot{\theta} \vec{e}_z \tag{4}$$

With κ being the shape factor of the cell (=6 for a sherical shape), η the viscosity of the medium, and V the total volume of the cell. The angular speed is then found to be given by:

$$\dot{\theta} = \frac{\chi''(\Omega)V_m B^2}{\kappa\eta V \mu_0} \tag{5}$$

Since the magnetic content (due to the embedded magnetic nanoparticles) of the cell does not significantly change over the course of the measurement, we can assume that V_m and $\chi''(\Omega)$ are constant in time. As a consequence, the rotational speed is inversely proportional to the product of the shape factor by the volume, namely the *effective volume of the cell*, V_{eff}:

$$\dot{\theta} \propto \frac{1}{\kappa V} = \frac{1}{V_{eff}}$$

$$(6)$$

We thus deduce for the rotation period, T:

$$T = \frac{2\pi}{\dot{\theta}} \propto V_{eff}.$$

$$(7)$$

In addition to the *sine qua none* use of the super-paramagnetic properties of the MNPs, it has to be noted that the magnetic torque is actually transmitted to the cell because the MNPs are localized in the endosomes, themselves attached to the cytoskeleton of the cell. Without these serial attachments, free particles or endosomes would not be able to move the cell's rotation.

Mechanical forces on the cell

While the cells are rotated, they experience levels of shear stress due to the rotation in the well, and also because of the presence of the trap walls that also contribute to it. In this situation, the level of shear stress that is found is similar to the levels encountered by a cell entering (or getting out) of the blood vessel, between 0.5 and 30dyne/cm2[58]**.** If we model the cell by a sphere of radius r = 10μm, assuming it is centered in the well, the maximum distance to the trap wall is d_{max} = 13μm and minimum distance d_{min} = 1.6μm (see **figure 1.1**), we get the following formula for the approximate shear stress, σ (with $\dot{\theta}$ the rotation speed, and η the viscosity of the fluid):

$$\sigma = 2\pi\eta\dot{\theta}\frac{r}{d}$$

With rotation speeds typically between 0.4s-1 and 0.03s-1 (rotation periods between 2.5 and 30 seconds), we get values of σ between 0.001 and 0.1 dyne/cm2. We will study how the rotation speed, well-size, and shear stress affect cell morphology feature classification. Such results will be used to adjust well size and rotation speed, so as to obtain the best differentiation between more and less malignant cancer cells.
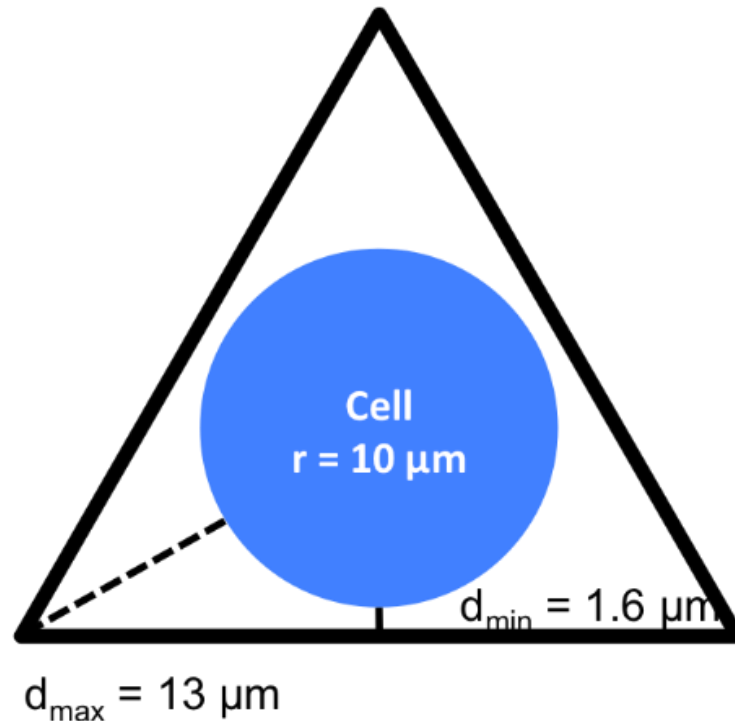
Figure 1.1: Geometry of a single cell trapped in a triangular microwell, with a schematic of a cell.

Outline of the Disseration

This thesis explores the development of the use of magnetic nanoparticles inside eukaryotic cells to induce rotation under the application of a low intensity rotating external magnetic field, which we called the Cell Magneto-rotation (CM) method. The development process took three stages. First was the proof of concept, then the multiplexing of the method to study the morphological changes in a cell population, and finally another layer of complexity in the multiplexing was added with the utilization of machine learning algorithms in order to separate two cancer cell populations based on their differential malignancy.

Chapter 2 presents the principle of the Magneto-rotation method. Cancer cells (HeLa) are made to passively (without specific targeting) uptake magnetic nanoparticles (MNPs). The successful uptake through endosomic pathways and the superparamagnetic properties of the nanoparticles where pivotal in the ability to rotate the cells, when placed at the center of magnetic coils. At this stage of development, cells were imaged with a brightfield upright miscroscope at a rate of 1fps, one single cell at a time, and we successfully measured the rotation

rate of the cells by analyzing the fluctuation of intensity of specific areas of the captions. Chapter 1 establishes the relationship between morphology changes and the changes in rotation rate, and the ability to use the CM method as a real-time drug sensitivity test.

Chapter 3 introduces several breakthroughs in the CM method. First, the quality of cell imaging is greatly improved using fluorescence microscopy, all the while moving from a 40X to a 10X objective. As the surface captured by the camera significantly increases, multiplexing becomes possible by trapping individual cells in the microwells of a microfluidic device. However, Chapter 2 shows the switch of purpose of the CM method. More than just a cytotoxic assay, it is sown that the method can also be used as a powerful system to analyze the dynamic changes of cells morphologies. This chapter also paves the way for more automation in the process. Notably, instead of the rotation rate, which is related to a single cell parameter (its "effective volume", V_{eff}) we measure many more mathematical parameters describing the morphological features of single cells, features extracted using object recognition algorithms. The 100-fold increase of cells available for measurements and of the measurements themselves also called for a distinctive analytical process. Measurements are processed in parallel on a High Performance Computing system (Flux at University of Michigan), and the high volume, high dimensional data sets are treated with *supervised* Machine Learning algorithms. This serves as a basis for the classification of morphological phenotypes inside a cells population, and relationship graph theory is used to visualize and find the different morphological clusters and behaviors present in the population.

In Chapter 4, automation and multiplexing go a step further. The CM setup is once more modified, with an automated and programmable motorized microscope stage and a high resolution monochromatic camera. The goal in Chapter 4 is to compare two populations of cancer cells: an epithelial Prostate Carcinoma cell line, and its malignant variant, the mesemchymal one, and build a classifier to identify the cells, based on their morphological features. Compared to Chapter 3, we managed a 20-fold increase in the number of cells simultaneously imaged and analyzed. In addition, additional features are measured, such as texture related features. Finally, *unsupervised* Machine Learning algorithms are now utilized, and this does not require the intervention of a human operator anymore. Chapter 4 shows that the phenotypes are correctly analyzed with a 94% accuracy.

In summary, the work in this dissertation shows the development and improvement of the Cell Magneto-rotation method as a diagnostic and identification tool for individual metastatic cancer cells in suspension. Combined with CTCs capture systems (such as filters) or biopsies, it can lead to the buildup of a new type of prognosis and evaluation of the evolution a cancer malignancy during therapy.

The fifth and last chapter of this dissertation presents our conclusions regarding the development of the magneto-rotation method, as well as future directions for its improvements and applications. Potential challenges and necessary steps towards the establishment of a reliable diagnostic technique are also discussed in this chapter.

References

1. Nguyen DX, Bos PD, Massagué J (2009) Metastasis: from dissemination to organ-specific colonization. Nat Rev Cancer 9: 274–284. doi:10.1038/nrc2622.

2. Yao D, Dai C, Peng S (2011) Mechanism of the Mesenchymal–Epithelial Transition and Its Relationship with Metastatic Tumor Formation. Mol Cancer Res 9: 1608–1620. doi:10.1158/1541-7786.MCR-10-0568.

3. Wong IY, Javaid S, Wong EA, Perk S, Haber DA, et al. (2014) Collective and individual migration following the epithelial–mesenchymal transition. Nat Mater 13: 1063–1071. doi:10.1038/nmat4062.

4. Bomken S, Fišer K, Heidenreich O, Vormoor J (2010) Understanding the cancer stem cell. Br J Cancer 103: 439–445. doi:10.1038/sj.bjc.6605821.

5. Boman BM, Wicha MS (2008) Cancer Stem Cells: A Step Toward the Cure. J Clin Oncol 26: 2795–2799. doi:10.1200/JCO.2008.17.7436.

6. Clevers H (2011) The cancer stem cell: premises, promises and challenges. Nat Med: 313–319. doi:10.1038/nm.2304.

7. Hüsemann Y, Geigl JB, Schubert F, Musiani P, Meyer M, et al. (2008) Systemic Spread Is an Early Step in Breast Cancer. Cancer Cell 13: 58–68. doi:10.1016/j.ccr.2007.12.003.

8. Korkaya H, Wicha MS (2007) Selective targeting of cancer stem cells: a new concept in cancer therapeutics. BioDrugs 21: 299–310.

9. Kuçi S, Wessels JT, Bühring HJ, Schilbach K, Schumm M, et al. (2003) Identification of a novel class of human adherent CD34- stem cells that give rise to SCID-repopulating cells. Blood 101: 869–876.

10. Longo D, Hasty J (2006) Dynamics of single-cell gene expression. Mol Syst Biol 2. Available: http://dx.doi.org/10.1038/msb4100110. Accessed 15 July 2010.

11. Krutzik PO, Crane JM, Clutter MR, Nolan GP (2008) High-content single-cell drug screening with phosphospecific flow cytometry. Nat Chem Biol 4: 132–142. doi:10.1038/nchembio.2007.59.

12. Valet G (2006) Cytomics as a new potential for drug discovery. Drug Discov Today 11: 785–791. doi:10.1016/j.drudis.2006.07.003.

13. Snijder B, Sacher R, Ramo P, Damm E-M, Liberali P, et al. (2009) Population context determines cell-to-cell variability in endocytosis and virus infection. Nature 461: 520–523. doi:10.1038/nature08282.

14. Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, et al. (2008) Dynamic Proteomics of Individual Cancer Cells in Response to a Drug. Science 322: 1511–1516. doi:10.1126/science.1160165.

15. Tay S, Hughey JJ, Lee TK, Lipniacki T, Quake SR, et al. (2010) Single-cell NF-[kgr]B dynamics reveal digital activation and analogue information processing. Nature 466: 267–271. doi:10.1038/nature09145.

16. Lahav G, Rosenfeld N, Sigal A, Geva-Zatorsky N, Levine AJ, et al. (2004) Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat Genet 36: 147–150. doi:10.1038/ng1293.

17. Abbott A (2003) Cell culture: Biology's new dimension. Nature 424: 870–872. doi:10.1038/424870a.

18. Justice BA, Badr NA, Felder RA (2009) 3D cell culture opens new dimensions in cell-based assays. Drug Discov Today 14: 102–107.

19. Mogilner A, Keren K (2009) The Shape of Motile Cells. Curr Biol 19: R762–R771. doi:10.1016/j.cub.2009.06.053.

20. Chen J, Li Y, Yu T-S, McKay RM, Burns DK, et al. (2012) A restricted cell population propagates glioblastoma growth after chemotherapy. Nature. Available: http://www.nature.com/doifinder/10.1038/nature11287. Accessed 2 August 2012.

21. Driessens G, Beck B, Caauwe A, Simons BD, Blanpain C (2012) Defining the mode of tumour growth by clonal analysis. Nature. Available: http://www.nature.com/doifinder/10.1038/nature11344. Accessed 2 August 2012.

22. Schepers AG, Snippert HJ, Stange DE, van den Born M, van Es JH, et al. (2012) Lineage Tracing Reveals Lgr5+ Stem Cell Activity in Mouse Intestinal Adenomas. Science. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1224676. Accessed 2 August 2012.

23. Luzzi KJ, MacDonald IC, Schmidt EE, Kerkvliet N, Morris VL, et al. (1998) Multistep Nature of Metastatic Inefficiency. Am J Pathol 153: 865–873.

24. Kienast Y, Von Baumgarten L, Fuhrmann M, Klinkert WEF, Goldbrunner R, et al. (2009) Real-time imaging reveals the single steps of brain metastasis formation. Nat Med 16: 116–122.

25. Stoletov K, Kato H, Zardouzian E, Kelber J, Yang J, et al. (2010) Visualizing extravasation dynamics of metastatic tumor cells. J Cell Sci 123: 2332. doi:10.1242/jcs.069443.

26. Wolf K, Mazo I, Leung H, Engelke K, von Andrian UH, et al. (2003) Compensation mechanism in tumor cell migration: mesenchymal-amoeboid transition after blocking of pericellular proteolysis. J Cell Biol 160: 267–277. doi:10.1083/jcb.200209006.

27. Lämmermann T, Sixt M (2009) Mechanical modes of "amoeboid"cell migration. Curr Opin Cell Biol 21: 636–644.

28. Hawkins RJ, Piel M, Faure-Andre G, Lennon-Dumenil A, Joanny J, et al. (2009) Pushing off the walls: a mechanism of cell motility in confinement. Phys Rev Lett 102: 58103.

29. Sahai E, Marshall CJ (2003) Differing modes of tumour cell invasion have distinct requirements for Rho/ROCK signalling and extracellular proteolysis. Nat Cell Biol 5: 711–719.

30. Friedl P, Wolf K (2003) Tumour-cell invasion and migration: diversity and escape mechanisms. Nat Rev Cancer 3: 362–374. doi:10.1038/nrc1075.

31. Prudent R, Vassal-Stermann E, Nguyen CH, Pillet C, Martinez A, et al. (2012) Pharmacological inhibition of LIM Kinase stabilizes microtubules and inhibits neoplastic growth. Cancer Res 72: 4429–4439.

32. Fantner GE, Barbero RJ, Gray DS, Belcher AM (2010) Kinetics of antimicrobial peptide activity measured on individual bacterial cells using high-speed atomic force microscopy. Nat Nano 5: 280–285. doi:10.1038/nnano.2010.29.

33. Zhang H, Liu K-K (2008) Optical tweezers for single cells. J R Soc Interface R Soc 5: 671–690. doi:10.1098/rsif.2008.0052.

34. Hochmuth RM (2000) Micropipette aspiration of living cells. J Biomech 33: 15–22.

35. Thoumine O, Ott A (1997) Time scale dependent viscoelastic and contractile regimes in fibroblasts probed by microplate manipulation. J Cell Sci 110: 2109–2116.

36. Guck J, Ananthakrishnan R, Mahmood H, Moon TJ, Cunningham CC, et al. (2001) The optical stretcher: a novel laser tool to micromanipulate cells. Biophys J 81: 767–784.

37. Guck J, Lautenschläger F, Paschke S, Beil M (n.d.) Critical review: cellular mechanobiology and amoeboid migration. Integr Biol 2: 575–583. doi:10.1039/C0IB00050G.

38. Gabriele S, Benoliel AM, Bongrand P, Théodoly O (2009) Microfluidic investigation reveals distinct roles for actin cytoskeleton and myosin II activity in capillary leukocyte trafficking. Biophys J 96: 4308–4318.

39. Brody JP, Han Y, Austin RH, Bitensky M (1995) Deformation and flow of red blood cells in a synthetic lattice: evidence for an active cytoskeleton. Biophys J 68: 2224–2232.

40. Lam WA, Rosenbluth MJ, Fletcher DA (2007) Chemotherapy exposure increases leukemia cell stiffness. Blood 109: 3505.

41. Franz CM, Taubenberger A (2012) AFM-Based Single-Cell Force Spectroscopy. In: Baró AM, Reifenberger RG, editors. Atomic Force Microscopy in Liquid. Wiley-VCH Verlag GmbH & Co. KGaA. pp. 307–330. Available: http://onlinelibrary.wiley.com/doi/10.1002/9783527649808.ch12/summary. Accessed 8 December 2014.

42. Ikai A, Afrin R, Watanabe-Nakayama T, Machida S-I (2012) Nanosurgical Manipulation of Living Cells with the AFM. In: Baró AM, Reifenberger RG, editors. Atomic Force Microscopy in Liquid. Wiley-VCH Verlag GmbH & Co. KGaA. pp. 331–354. Available: http://onlinelibrary.wiley.com/doi/10.1002/9783527649808.ch13/summary. Accessed 8 December 2014.

43. Moreno-Herrero F, Gomez-Herrero J (2012) AFM: Basic Concepts. In: Baró AM, Reifenberger RG, editors. Atomic Force Microscopy in Liquid. Wiley-VCH Verlag GmbH & Co. KGaA. pp. 1–34. Available: http://onlinelibrary.wiley.com/doi/10.1002/9783527649808.ch1/summary. Accessed 8 December 2014.

44. Johnson BN, Mutharasan R (2012) Biosensing using dynamic-mode cantilever sensors: A review. Biosens Bioelectron 32: 1–18. doi:10.1016/j.bios.2011.10.054.

45. Ozkan M, Wang M, Ozkan C, Flynn R, Esener S (2003) Optical Manipulation of Objects and Biological Cells in Microfluidic Devices. Biomed Microdevices 5: 61–67. doi:10.1023/A:1024467417471.

46. Kovac JR, Voldman J (2007) Intuitive, Image-Based Cell Sorting Using Optofluidic Cell Sorting. Anal Chem 79: 9321–9330. doi:10.1021/ac071366y.

47. Dochow S, Krafft C, Neugebauer U, Bocklitz T, Henkel T, et al. (2011) Tumour cell identification by means of Raman spectroscopy in combination with optical traps and microfluidic environments. Lab Chip 11: 1484–1490. doi:10.1039/C0LC00612B.

48. Ashcroft RG, Lopez PA (2000) Commercial high speed machines open new opportunities in high throughput flow cytometry (HTFC). J Immunol Methods 243: 13–24.

49. Chikazumi S, Taketomi S, Ukita M, Mizukami M, Miyajima H, et al. (1987) Physics of magnetic fluids. J Magn Magn Mater 65: 245–251. doi:10.1016/0304-8853(87)90043-6.

50. Yang L, Peng X-H, Wang YA, Wang X, Cao Z, et al. (2009) Receptor-Targeted Nanoparticles for In vivo Imaging of Breast Cancer. Clin Cancer Res 15: 4722–4732. doi:10.1158/1078-0432.CCR-08-3289.

51. Yang L, Mao H, Cao Z, Wang YA, Peng X, et al. (2009) Molecular imaging of pancreatic cancer in an animal model using targeted multifunctional nanoparticles. Gastroenterology 136: 1514–1525.e2. doi:10.1053/j.gastro.2009.01.006.

52. Yang J, Gunn J, Dave SR, Zhang M, Wang YA, et al. (2008) Ultrasensitive detection and molecular imaging with magnetic nanoparticles. The Analyst 133: 154. doi:10.1039/b700091j.

53. Kaushik A, Khan R, Solanki PR, Pandey P, Alam J, et al. (2008) Iron oxide nanoparticles–chitosan composite based glucose biosensor. Biosens Bioelectron 24: 676–683. doi:10.1016/j.bios.2008.06.032.

54. Sun S, Murray CB, Weller D, Folks L, Moser A (2000) Monodisperse FePt Nanoparticles and Ferromagnetic FePt Nanocrystal Superlattices. Science 287: 1989–1992. doi:10.1126/science.287.5460.1989.

55. Rabias I, Tsitrouli D, Karakosta E, Kehagias T, Diamantopoulos G, et al. (2010) Rapid magnetic heating treatment by highly charged maghemite nanoparticles on Wistar rats exocranial glioma tumors at microliter volume. Biomicrofluidics 4: 024111. doi:10.1063/1.3449089.

56. Janssen XJA, Schellekens AJ, van Ommering K, van Ijzendoorn LJ, Prins MWJ (2009) Controlled torque on superparamagnetic beads for functional biosensors. Biosens Bioelectron 24: 1937–1941.

57. Kinnunen P, Sinn I, McNaughton BH, Newton DW, Burns MA, et al. (2011) Monitoring the growth and drug susceptibility of individual bacteria using asynchronous magnetic bead rotation sensors. Biosens Bioelectron 26: 2751–2755. doi:10.1016/j.bios.2010.10.010.

58. Turitto VT (1982) Blood viscosity, mass transport, and thrombogenesis. Prog Hemost Thromb 6: 139–177.

Chapter 2 :
Nanoparticle induced Cell Magneto-Rotation: Monitoring Morphology, Stress and Drug Sensitivity of a Suspended Single Cancer Cell.

This chapter has been adapted with minor modifications from the following published article:

Abstract

Single cell analysis has allowed critical discoveries in drug testing, immunobiology and stem cell research [2–5]. In addition, a change from two to three dimensional growth conditions radically affects cell behavior. This already resulted in new observations on gene expression and communication networks and in better predictions of cell responses to their environment [6–8]. However, it is still difficult to study the size and shape of single cells that are freely suspended, where morphological changes are highly significant [9]. Described here is a new method for quantitative real time monitoring of cell size and morphology, on single live suspended cancer cells, unconfined in three dimensions. The precision is comparable to that of the best optical microscopes, but, in contrast, there is no need for confining the cell to the imaging plane. The here first introduced *cell magnetorotation* (CM) method is made possible by *nanoparticle induced cell magnetization*. By using a rotating magnetic field, the magnetically labeled cell is actively rotated, and the rotational period is measured in real-time. A change in morphology induces a change in the rotational period of the suspended cell (e.g. when the cell gets bigger it rotates slower). The ability to monitor, in real time, cell swelling ordeath, at the single cell level, is demonstrated. This method could thus be used for multiplexed real time single cell morphology analysis, with implications for drug testing, drug discovery, genomics and three-dimensional culturing.

Introduction

The heterogeneity, i.e. non-uniformity, found in cancer cell populations, and the ubiquitous cell differentiation, has led to increased interest in individual cell studies. Historically, a tumor was thought to originate from the successive divisions of a single 'mother cell', leading to the assumption that all the cells in a tumor shared the same genetic code. However, recent findings have altered this theory, stressing the need for tools that can monitor and track single cells in a high throughput fashion. Currently, standard assays performed on cell populations make individual patterns difficult to access, due to effects of averaging [10]. Flow cytometry, for instance, has been massively used in the last 20 years, for its ability to perform fast analysis on a very high number of cells at a time (10000 cells/s). Time point analysis can also be performed using this technique, but it is not possible to track each cell individually.

Then again, it is especially important that even a small minority of cells, such as stem cells, whose behavior could be considered to be statistically irrelevant compared to the large majority of the population, can have a critical biological and medical impact. For instance, the use of the *Imatinib* drug that targets the *BCR-abl* fusion protein in patients with chronic myelogenous leukemia (CML) first seemed to be one of the most successful targeted therapies. However, the treatment does not eliminate the CML stem cells, and with the withdrawal of Imatinib the disease reappeared [11,12]. As a consequence, the focus on cell-to-cell variations has also allowed important breakthroughs in the understanding of cell differentiation, drug response, protein mechanisms and dynamics, as well as of the important role played by stem cells, especially for cancer stem cells [13]. Metastasis relies on cancer cells circulating in the vascular network. The cells responsible for cancer propagation to secondary tumor sites are extremely rare (a few cells per million in the blood), and they go through a circulating stage before populating other tissues. Therefore, along with single cell analysis, three dimensional assays also permit a better comprehension of cellular dynamics [14–16], by narrowing the gap between *in vitro* and *in vivo* behavior [8]. However, all previously mentioned single cell analysis techniques are restricted by

their confinement of the cell in two dimensions. To overcome this limitation, we employ a new approach using *suspended cell magneto-rotation* (CM).

Specifically, we use a *nanoparticle induced Cell Magneto-rotation method*, where the driving magnetic field and the rotating cell are *out-of-synch* with each other. The cells are embedded with 30 nm commercial magnetic nanoparticles (Ocean Nanotech®) and are rotated under an external magnetic field of about 1mT, at about100 Hz. We note that a thousand times (1000x) higher fields, on the order of 1T, are used for MRI. Also, magnetic nanoparticles have been widely used in biology [17–21]. Thus the CM method is designed to be biocompatible and non-toxic. The live cell is rotated *asynchronously* (see **Figure 2.2**) in suspension, and its rotational frequency is highly sensitive to any morphology change. As reported here, magneto-rotation does not affect the cell's viability, and allows for real time analysis to be performed. Changes in cell morphology are indicated quantitatively by the single cell's rotation period. The trends in the rotation rate allow discrimination between a healthy cell, a dying cell or a swelling cell. In addition, this new technique is easily adaptable to any microscope set-up, is fluorescent-label free, and is compatible with simultaneous fluorescence and/or other optical imaging and spectroscopy methods as well as magnetic separation and enrichment techniques. Other methods used to track morphological changes of single biological cells include Atomic Force Microscopy [22] (AFM) and Optical Tweezers [23] (OT). These methods may offer higher resolution, but are limited by the attachment of cells to a surface (AFM), or by the irreversible damage caused by laser trapping (OT). Furthermore with OT, for each cell line, viability studies have to be done for each cell type in order to prevent photodamage, which limits its applicability [24]. The use of cantilevers has also been reported to track the mass of live cells [25], but there are no publications yet on single cancer cells in suspension.

Results

Model for the rotation of magnetically labeled cells

To verify that cells could be magnetically manipulated, we placed them in the center of magnetic coils with magnetic field amplitudes of 1mT, as shown in **Figure 2.1b**. The coils themselves are adapted to the platform of a microscope in order to record videos (see **Figure 2.1d**). The single cells rotate at frequencies ranging from 0.05 Hz to 2 Hz in this setup (much lower than the 100 Hz driving fields, due to operating in the *asynchronous regime*, see below). Focusing a low power, 1.45 mW HeNe laser through the microscope, the forward scattered signal is recorded with a photodiode [26]. The cell viability is not affected by this low-intensity laser, as shown in **Figure 2.7**. When the cell rotates, it produces rotational-dependent modulation that can be measured with the photodiode. With real-time signal processing, the rotation period of the cell and therefore its size/morphology can be monitored in real-time.

a

CCD

Condenser

He-Ne Laser

Filter

Microscope

Computer Workstation

Function
Generator
(cos)

Amplifier

Coils

Objective

Function
Generator
(sin)

Cell Plate

Bright
Field Lamp

Photodetector

b

1

2

3

Electromagnet

Photodetector

Magnetized Cell

Iron Oxide
Nanoparticle

c

Rotation Period (s)

1.80

1.55

1.30

Voltage (mV)

0.32

0.29

100   102   104   106   108   110
Time (s)

100   130   160   190   220   250

Time (s)

d

Figure 2.1: Magneto-Rotation of a single cell. **a)** Schematics of the complete setup. A Live Cell Array® plate, with 100µm wells, is placed on the platform of a microscope, for which a set of electromagnetics has been adapted. Note that the cell is not stuck to the bottom of the well. Under the 60x objective, the laser beam undergoes forward scattering from the rotating cell (15 to 20µm), and the variations in the forward scattered light is captured in real-time by a photo-detector, and analyzed on a computer. **b)** Schematics of a rotating cell placed inside the magnetic coils: two identical sinusoidal signals, with a phase shift of 90°, pass through the two pairs of coils. The applied magnetic field and the magnetic moment of the cell are not aligned, creating a torque that drives the cell's rotation. **c)** Rotational period of a fixated cell in DMEM. The inset represents the raw signal from the photodetector, showing the periodicity over a given time window. The treatment of the signal then gives the rotational period (See **Methods** section on the optical setup for the signal treatment description). **d)** Caption of the setup. Custom Helmoltz Coils with NUNC Live Cell Array Plate on the microscope stage.

The cell is found to exhibit magnetic rotational behavior very similar to that of a magnetic microparticle (**Figure 2.2**). As shown by McNaughton et al. [27], and extended to the case of superparamagnetic particles [28], there exists a critical frequency of the external magnetic field above which the particle does not rotate synchronously with the field, i.e. the particle cannot keep up anymore with the driving frequency. In this asynchronous regime, the mean value of the rotation speed of the single cell is given by $\left\langle \dfrac{d\theta}{dt} \right\rangle = \dfrac{Torque}{Drag} = \dfrac{\Gamma}{\kappa\eta V}$, where Γ is the magnetic torque and $\kappa\eta V$ is the drag due to viscosity forces. Here, κ is its Einstein's shape factor, V the volume and η the coefficient of viscosity. We note that Γ is proportional to the magnitude of the magnetic field, the magnetic moment of the cell and the volume of the magnetic *contents* of the cell; however, all these parameters are kept constant in the experiments. Therefore, in the asynchronous regime, any change in the cell's shape or volume, i.e. in its effective volume, $V_{eff} = \kappa V$, induces a change in the rotation speed, given by the above formula. This model has been further refined for the case of paramagnetic particles [29,30], wherein the rotational period, T, is found to be proportional to the effective volume, $T \propto V_{eff}$ (this is true in the asynchronous rotational regime; for a complete derivation, see ref. 27 and equations in introductory chapter). As can be seen from this dependence, if the volume increases, the rotation period increases proportionally. The same goes for the shape factor, and, as a consequence, one can detect morphology changes.

Magnetic behavior of the cells.

To first explore the incubated cells' magnetic behavior, cells were fixated with a solution of 3.7% formaldehyde in phosphate buffered saline. This way, the cross-linking of the proteins, both in the membrane and in the cytoplasm, removed any potential effects of changes to the drag, resulting from morphology changes. A single cell was thus tested under different field frequencies, keeping the other conditions constant. The cell mimicked the magnetic response behavior of a superparamagnetic micro-bead that was manufactured in an analogous way, e.g. iron-oxide nanoparticles (magnetite) loaded into a polystyrene bead, such as the DynaBeads™. The magnetic response is shown in **Figure 2.2**. At frequencies ranging from 1Hz to 1 kHz, cells rotate in an asynchronous manner [27]. Indeed, while the field rotates at frequencies above 1Hz, cells rotate at much lower frequencies. Being in the asynchronous regime is crucial to the described magneto-rotation method. When in this regime, any small change in the cell shape (or in the liquid medium's viscosity) immediately translates into a change in rotational period. Otherwise, in the synchronous regime, the cell would keep the same rotation rate, i.e. would faithfully follow the driving magnetic field, with an identical rotation frequency, irrespective of cell or medium changes. As previously reported, we note the presence of a maximum synchronous rotational frequency beyond which the rotation frequency of the cell decreases with an increase in the applied driving frequency.

Given that the cell is already in the asynchrous regime at driving frequencies of 1-1,000 Hz, we can set the applied field frequency so that the asynchronous cell rotation frequency is at its highest (which occurs for a driving frequency at about 100 Hz), compared to other applied frequencies. This way, the effects of surface friction on rotation are reduced to a minimum, and the sensitivity to drag and shape increases, as well as the real time resolution.
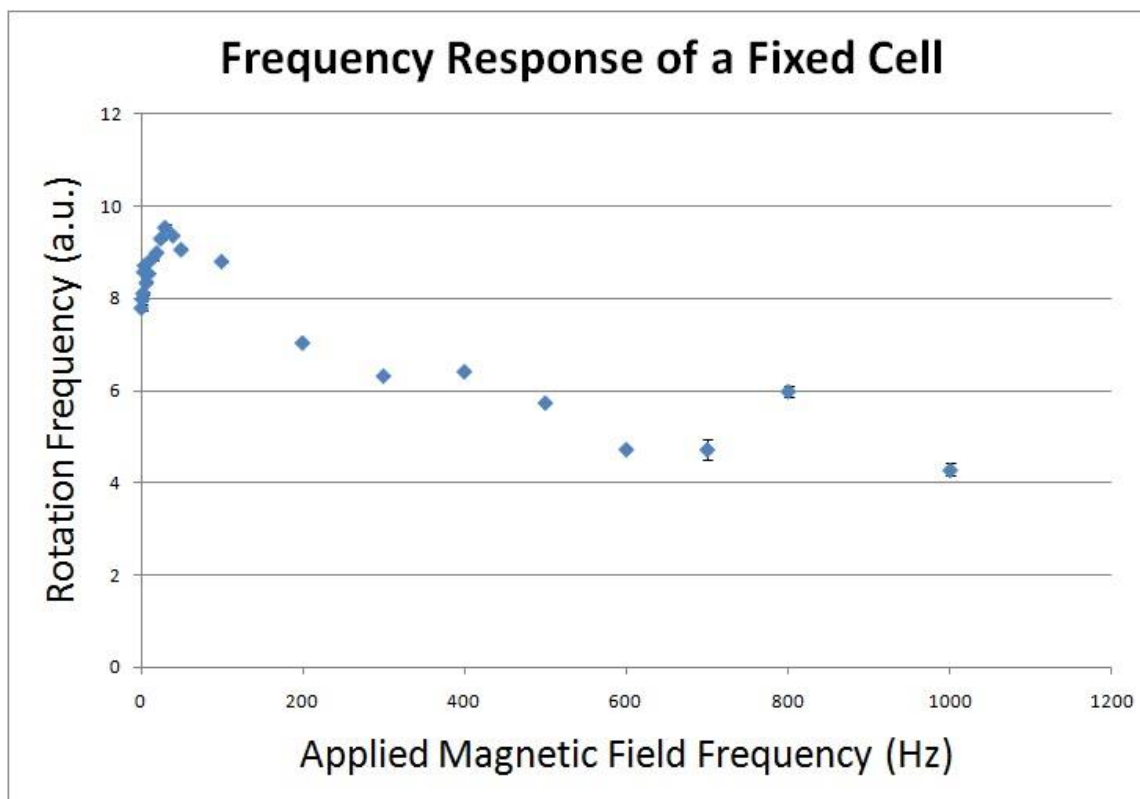
Figure 2.2: Frequency response of a fixated cell (error bars are inside the dots, values represent mean +/- 0.5*s.d. , n=18).

Magnetic characterization of the cells

To characterize furthermore the magnetization of the cells, we looked at the localization of the nanoparticles after incubation, to determine whether they stayed attached to the surface, got internalized, and, if they did, if the nanoparticles were free to move in the cytoplasm or trapped in vesicles (endosomes). To do so, we attached HPTS fluorescent dyemolecules (8 - Hydroxypyrene - 1,3,6 - trisulfonic acid, trisodium salt) to our nanoparticles, using the electrostatic attraction forces between the particles and the dyes. HPTS is a membrane impermeant dye, and thus it needs a vector to get internalized by the cells. Following the standard protocol of incubation, we washed the cells three times in PBS, and the cells were observed under excitation at 450nm with fluorescence being checked at 510nm. The results are shown **figure 2.3a**. As we can see, the magnetic nanoparticles are internalized by the cell through

endocytosis. In addition, neither the nucleus nor the cytoplasm shows fluorescence, which indicates that the nanoparticles remain in the vesicles. Moreover, we assessed the iron content of the cells by Inductively Coupled Plasma (ICP) measurement (see **Methods** section). As expected, the iron content increases with the MNP concentration in the culture media, and the trend appears to be linear in the concentration window that we used (**figure 2.3b**). For our rotation experiments, we estimated that the iron content is around 14 pg/cell. Compared to the mass of a nanoparticle, this means that, on average, less than 20,000 nanoparticles have gotten into the cell. Other sizes of magnetic nanoparticles were also tested (10nm, 100 nm and 200 nm), but internalization was maximized for particles with a diameter of 30 nm (data not shown).
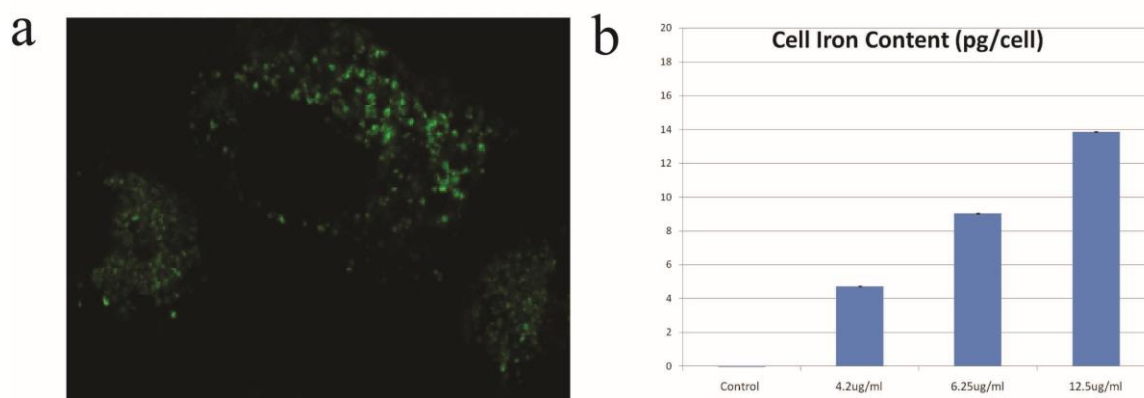
Figure 2.3: **a)** Fluorescence Image (40x) of a HeLa cells after incubation with dyed magnetic nanoparticles at an extracellular iron concentration of [Fe] = 12.5ug/ml (0.22mM). **b)** Cellular iron content in picogram per cell. The concentrations of particles in the media are given in iron concentration (error bars values represent mean +/- 0.5*s.d. , n=3)

Cytotoxicity assay and drug sensitivity

In this study, cancer cells loaded with nanoparticles were magnetically separated and resuspended in different media, such as culture medium (DMEM), DMEM with 5% Ethanol, DMEM with 100ug/ml Cisplatin or DMEM with 75% deionifzed water. Each medium was used to verify different aspects of this method: DMEM was used as a control, ethanol was used as a cytotoxic agent, Cisplatin was used to model a drug assay; also, to promote stress through cell

swelling, we used a large proportion of DI water, reversing the ionic balance between the inside and outside of the cell. Note that a large concentration of salt in solution has the opposite effect on the cell, namely shrinking it. The cells in suspension were then pipetted onto a Live Cell Array™ plate (NUNC™), where the array has 100 μm wide wells, which provide adequate compartments for single cells to rotate and be analyzed. Optical scattering signals (from the rotating cells) were recorded and the changes in the rotation period were measured for the different media (**figure 2.4**). Magneto-rotation was performed under numerous conditions, with different cell samples; the following results show typical examples of cell behavior that have been reproduced multiple times in our system.
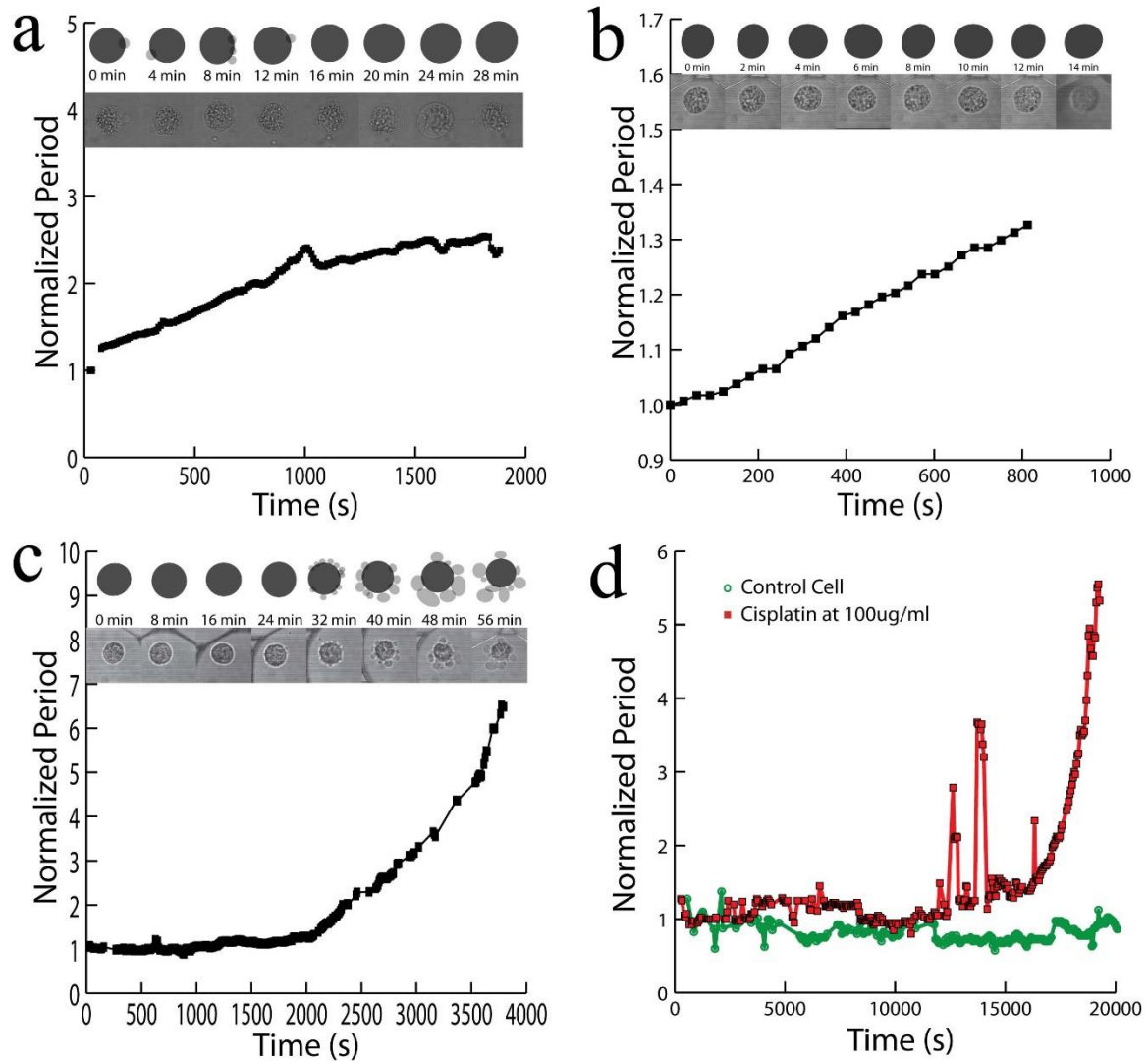
Figure 2.4: Changes in the rotation period of a single HeLa cell **a)** In DMEM on an agarose layer **b)** In a mixture of 75% DI water and 25% DMEM **c)** In a mixture of DMEM with 5% Ethanol and **d)** for a live cell in DMEM (green circles) compared to a HeLa cell (red squares) in DMEM with a 100ug/ml of Cisplatin. The Y axis is the normalized period, and the X axis is time in seconds. Lines show trend between connected points. For each graph, in the pictures above it, the bottom pictures show snapshots of the rotated cell at each indicated time, while the schematic pictures on top of it show the corresponding cell shapes (fixated cell not shown). Dark discs represent the cell cytoplasm and membrane, while grey spots show the vesicles formed at the surface, if any.

Figures **2.4a** and **2.4b** show two cases of cell swelling. Cell swelling generally occurs because of the osmotic pressure created either by an ionic imbalance, as mentioned earlier, or by a lack of nutrients. Either way, the cell expands to cope with the imbalance of the chemicals it needs for maintaining its metabolism. To reach ionic disparity, we used DI water (**figure 2.4b**). We also observed that cells would also swell when placed on an agarose layer (2% agarose in DI water) (**figure 2.4b**). Agarose gel is porous, a property that is used in the electrophoresis of proteins, and this property might be at the origin of the swelling. Indeed, the nutrients present in the growth media, mainly glucose, can diffuse into the agarose gel while the cells rotate above it. The cells would therefore swell to balance the reduced concentration of nutrients available in solution, as observed by Goldberg et al. in cortical cells [30]. Since the cell volume increases, the rotation period increases. Alternatively, cell death is provoked when placed in a solution with 5% ethanol (**figure 2.4c**) or using a concentration of 100 ug/ml of Cisplatin in solution (**figure 2.4d**, red line connecting squared dots). However, the mechanisms of these kinds of cell deaths are different from the cases above, since blebs appear at the surface of the cell. In 5% ethanol, it takes only around 30 minutes (**figure 2.4c**) for blebs to appear, while in the case of the treatment by Cisplatin at 100ug/ml, it takes several hours. Contrary to the swelling case, it is the changes in shape of the cell membrane that increase the effective volume. Blebbing and the formation of vesicles at the surface of the cell indicate that the cell contents are being broken down and separated into several vesicles. As the death process continues, the vesicle sizes increase. This kind of phenomenon does not only add to the volume, but it critically affects the shape factor of the cell. The combination of these two parameters, namely the *effective volume*, is what is tracked with magnetorotation, thus amplifying the blebbing effect. Eventually, the drag on the cell becomes so high, compared to the initial state of the cell, that the cell rotation period rises drastically (by 550%), in a non-linear way (see **figures 2.4c**, red line on **figure 2.4d** and see **figure**

30

2.5 for a comparison with microscope measurements). Thus both cell death mechanisms, though very different, can be observed and differentiated with Cell Magnetorotation.
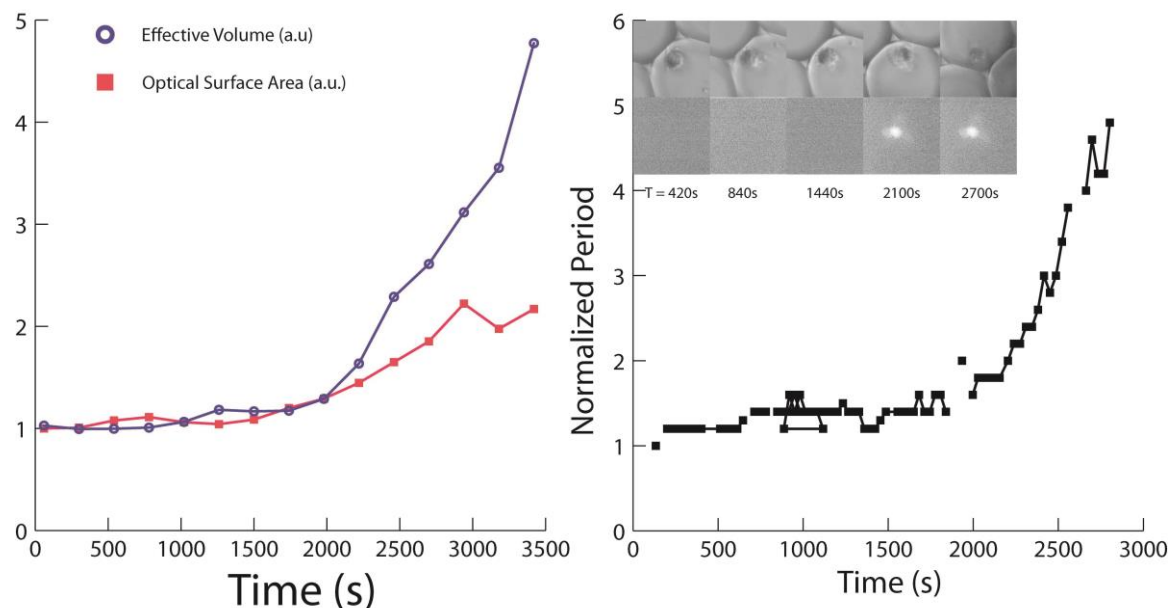
Figure 2.5: **a)** Comparison of sensitivities between microscope and magneto-rotation in measuring cell death (HeLa cell in DMEM, 5% Ethanol). In red is the normalized surface area as measured with the microscope, and in blue is the normalized effective volume period as measured with Magneto-Rotation using **Supplementary Equation 7**. **b)** Comparison of cell death monitoring using magneto-rotation and Live/Dead cell assay.

We also performed magnetorotation of a healthy cell (**figure 2.4d, green line**), in growth media. In the absence of a toxic agent, the rotation period did not significantly change (the standard deviation of the rotation period was 15%). A fixed morphology control test was realized by fixating the cells in a 4% formaldehyde vial (1.5 ml) for 10 min, under end-over-end vial rotation (see **figure 2.6, red line**). Since the membrane and the cell contents were cross-linked, the cell morphology did not change, under isotonic conditions, and thus, as expected, the rotation period did not change. As compared to a fixated cell, where the rotational period is very flat, for live cells we observe that the rotation period, over time, exhibits significant short-time fluctuations. This may be a result of the cell metabolism, which is still active during rotation. Overall, this shows that when the rotation period is constant, it corresponds to a cell that is not significantly changing in its effective volume.
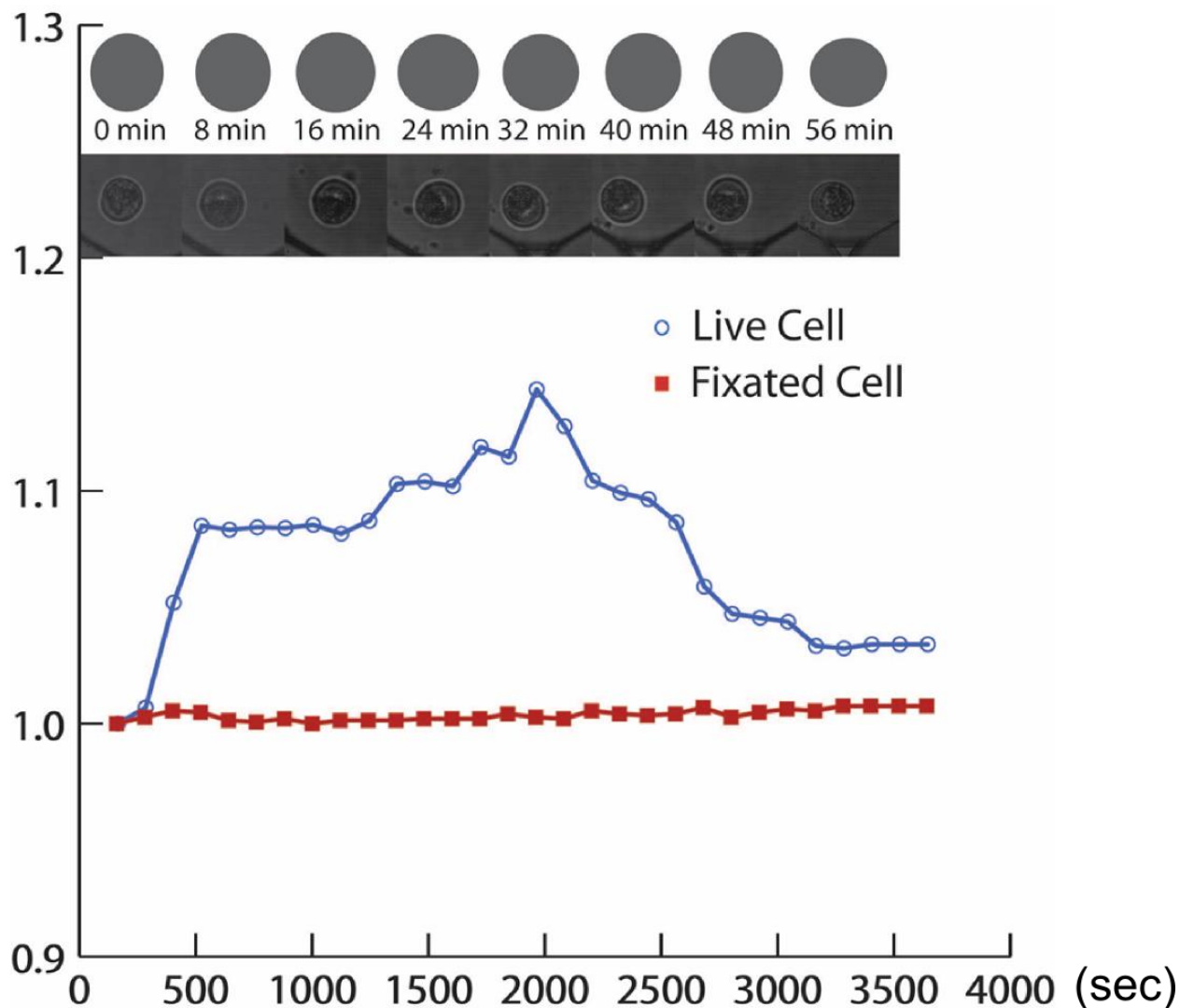
Figure 2.6: **Supplementary Figure S2:** Changes in the rotation period of a single HeLa cell in DMEM (blue circles) compared to a fixated HeLa cell (red squares) in DMEM.

To assess the accuracy of the method regarding effective volume modifications, we compared the trends in the effective volume (proportional to the rotation period) with those of the surface area, as estimated from microscopy images (the surface area being a standard indicator of the cell morphology/shape factor). With an imaging software (Adobe® Photoshop®), we estimated the surface areas of the cells at regularly spaced intervals. As can be seen on **figure**

2.5a, magneto-rotation is as effective as an optical microscopy setup for observing small changes. However, for bigger changes, magneto-rotation amplifies the response, compared to the optical microscope setup. Any significant loss of magnetic content takes several days, according to Arbab et al. [31]. Thus its impact on the interpretation of the results, after several hours, can be ignored. Also, the steady rotation rate of a magnetized control cell tends to confirm that the loss of magnetic content is not significant over the time-span of the measurement. Otherwise, the magnetic moment of the cell would critically decrease, and the cell would slow down significantly, which is not the case (**figure 2.4d**). Therefore, we can safely assume that the cell's effective volume is indeed proportional to the rotational period. Magneto-rotation can also be compared to Live/Dead cell assays. Cells were prepared following the protocol described earlier. Before pipetting into the microwell plate, we added 2.0 ul of calcein and 5.0 ul of propidium iodide (PI) to a 1 ml sample containing cells. Cells were left sit in the incubator for 10 min, and then resuspended in DMEM with 5% ethanol and the same amount of dyes, after which, they were pipetted and rotated. As can be seen on **figure 2.5b**, the cell undergoes morphology changes well before PI fluorescence can be seen, and by the time cell death (PI defined) occurs, the rotation period has slowed down by a factor of about 2. This not only shows that the magneto-rotation method' compares well with fluorescent assays, but also shows it to be more sensitive. It is not surprising to see the rotation rate slowing down *well before* one is being able to detect fluorescence from the PI dyes. Indeed, the PI dyes only make their way into the cytoplasm after the cell walls are have been destroyed. However, well before, other processes take place, one of them being the formation of blebs at the surface of the cell, a phenomenon that cell magneto-rotation can accurately monitor, which is not the case for the MTT assay for instance.

Effects of magneto-rotation on cell viability and division

To investigate the ability of the setup to monitor cell death, without causing cell death, we conducted several viability tests (laser exposure, short term and long term effects of rotation on viability, cell division and clonogenicity).

We first tested the effect of the uptake of magnetic nanoparticles [yellow (RHS) and red (middle) bars in **figure 5a**], and of the presence of a magnetic field, on cell viability [red (middle) and blue (LHS) bars in **figure 5a**]. We performed the viability test on three different HeLa cell populations. After an hour at 37°C, with humidity and CO_2 control, a cell count was made using Trypan blue. There was no significant difference in viability among the three cell groups (**figure 5a**). This shows that neither the incorporation of the particles nor the rotation under a magnetic field affected the cells viability over the time scale of an hour. Indeed, the same kind of magnetic iron oxide nanoparticles are quite commonly used [17,31] to magnetophoretically separate certaincell populations from heterogeneous populations, as well as during MRI scans on patients (for contrast enhancement), without causing harm to cells. In the above viability tests, the field intensity and the magnetic particle concentrations were purposely set at higher values (0.5mT and 40ug/ml) than those described in this paper for magnetorotation (0.1mT and 25ug/ml), in order to keep a safety margin in the protocol.

Another possible concern we addressed is the effect of the laser exposure on the cell's viability (**figure 2.7b**). The viability test shows no significant cell death and no significant difference after two hours, between control cells and magnetic cells that were exposed to the laser. Both the interaction of the cells with light and the possible interaction of the magnetic nanoparticles with the laser do not affect the viability of the cells.

Finally, we investigated the possible impact of the physical rotation of the cells on their viability. Indeed, in order to accurately monitor toxicity effects, cell rotation has to be harmless. **Figure 2.7c** adresses this latter point. Comparing the death rate of rotating cells and the death rate of non-magnetic cells, we found no statistical difference in the two trends (n=4, p=0.245>0.05, F=1.65<5.98 = F_{crit}). In addition, as we observed (data not shown) and as described in other publications [31], cells containing magnetic nanoparticles can be subcultured. Also, to assess the cells' clonogenicity, we performed a clonogenic assay where cells were first magnetically rotated for 24 hours in an incubator, and then let to grow on agarose for three weeks. We found no significant difference between the control samples and the rotated samples (n=3, t = 1.37 < 2.77 = t_{crit}, p = 0.24 > 0.05 = p_{crit}, see **figure 2.8**).
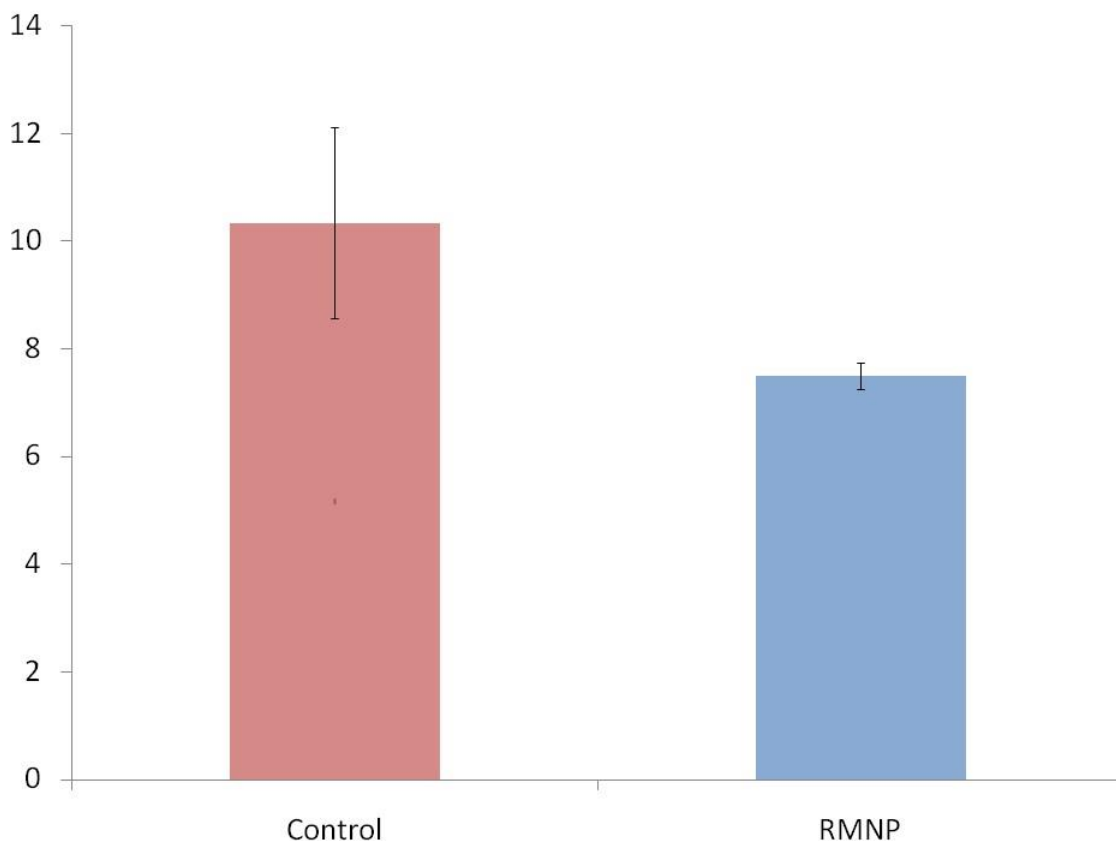
Figure 2.7: Clonogenic assay on HeLa cells incubated with magnetic nanoparticles (12.5ug/ml, unfiltered) and rotated for 24hrs in an incubator. For each sample, after incubation with magnetic nanoparticles following the standard protocol, cells were washed, detached and counted. 10000 cells were then rotated for 24hrs at 37°C, in a 5% CO2 environment with humidity control. Using a 6-well plate, 200 cells were put to grow on an agarose layer (1.3% agarose in DMEM) for 3 weeks. Control cells were not exposed to nanoparticles nor toany magnetic field. Control cells were washed, detached, counted and for each well, 200 cells were put to grow on agarose. Values represent mean +/- 0.5* s.d. n=3.

Effect of rotation on cell division.

Finally, we also tested the effect of magneto-rotation on cellular division. The question was: does magneto-rotation impede immediate cell division? To investigate the short term impact, we rotated cells on agarose for 72 hours, and compared cell growth with two other controls (non labeled and magnetically labeled cells in the absence of magnetic field). We found no difference between the two different groups of magnetically labeled cells (see **figure 2.7**). This also rules out any potential magnetic hyperthermia happening during rotation.

Supported by our observations of cell divisions during (the very slow) rotation and due to the low magnitude of the shear stress, we believe that rotation does not significantly affect cell doubling. To verify this statement in a more quantitative manner, we placed three sets of 100,000 cells in an incubator: a control one (no magnetic labeling, no rotation), and two sets of magnetically labeled cells, one being rotated, and the other one being still. All of the cell samples were put on top of an agarose layer to prevent cell adhesion. After 24 hours at 37oC, 5% CO_2 and humidity control, we observed a loss of population for the magnetically labeled cells (56,000 cells counted for the rotated ones, 60,000 for the non rotated), and only a slight increase (104,000) for the control cells. This is probably due to the change of growth conditions, from adherent plates to a non adherent one. However, after a total of 3 days, all sets showed population growth, with a count of 72,000 for either set of magnetically labeled cells, and 144,000 for the control cells. Therefore, the growth rate was significantly higher for control cells than for the magnetically labeled cells, but between the two sets of magnetically labeled cells, there was no significant difference. From this and the subculturing assay we performed, we can infer that the factors that affect cell growth during rotation are the presence of embedded magnetic nanoparticles in the cells and the growth being in suspension on an agarose layer. In addition, growth in suspension was probably altered because the cells we used were plated cells, and these were not let to adapt to a suspension state. We thus conclude that rotation at sub-hertz frequencies does not affect cell division. This test also adds more evidence regarding the relative harmlessness of the method, given that, after 72 hours, cells were proliferating. Also, the risk of magnetic hyperthermia is nonexistent.
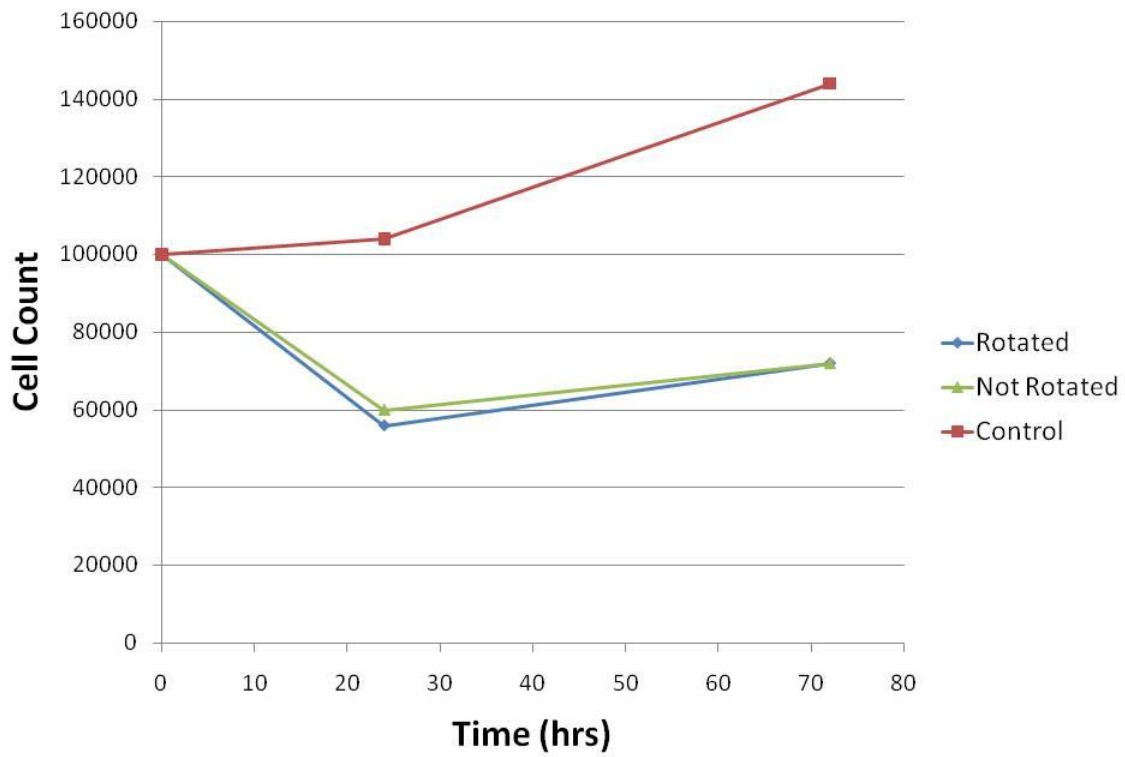
Figure 2.8: Effect of rotation on cell division.
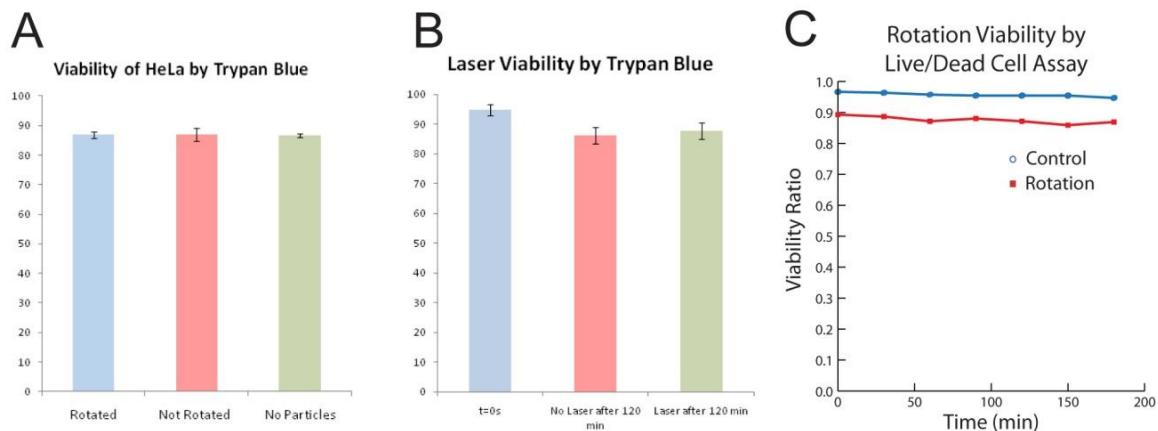
Figure 2.9: **a)** HeLa cells viability after incubation with nanoparticles and rotation under a rotating magnetic field. All the cells came from the same cell line, and were cultured at the same time, each for 4 days. HeLa cells were grown until reaching 70% confluency, and the first sample constituted the control group (RHS). The two other groups, incubated with magnetic nanoparticles, originated from the same cell batch, cells grown in the presence of 40ug/ml in DMEM, until reaching 70% confluency. Each group was made of two samples containing 50,000 cells each. While the second sample was not rotated, the third one (control) was put under a field of 0.5mT and rotated at a driving frequency of 100Hz (LHS). During the experiment, cells were maintained at $37°C$, with 5% CO_2 and humidity control. For every group, n=3. Values represent mean +/-s.d. **b)** Magnetic HeLa cells viability before and after laser exposure. HeLa cells were incubated with magnetic nanoparticles, for 48 hours, following the protocol described before. In a 96-well plate, 150ul of each set of cells was pipetted. Control measurement (blue) was realized after cells were washed, detached and resuspended in fresh media at $37°C$. Non-exposed (red) and exposed cells (green) were kept on the microscope stage for 120 min at room temperature. Each well contained 25,000 cells. Values represent mean +/- 0.5* s.d. n=3. **c)** HeLa cells viability during magneto-rotation at $37°C$, with humidity and 5% CO2 control. HeLa cells were pipetted onto a Live Cell Array (NUNC™). The cells trapped in the 100um wells were counted using Calcein. For both the control and the rotated cells groups, n=4. Cell death was monitored using Propidium Iodide. Standard deviations are within the dots.

Discussion

38

Harmlessness of the method

The use of magnetic nanoparticles and alternating magnetic fields has been commonly associated with hyperthermia, a process where the vibrating nanoparticles inside the cells produce heat, eventually killing the labeled cells through a rise in temperature. As a consequence, the ability to rotate cells through the internalization of similar magnetic nanoparticles and the application of a rotating magnetic field, i.e. alternating in two directions at the same time, without causing harm to the cell has been a concern, even though we are using much lower fields by an order of magnitude, and frequencies in the ranges of a few dozen Hz instead of a few 100 kHz [32,33].

Our first concern was then to assure that the rotation in itself did not kill the cells. Our results show that the viability of the cells is preserved while they are rotated. Also the exposure to a (weak) laser (in order to capture a scattering signal from the rotating cell) does not have any effect on short term cell viability, as shown in **figure 2.9b**. However, the presence of a laser is not necessary, and the signal can also be analyzed through a camera, removing any long time risk that a long term exposure to a laser beam could cause. Indeed, another way to measure the rotation rate is using a CCD camera and measuring the intensity fluctuations of a specific area.

Our results also show that the internalization of magnetic nanoparticles does not cause any effect on cell viability, and it only affects cell division by reducing the growth rate for a short time, over a limited number - at most 3 - of cell cycles, before reaching normal rates. Indeed, our magnetically labeled cells have been successfully subcultured in petri dishes, and we observed no difference in viability or in proliferation rates after three division cycles (data not shown). In accordance with previously published data [34], we also found that magnetically labeled cells grew at a slower rate than non-labeled cells, up until three division cycles, from which point onwards the growth rates were back to normal (see **figure 2.8**). Also, as mentioned, according to Arbab et al. [31] the presence of cell internalized magnetic nanoparticles does not cause deleterious long term effects on the viability of the cells (over a period of 5 to 7 division cycles, i.e. over several weeks).

The presence of a rotating magnetic field, and the induced sub-hertz frequency rotations that were induced in the magnetically labeled cells did not have along term impact on cell division, as shown by our clonogenic assay and by thecell count, after rotating cells for 24 to 72 hours.

Therefore, we have shown that for magnetorotation any cell death observed was the consequence of a purposely-induced toxic environment. In addition, we anticipate that since cells do not die as a result of rotation, cell growth, and even critical dormancy studies could be performed (work in progress). It is worth noting that cell division has been observed during rotation, and rotating cells do not seem to have a different division rate compared to magnetically labeled non-rotating cells (see **figure 2.8**) All in all, the difference in growth rate observed during rotation can be definitely associated with the labeling of the cells with nanoparticles, and not the impact of rotation itself.

This study presents a major difference in cell viability compared, for instance, to the *cell electro-rotation* method, which uses the cytolplasm non-uniformity to induce an electric dipole. The latter, at low frequencies, can cause the rupture of the plasma membrane, resulting in cell death [35].

In conclusion, cell magneto-rotation preserves the viability of the cell, both on a short and long term perspective (3 weeks). The rotation in itself does not affect cell growth. Our results hence demonstrate that if cells are harmed while they are rotated, it is caused by a harmful change in the cells' environment.

Cell magneto-rotation method potential relevance as a cytotoxicity and drug sensitivity assay

As described in the former section, we have demonstrated the ability to monitor cell death using the change in rotation rate of a magnetically labeled cell. The morphology of the cell has been successfully linked to cell fate, since we could associate the formation of blebs during cell death with a significant slowdown in rotation rate. We were also able to characterize cell death with a typical rotation trend, namely the exponential-shaped curve of the rotation period over time. Compared to a live/dead cell assay, we can detect cell slowdown as early as with fluorescence methods, if not earlier. Indeed, blebs are formed while the cell is dying, at a point where the cell membrane is still impermeable to the fluorescent dyes (here, propidium iodide). These results not only show the ability to discriminate cell death from the rotation curve shape, but also the compatibility of the method with a fluorescence assay. To this end, cell magneto-rotation can also be used as a way to maintain single cells in a non-adherent and localized fashion.

Another advantage of the presented method is its ability to track the very same cells over extended period of times. Indeed, fluorescent dyes are subject to photobleaching, affecting the evanescence of the intensity of the light emitted by the dyes. In order to monitor a phenomenon over time, it is then necessary to use different groups of cells that will be stained at different points of time.

As much as cell-to-cell variation can be screened by variations in fluorescence intensity in a cell sample, variations in the trends of cells' rotational periods can also give insights into cell-to-cell variability/heterogeneity. For instance, we can track this heterogeneity not only through the amount of iron-nanoparticles loading into the cell, but also through the time it takes for the rotational period to double under toxic conditions, in a similar fashion to the way the radiation half-life is measured for radio-active atoms. This way, the average « doubling time » will give a frame of reference for the entire cell population, while its distribution among cells in the same population will be a source of information regarding its heterogeneity.

Though we only show here single cells being studied, either separately or at a small throughput (between five to ten at the same time), this study still serves as a proof-of-concept

41

for the method, and our future work will be focused on more robust and perfected multiplexed arrays, with at least a few dozen cells, which would be the relevant quantity regarding circulating tumor cells. Cell magneto-rotation, rather than competing with techniques such as flow cytometry, complements them by extending the reach of the assays to rare cell populations that are naturally found in suspension, and by preserving them in this state while performing the assay.

In this study, our intention was to show that magneto-rotation could potentially be used as a novel method to monitor morphology changes of circulating tumor cells (CTCs)in suspension, at the single cell level. These cells are both very rare and, as stated by their name, are in suspension. They can even circulate in the bloodstream for months or longer [32] without attaching to any surface. This phenomenon, coupled with dormancy and repopulating potential, explains why patients who seemed apparently cured had developed one or several new tumors. In terms of adaptability, this new method can equally be used in serum.

We have here introduced a new method to monitor morphology changes occurring in single cells in suspension. By keeping the cells in suspension, magneto-rotation could help bridge the gap between petri dish and bloodstream environment. Even though flow is not present in our system, the magnitude of the shear stress acting on the cell while rotating, is of the same order of magnitude as that in the bloodstream (20 to 40 dynes/cm^2). It has to be noted though that shear stress in the bloodstream is not uniformly distributed in space and in time (due to heart pulses). Instead of a moving environment, the cell itself performs a relative motion, the advantage being that the cell stays highly localized, without the need to be attached or constrained, which would be the case if we wanted to track single cells in a flowing stream. In addition, it has been shown that gene expression and cell signaling are significantly different for cells grown on a 2D petri dish compared to those grown in 3D [6,7]. Once plated, clinical samples might also express a different phenotype than their suspended counterparts, a phenomenon that could be studied using Cell Magneto-Rotation. In the meantime, traditional assays, such as flow cytometry and MTT assays, have been relying on mass numbers and plated cells.

Hence, we see their potential inadequacy when it comes to toxicity assays of CTCs: the impossibility to perform these assays on a reduced number of cells (a few dozens), and, more seriously, the risk of being irrelevant because of the difference in gene expressions, if not mutations, that occur if these circulating cells are plated. Applied to the rare CTCs, where every single cell could be a repopulating one, the one that we want to target, and thus one cannot afford to lose a significant amount of cells at every time point of the monitoring. Another important feature that these cells exhibit is dormancy. They can stop growing for prolonged periods of time [36]. What is the point of plating these cells if they are not supposed to grow? And if they do grow, what conclusions can be drawn from assays made on cells that have been denatured in the process? As much as it is vital to « eradicate all intratumoral subclones », as stated by Notta et al. [12], the next anti-cancer therapies will also have to eradicate all the subclones in the circulating cell population so as to prevent metastasis: such drug sensitivity tests could be performed using the CM method, as a complementary technique. In addition, the magneto-rotation test can be used coupled with a camera instead of a laser beam (or an LED), and thus does not necessitate a complex optical setup besides the microscope. Since a dormant cell is alive but does not grow, its rotation rate should not vary under non-toxic conditions, even after a period of time corresponding to a full cell cycle. Thus our approach could allow us to discriminate dormant cells from the general population.

We chose to work on HeLa cells because of their ability not only to survive but also to grow in suspension. As such, for this proof of concept, they served as a model for CTCs. Many questions are worth being asked then. Do circulating or disseminating tumor cells divide in the bloodstream? If yes, which cells tend to divide? Do they mutate? Compared to plated cells, how do they respond to drugs? This Cell Magneto-Rotation method could potentially offer researchers a valuable tool to answer such questions.

We also observed the formation of filopodia in healthy cells during rotation. Filopodia are spikes that are responsible for cell motility, migration and fixation to a substrate. However, because filopodia are oriented toward the outside of the cell, these morphology changes were

43

sufficient to affect the rotation rate. It is not clear yet whether filopodia formation is a result of rotation or a process that would occur anyways to cells in suspension. However, filopodia, or other protrusions, might not be formed in cells while circulating, but it is very likely that they appear when these circulating cells try to attach to the endothelium in order to reach for tissues and/or secondary tumors [37]. As such, if magneto-rotation actually permits the formation of protrusions, it could add a tool to the research effort on cell adhesion. Again, to our knowledge, no other method could do that for cells in suspension, which is critical when it comes to cells invading new tissues from the bloodstream.

Conclusion

In summary, we have described a single live cell analysis system that can monitor cell morphology through the related effective volume changes, *in suspension*; it does not affect cell viability. Specifically, we have demonstrated the ability to use cells as rotating magnetic microplatforms, through the uptake of functionalized magnetic nanoparticles, and the ability to control and measure their rotation under near real-time conditions. Cell death, and the dying process can simply be monitored through changes in the cell's rotational period. This lends itself to rapid drug sensitivity testing on cancer cells, with no need for cell culturing. Potentially, it could be used for tests on the rare and fleeting (due to differentiation) cancer stem cells. While circulating, the dormancy of these cells could also be evaluated this way, via the observed stability of their rotation rate. The methodology used here is very general, and can be used with various cell types (tumor, stem cells, red blood cells), and in various media. Also, this micro-system could be operated on a range of supports (cell imaging plate, agarose layer, inverted droplet, PDMS micro-channel), and we anticipate that this magneto-rotation method can also be applied to the rotation of other systems, such as cell clusters or spheroids (work in progress). The CM method here described could be adapted to various biotechnology applications, e.g. drug discovery or testing, and to growth assays, all performed in a three dimensional environment. We also envision CM integration into an *in vivo* magnetic enrichment process, followed by *ex vivo*

monitoring, for tailor-made therapies. Ongoing work is focused on live cell analysis, on cell growth and on studies on clinical samples.

Materials and Methods

Functionalization of magnetic nanoparticles

To magnetically label HeLa cells, 30nm amine coated magnetic nanoparticles (Ocean Nanotech®) were functionalized using poly-L-lysine (PLL, Sigma Aldrich GmBH), a transfection agent that improves the internalization in cells [30]. A solution of 200ug/ml of nanoparticles in Dublecco's Modified Eagle Media DMEM was mixed with 10uL of PLL, and rotated end-over-end in a vial at room temperature for 1 hour. The particles solution was then filtered using a 0.2um filter (Whatman® Nylon Filter Media) to remove any biological agents that could contaminate the sample. The filtered solution was immediately used.

Cell culture and labeling

HeLa 229 cells (American Type Culture Collection) were cultured for four days in Dublecco's Modified Eagle Media (DMEM 11995, Invitrogen™), 10% Foetal Bovine Serum (FBS), 1% Penicillin–Streptomycin–*Glutamine (*PSG) and 25ug/ml (prior to filtration) of functionalized magnetic nanoparticles (Ocean Nanotech).   The growth medium was removed, and cells were washed once, using PBS, before adding Cell Detachment Buffer (Gibco™). This enzyme free buffer does not affect surface proteins during cell removal from the dish, and allows the nanoparticles which could have attached the surface of the cell to be retained. After 30 min of incubation in the detachment buffer, cells were washed with DMEM, and centrifugated (for the preparation of fixated cells, this step was replaced by magnetic separation in order to keep the cells from forming clusters). Cells were resuspended in fresh media.

Nanoparticles preparation for fluorescent imaging.

3 ml of magnetic nanoparticles (tagged with poly-L-lysine) at a concentration of 200ug/ml in DMEM were mixed with 3mg of HPTS fluorescent dyes. The mixture was vortexed and then put under end over end rotation for one hour before being centrifuged at 9000rpm for ten minutes in Amicon® Ultra centrigugal filters Ultracel® 3k. The particles tagged with the fluorescent dyes were then resuspended in DMEM at the initial concentration of 200ug/ml.

Setup for rotation measurements

Before rotation, 300uL of the cell solution was introduced into a Live Cell Array™ plate (NUNC®), with 100um wells. Cells were then pulled to the bottom of the plate using a permanent magnet. Once cells were pulled down to the wells, the plate was placed inside the coils, with the wells in the center.

Coils description

Custom Helmholtz coils (see **figure 2.1d**) were integrated on the platform of an Olympus® BX50WI microscope. Each pair of coils produced a field parallel to the imaging plane and was plugged into an amplifier (amplification factor during rotation was set to 1), which, in turn, was plugged to two function generators with a 90 degree phase shift (Agilent Technologies Arbitrary Waveform Generator 33220A, 20 MHz function). Both power supplies were set to provide a sine wave function, with amplitude of 3V. The phase shift was controlled with an oscilloscope (Agilent Technologies, DSO5012A). Finally, the magnitude of the magnetic field was measured using a magnetic probe placed in the center of the magnetic coils (3 Axis Magnetic Field Transducer, C-H3A-2m_E3D-2.5kHz-1%-2T, Sensitivity 5[V/T], SENIS GmbH).

Optical setup

The laser used was an unstabilized HeNe laser (Spectra-Physics® 136/P), with a wavelength of 632nm (1.8mW/cm^2). Data were acquired using a Labjack UE9 data acquisition

device, receiving the diffraction signals from a non amplified photosensor. The data were recorded analyzed on a computer (DELL©, *Intel® Core™2 Duo CPU E6550 at 2.33Ghz, 1.98GB RAM, Microsoft Windows® XP Professional Version 2002 SP3) using customized software (LabView).*

The modulated signal is then treated using an algorithm that recognizes the peak to peak variations. From there, the rotation period is extracted averaging the peak to peak period over a defined time window that moves over time. For instance, the time window over which we average the period could be 60s, and it would be translated by twenty seconds to calculate the next point.

The longest cell rotation period used is on the order of one minute, which is the case when the cell's blebbing created a large cell and a high effective volume. At the beginning of the experiments, the rotation period was usually comprised between 1s and 15s. To analyze the signal, we measured and average the rotation period over a moving time window of at least 10 periods. In the early stages, we needed a 30s time window, and when the rotation rate becomes very low (30s), we used a time window of around 3 min (even though at this point, a statistical averaging of the rotation period is not relevant since the length of the period reduces the error made on the measure).

Image acquisition was made through a Digital Camera (Mightex Monochrome Camera MCE-B013-US, 1.3 MegaPixels), and images were recorded with the Mightex acquisition software (v1.1.0, 1280x1024, Exposure Time 35ms). Image capture was realized via an external trigger, programmed on LabView.

Laser wavelength, power

The laser power was measured using a power-meter (Coherent Calibration Tag, MIL-STD-45662-A). Before reaching the microscope's mirror (namely after its transmission through the

47

condenser), the power measured was of 1.45mW. On the microscope platform, the power was between 125uW +/-2uW (1.8mW/cm^2).

Inductively Coupled Plasma

To measure the iron content of the cells, we used ICP measurements. After the standard incubation protocol, cells were washed three times in ice-cold PBS, detached, and counted. Afterward, cells were digested for three hours in 70% nitric acid in a water bath at 90°C and the iron content was then measured using an Inductively Coupled Plasma (ICP). For these measurements, the magnetic nanoparticles were not filtered before incubation, so that the exact density in solution was known. To make sure the MNP concentration was the same with unfiltered particles than with filtered ones, we measured the iron content of the MNP solution before and after filtration, resulting in a loss of 50% of the particles (data not shown).

Clonogenic Assay

HeLa cells were cultured in DMEM complemented with 10% FBS and 1% PSG until confluent in an incubator at 37oC, 5% CO2 and 100% humidity control. Cells were then washed in PBS, trypsinized and resuspended in fresh culture media. From this same batch, one control petri dish and one petri dish with nanoparticles were prepared (cf. Methods section in manuscript). After 4 days of incubation, cells were washed with PBS (three times in ice-cold PBS for labeled cells), resuspended in fresh media and counted. A 6-well plate was coated with a mix of 1.2% agarose in transparent DMEM, and let sit at room temperature for 30 min. Afterward, we placed 200 control cells in two different wells, and placed 10,000 magnetically labeled cells on an agarose coated plate, under a rotating magnetic field, for 24 hours. Afterward, in two different agarose coated wells, we placed 200 of the labeled cells that were rotated. After three weeks, we carefully pipetted the formed spheroids, and counted the formed spheroids (excluding small cellular aggregates).

References

1. Elbez R, McNaughton BH, Patel L, Pienta KJ, Kopelman R (2011) Nanoparticle Induced Cell Magneto-Rotation: Monitoring Morphology, Stress and Drug Sensitivity of a Suspended Single Cancer Cell. PLoS ONE 6: e28475. doi:10.1371/journal.pone.0028475.

2. Longo D, Hasty J (2006) Dynamics of single-cell gene expression. Mol Syst Biol 2. Available: http://dx.doi.org/10.1038/msb4100110. Accessed 15 July 2010.

3. Krutzik PO, Crane JM, Clutter MR, Nolan GP (2008) High-content single-cell drug screening with phosphospecific flow cytometry. Nat Chem Biol 4: 132–142. doi:10.1038/nchembio.2007.59.

4. Valet G (2006) Cytomics as a new potential for drug discovery. Drug Discov Today 11: 785–791. doi:10.1016/j.drudis.2006.07.003.

5. Snijder B, Sacher R, Ramo P, Damm E-M, Liberali P, et al. (2009) Population context determines cell-to-cell variability in endocytosis and virus infection. Nature 461: 520–523. doi:10.1038/nature08282.

6. Justice BA, Badr NA, Felder RA (2009) 3D cell culture opens new dimensions in cell-based assays. Drug Discov Today 14: 102–107.

7. Abbott A (2003) Cell culture: Biology's new dimension. Nature 424: 870–872. doi:10.1038/424870a.

8. Molina JR, Raphael RM, Ozawa MG, Stark DJ, Levin CS, et al. (n.d.) Three-dimensional tissue culture based on magnetic cell levitation. Nat Nanotechnol 5: 291–296.

9. Hsieh HB, Marrinucci D, Bethel K, Curry DN, Humphrey M, et al. (2006) High speed detection of circulating tumor cells. Biosens Bioelectron 21: 1893–1899. doi:10.1016/j.bios.2005.12.024.

10. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature 440: 358–362. doi:10.1038/nature04599.

11. Wicha MS, Liu S, Dontu G (2006) Cancer stem cells: an old idea—a paradigm shift. Cancer Res 66: 1883.

12. Notta F, Mullighan CG, Wang JCY, Poeppl A, Doulatov S, et al. (2011) Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. Nature 469: 362–367. doi:10.1038/nature09733.

13. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, et al. (2010) Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. Dev Cell 18: 675–685. doi:10.1016/j.devcel.2010.02.012.

14. Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, et al. (2008) Dynamic Proteomics of Individual Cancer Cells in Response to a Drug. Science 322: 1511–1516. doi:10.1126/science.1160165.

15. Tay S, Hughey JJ, Lee TK, Lipniacki T, Quake SR, et al. (2010) Single-cell NF-[kgr]B dynamics reveal digital activation and analogue information processing. Nature 466: 267–271. doi:10.1038/nature09145.

16. Lahav G, Rosenfeld N, Sigal A, Geva-Zatorsky N, Levine AJ, et al. (2004) Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat Genet 36: 147–150. doi:10.1038/ng1293.

17. Galanzha EI, Shashkov EV, Kelly T, Kim J-W, Yang L, et al. (2009) In vivo magnetic enrichment and multiplex photoacoustic detection of circulating tumour cells. Nat Nano 4: 855–860. doi:10.1038/nnano.2009.333.

18. Bertorelle F, Wilhelm C, Roger J, Gazeau F, Ménager C, et al. (2006) Fluorescence-Modified Superparamagnetic Nanoparticles: Intracellular Uptake and Use in Cellular Imaging. Langmuir 22: 5385–5391. doi:10.1021/la052710u.

19. Yu MK, Jeong YY, Park J, Park S, Kim JW, et al. (2008) Drug-Loaded Superparamagnetic Iron Oxide Nanoparticles for Combined Cancer Imaging and Therapy In Vivo*. Angew Chem Int Ed 47: 5362–5365.

20. Ho VHB, Müller KH, Darton NJ, Darling DC, Farzaneh F, et al. (2009) Simple magnetic cell patterning using streptavidin paramagnetic particles. Exp Biol Med 234: 332–341.

21. Wilhelm C, Gazeau F, Bacri J-C (2003) Rotational magnetic endosome microrheology: Viscoelastic architecture inside living cells. Phys Rev E 67: 061908. doi:10.1103/PhysRevE.67.061908.

22. Fantner GE, Barbero RJ, Gray DS, Belcher AM (2010) Kinetics of antimicrobial peptide activity measured on individual bacterial cells using high-speed atomic force microscopy. Nat Nano 5: 280–285. doi:10.1038/nnano.2010.29.

23. Zhang H, Liu K-K (2008) Optical tweezers for single cells. J R Soc Interface R Soc 5: 671–690. doi:10.1098/rsif.2008.0052.

24. Ramser K, Hanstorp D (2010) Optical manipulation for single-cell studies. J Biophotonics 3: 187–206.

25. Gfeller KY, Nugaeva N, Hegner M (2005) Rapid Biosensor for Detection of Antibiotic-Selective Growth of Escherichia coli. Appl Environ Microbiol 71: 2626–2631. doi:10.1128/AEM.71.5.2626-2631.2005.

26. McManus M, Fischbarg J, Sun A, Hebert S, Strange K (1993) Laser light-scattering system for studying cell volume regulation and membrane transport processes. Am J Physiol- Cell Physiol 265: C562.

27. McNaughton BH, Kehbein KA, Anker JN, Kopelman R, others (2006) Sudden breakdown in linear response of a rotationally driven magnetic microparticle and application to physical and chemical microsensing. J Phys Chem B 110: 18958.

28. Janssen XJA, Schellekens AJ, van Ommering K, van Ijzendoorn LJ, Prins MWJ (2009) Controlled torque on superparamagnetic beads for functional biosensors. Biosens Bioelectron 24: 1937–1941.

29. McNaughton BH, Agayan RR, Wang JX, Kopelman R (2007) Physiochemical microparticle sensors based on nonlinear magnetic oscillations. Sens Actuators B Chem 121: 330–340.

30. Goldberg M, Choi D (1993) Combined oxygen and glucose deprivation in cortical cell culture: calcium-dependent and calcium-independent mechanisms of neuronal injury. J Neurosci 13: 3510–3524.

31. Arbab AS, Bashaw LA, Miller BR, Jordan EK, Lewis BK, et al. (2003) Characterization of Biophysical and Metabolic Properties of Cells Labeled with Superparamagnetic Iron Oxide Nanoparticles and Transfection Agent for Cellular MR Imaging1. Radiology 229: 838–846. doi:10.1148/radiol.2293021215.

32. Villanueva A, de la Presa P, Alonso JM, Rueda T, Martínez A, et al. (2010) Hyperthermia HeLa Cell Treatment with Silica-Coated Manganese Oxide Nanoparticles. J Phys Chem C 114: 1976–1981. doi:10.1021/jp907046f.

33. Wilhelm C, Gazeau F (2009) Magnetic nanoparticles: Internal probes and heaters within living cells. J Magn Magn Mater 321: 671–674. doi:10.1016/j.jmmm.2008.11.022.

34. Wilhelm C, Gazeau F (2008) Universal cell labelling with anionic magnetic nanoparticles. Biomaterials 29: 3161–3174. doi:10.1016/j.biomaterials.2008.04.016.

35. Markx GH, Davey CL (1999) The dielectric properties of biological cells at radiofrequencies: applications in biotechnology. Enzyme Microb Technol 25: 161–171.

36. Aguirre-Ghiso JA (2007) Models, mechanisms and clinical evidence for cancer dormancy. Nat Rev Cancer 7: 834–846. doi:10.1038/nrc2256.

37. Machesky LM, Li A (2010) Fascin. Commun Integr Biol 3: 263–270.

Chapter 3 :
Morphology Heterogeneity in Metastatic Breast Cancer Cells by Dynamic Histology using Multiplexed CMR (Cell Magneto-Rotation) and Graph Analysis.

Abstract

Cancer metastasis is driven by a small minority of cells that have the potential to migrate and populate new tissues. Metastatic cells exhibit their higher plasticity with the use of protrusions, blebs or amoeboid movements during migration. Can we quantify metastatic potential? Here we describe a new *biomarker free, dynamic histology method, which we designed* for classifying and quantifying the inherent plasticity of cancer cells, based on the way they change their shapes when suspended in tight wells, towards the determination of their metastatic potential. MBA-MD-231 breast cancer cells were stained to be magnetic by endosomic uptake of magnetic nanoparticles. Trapped individually in 3-dimensional wells, cells are then observed at the single cell level, while being slowly rotated and thus freely suspended, using a low intensity magnetic field. These cells tend to adopt different shapes and shape evolutions, and thus alternate between different morphological phenotypes over time. Using object recognition and machine learning algorithms, we were able to measure the shape dynamics of each cell over time, continuously, and to successfully map the inherent broad heterogeneity of the morphological phenotypes found in a given cancer cell population. Cells were then clustered into groups sharing similar dynamics in their morphological behavior, and *graph theory analysis* revealed distinct clusters of highly *vs.* weakly plastic cells. We believe that extending this method to animal samples, together with a gene expression analysis, will enable, in a given cell population, a reliable identification of cells having a distinct metastatic potential. Such quantification and visualization of malignancy in a *relationship graph* could be used for an early assessment of the efficacy of therapy, during personalized treatment.

Significance

A most critical challenge in treating cancer is to identify the most aggressive and proliferative cells. They are characterized by a superior ability to change their shape, an ability that helps them migrate to new tissues, resulting in metastasis. We thus developed a *dynamic histology* method that tracks and monitors each cancer cell's morphology over time, on the individual cell level. By trapping single cells in 3-dimensional wells and magnetically rotating them, each cell's shape is continuously extracted by object recognition algorithms. Cells are then classified, based on the similarity of their behavior, and the cells with higher morphological activity are identified, using relationships graph theory. When highly parallelized, as described here, this method is likely to provide both faster and more quantitative diagnostic capabilities and assessments for personalized cancer therapies.

Introduction

The overwhelming part of lethality in cancer finds its source in metastasis. Metastasis is the process during which cells from the original cancer tumor leave their environment and disseminate so as to colonize new tissues. When the cancer has significantly metastasized, the prognosis is usually very pessimistic. Presented here is a new *dynamic histology* method, enabling a new classification of cancer cells that correlates with their *metastatic potential*, i.e., their probability of causing metastasis. Cell morphology is strongly correlated to a cancer cell's metastatic potential[1–5]. Therefore, using morphology to classify cells and thereby identify the cells with the highest malignancy (metastatic potential) could help circumventing the limitations of solely relying on surface biomarkers, which may alter during therapy[6]. Even though there are clear morphological distinctions among cells when plated, such plating of the cells might change their phenotypes and alter the quality of diagnostics[7,8]. Here, using the Cell Magneto-Rotation (CMR) method, on breast cancer MDA-MB-231 cells, we show that we can monitor, identify the shape evolution of individual cells over time, in a multiplexed fashion, and all of this while they are floating freely, just magnetically rotated, with no cell attachment inside individual wells. Based on the measurements of parameters characterizing individual cell morphology (>120 parameters) and its time evolution, the cancer cells' heterogeneity is visualized in a relations

54

graph, and single cells are then classified according to their metastatic potential. In addition to separating aggressive from benign morphological phenotypes, the method also identifies transitional cell phenotypes that serve as bridges towards the development of highly aggressive ones. Such information is expected to have both theoretical and clinical impact. We note that while this new method may be ideal for the analysis of captured CTCs (circulating tumor cells), it could also be used on cells from biopsied tissues, thereby complimenting, and potentially supplementing, traditional histology.

Over the past decades, the wide-ranging heterogeneity among a previously supposed to be homogeneous cell population has been recognized and intensely studied. For instance, it is now well understood that various problems in medicine, such as in cancer metastasis and resistance, or even in immunology[9–15], originate from cell heterogeneity. Tissues are no longer seen as assemblies of identical cells, sharing the exact same genome, phenotype and behavior, but as hierarchical societies where each cell plays its own role, as is, for instance, hypothesized in the cancer stem cells paradigm[2,16–18]. The new global understanding has been helped by numerous new technologies, which allow one to analyze cell populations at the single cell level. However, one of the major bottlenecks left is the ability to link the single cell analysis data with the actual behavior of this single cell as observed; specifically, the ability to track, over time, in a dynamic population of cells, every given single cell, retrieve the information collected, and associate it to the particular cell, and to do this for each cell in the entire population in a sample. Indeed, even though single cell genomics has seen a lot of progress, it is still highly challenging to analyze the genome or the RNA of one given single cell, since part of the DNA, or RNA, might get lost, or the cell may be damaged during manipulation. In addition, such measurements require a lot of computational power. By the time one gets the results, cells are not retrievable, and one cannot keep track of dynamic changes happening in the cell.

Cell morphology has been strongly correlated with the cancer cell's metastatic potential[1,2,19]. For instance, there is very strong evidence in animal models that the amoeboid stage significantly increases the metastatic potential of tumors[4]. Therefore, using morphology

to identify the cells with the highest malignancy is not only label-free but could help bypass the limitations of solely relying on surface biomarkers and on static, one-time snapshots of the morphology, which may get altered during therapy[6,20]. Even though there are clear morphological differences among plated cells, plating is not always informative. Here, using the Cell Magneto-Rotation (CM) method on breast cancer MDA-MB-231 cells, we show that we can monitor, identify and classify the shape evolution of individual cells over time, in a multiplexed fashion, and all of this while these cells are floating freely, just magnetically rotated, without surface attachment to the wells.

The development in recent years of high-throughput analysis methods[21,22] has resulted in an exponential increase in the amount of retrievable data[23]. Traditional statistical analysis methods are not optimal for treating large amounts of data, for the following reasons. On one hand, the data collected from even single experiments can be overwhelming for regular computers, and parallel computing is often needed. On the other hand, given the change of scale, not only in the amount of samples but also in the amount of possible measurements, it is now possible to extract relationship patterns, in particular with the use of graph theory. The very same principles used by web sites to capture similarities among users and their personalized content can be applied to many fields in biology, even to predictions of epidemics, using graphs and social networks[24–27]. For instance, machine learning models have been shown to be good predictors of drug efficacy[28]. These techniques, often referred to as "Big Data", could have an important impact if used intelligibly for single cell analysis. Indeed, theoretically, all the cells in a tumor sample could be identified, analyzed, and presented in a *relations graph* that would provide a key for understanding and monitoring tumor development. Notably, with traditional methods for single cell characterization/classification, based on genetic/biochemical analysis, there is still a void when it comes to medical applications, because such single cell analysis is still too much of a challenge. Notably, the fraction of cells actually individually analyzed is low, and reliable cell identification and tracking is still to be developed.

Compared to the traditional biological/biochemical methods of monitoring single cells, here, in the CMR method, the cells are kept floating in suspension, which is a step closer to the situation cells are facing when travelling within the bloodstream. This is particularly relevant to circulating tumor cells (CTCs). Present techniques that rely on suspended cells, such as Flow Cytometry, can provide single cell analysis, at a high throughput as well, but are unable to keep track of where the cells went, so as to identify them, which makes it less than optimal for use in a dynamic system. Here, because cells are located in specific wells, and because we identify each well with a specific tag, we can keep track of each cell over a long period of time. Our CM method, presented below, allows us to do single cell analysis in a dynamic way, harmlessly and with high throughput.

Regarding CTCs or cells dissociated from primary tumors, what happens to cancer cells when facing a brutal change in environment, from being attached to a matrix to being suspended in a flow, is relatively unknown. The way a cell's morphology changes when "thrown" into this new type of environment could potentially help us understand some of the processes associated with metastasis, and in particular, what happens when cells enter the bloodstream and have to face hydrodynamic stress. For example, it is not clear what fraction of the CTCs captured from the bloodstream will eventually metastasize. First, their origin is not clear, given that tumor cells shed millions of cells every day, and the ones captured using current systems might not be viable by the time they are recovered[29]. Indeed, as explained by Bednarz-Knoll and colleagues[29], one could argue that most of the cells being captured are the ones that spent too much time in the bloodstream without anchoring to the endothelium, namely the ones that did not make it into a new tissue environment, thus inducing a reverted *survivor bias*. In turn, the CTCs remaining in the bloodstream could be either damaged and/or unfit to colonize, which might explain why it is so difficult to culture captured CTCs *in vitro*[30]. In addition, with the exception of morphology based separation techniques of CTCs, such as ScreenCell®[31] or ISET®[32], other techniques rely on surface biomarkers, most notably EpCAM, that may not represent the full scope of the diversity found in the CTCs population[29]. The subpopulation of CTCs thus captured may therefore not be representative of the CTC population as a whole. And, finally, the

mechanisms and mechanical cues involved during anchoring of CTCs to the endothelium, prior to invading a new tissue, are still mostly unknown. To our knowledge, so far there is no *in vitro, or ex vivo,* tool that could be used to study such a course of action.

We believe that the ability to classify and track cells based on their morphological changes, over time, could be exploited for finding the precursors of the metastatic process, by directly looking into the behavior dynamics of the tumor cells. Our method could then be used as a diagnostic tool, and improve the selection process and capture of *meaningful* CTCs, by refining the criteria used in the existing capture technologies, i.e. by modifying the targeted biomarkers or the way to capture the cells in the bloodstream.

Specifically, based on the CMR method, we present here a new *dynamic histology* technique that enables one to track, monitor and *classify* single cells in a multiplexed fashion, by dynamically measuring geometrical features of cells, over an extended period of time. Using, mathematical, unsupervised clustering techniques, we were able to cluster cells into distinct morphological and behavioral phenotypes.

Results

Human Breast Cancer (MDA-MB-231) of mesenchymal phenotype cells are magnetized as described by Elbez et al.[33]. They are then introduced into a microfluidic chip and trapped, where the surface treatment significantly prevents cell attachment to the surface, as described in the **Materials and Methods Section**[34]. The cells are then ready to be imaged (**Figure 3.1**). Trolox is used here as an oxygen radical scavenger, in order to limit photodamage[35]. The media is then supplemented with 25 mM HEPES, and cells are then placed inside a homemade stage incubator, with temperature and humidity control. Exposure of the cells is controlled by a programmable shutter, and the total exposure of the cells is only 900 ms per minute. Under these conditions, cells survive for more than 24 hours (without exposure[33]), and are not harmed by exposure during imaging (see **Figure 3.2**). More importantly, we have shown previously the ability of CMR to detect apoptotic behavior, by comparing the evolution of the rotation rate of a cell with the presence (or not) of cytoplasmic Propidium Iodide (PI) fluorescence, in a cell/death

assay. Basically, when the cell's constitution deteriorates, its rotation rate sharply decreases, until rotation fully stops. We detected that PI fluorescence, and thus the beginning of cell necrosis, correlated well with a significant increase in the cell rotation period[33]. Therefore, as long as cells produce GFP and rotate in a relatively stable fashion, we can be confident that the cells are still healthy. Notably, we did not notice any qualitative difference in behavior by imaging cells for just 90 minutes. The use of this shorter experiment time has the advantages of speeding up the diagnostic capabilities of our system, and limiting cells' exposure to the fluorescence excitation light.
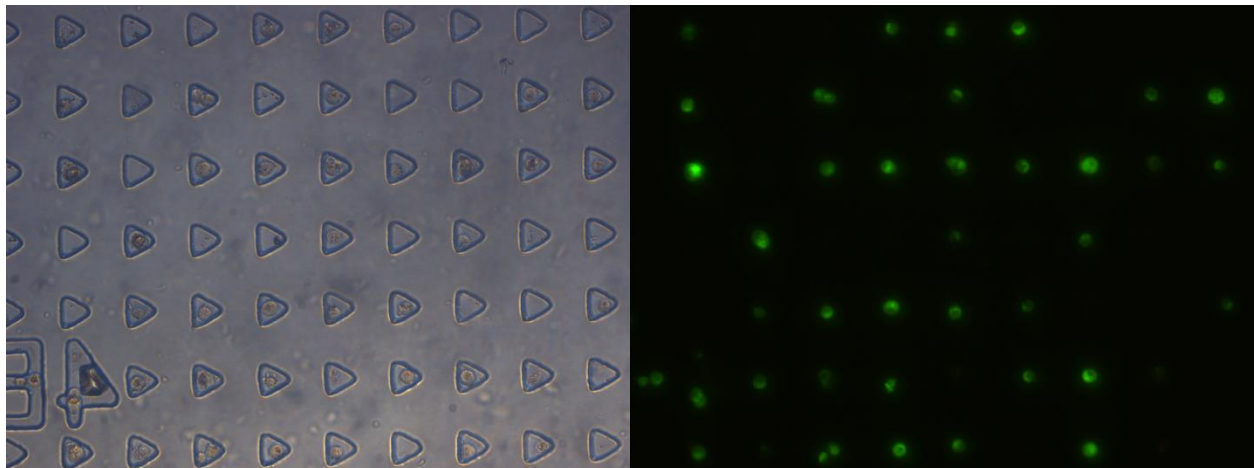
Figure 3.1 a) Brightfield image of the trapping microfluidic chamber with trapped cells and **b)** Fluorescent imaging caption of the same area. Both captions were taken with a 10X magnification. Each triangular trap has a side length of 40μm.

To assess the ability of our method to successfully monitor cells while preserving viability, we performed a live/dead cell assay using Propidium Iodide. To do so, cells were magnetically rotated as described above, and exposed to fluorescence exciting light; viability was assessed at time 0 min and time 90 min (see **Figure 3.2**).
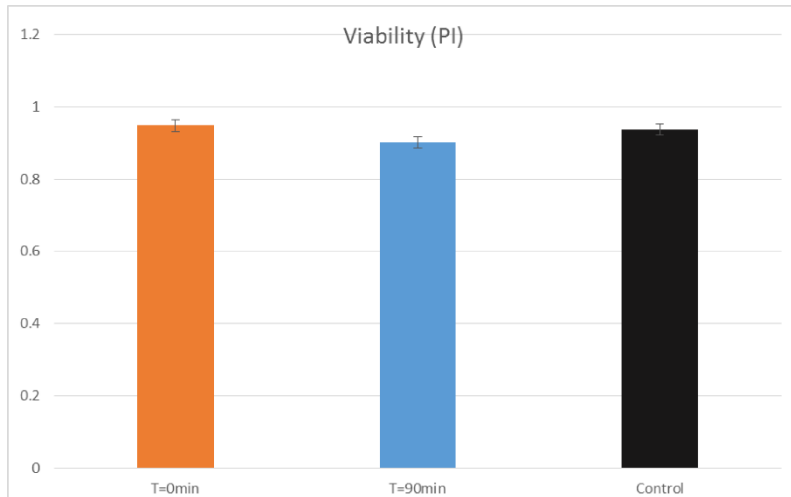
Figure 3.2: Live/Dead cell assay of rotating cells using PI, after 90mins exposure to illumination on the microscope stage.

To extract morphological features from the cell images, we used the open-source software CellProfiler[36]. For each cell, the area around the well where it is located is cut out from the original picture, and the corresponding series of images is analyzed. The time lapse images of the cells are run through a CellProfiler pipeline (series of procedures), where the threshold is adjusted so that the delineation fits best the shape of the cell. Because of the heterogeneity in the expression of the fluorescent protein by the cells, we group cells by the intensity threshold set to delineate the cells. This step has no effect on the clustering that we apply later on. Once the outline of the cell is detected, the software measures several geometrical features of the cell, such as its *area*, its *perimeter*, *eccentricity, roughness*, etc., as well as its *texture* or *angular repartition*. In total, more than 80 parameters are measured for each cell (see **Appendix A**), at every time point. As shown in **Figure 3.3**, the cell outline is detected rather well for features on the scale of a few micrometers. For instance, even though uneven contrast inside the cell can add to the difficulty of delineating the cell outline, we successfully extracted the shape of different types of cells (*round, amoeboid, blebbing, oblong and protrusive*), using the same pipeline and the same intensity threshold. These phenotypes were chosen because they are associated with different motility and migratory behaviors of the cell. For instance, protrusions are used by cancer cells to progress through the extra-cellular matrix (ECM) by attaching to the collagen fibers and pulling themselves through (if needed, the cell can also digest

60

the collagen fibers to make way). If not possible, the cells can revert to an amoeboid or a blebbing phenotype and use fast shape changing abilities to squeeze in between the fibers. We also observed that cells could sometimes have a clearly oblong shape, and decided to include this class to the list of phenotypes.

Most importantly, we show here that we can track fast shape changing events (events on the order of one second), and this in a completely automated fashion. However, some parts of the cells, especially very thin protrusions, are hard to delineate, as can be seen in the sequential captions on **Figure 3.3**.
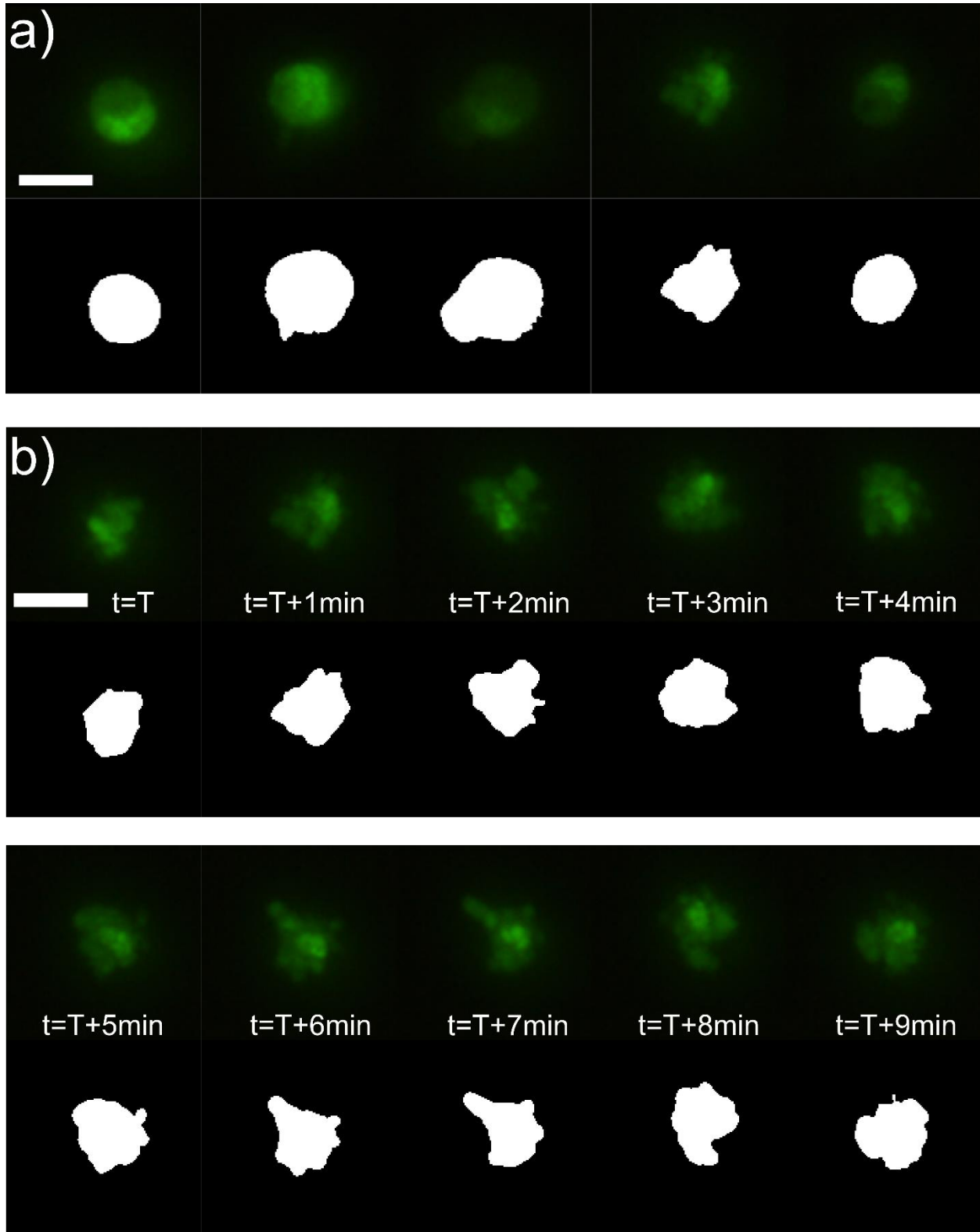
Figure 3.3: Image analysis and features detection using CellProfiler. **a)** Different phenotypes of cells are represented here (round, protrusive, amoeboid, blebbing and oblong). In green are the fluorescent images as taken by the CCD camera, and the bottom captions of each series show the

result after image detection by the CellProfiler software. **b)** (2 parts) shows a sequence of the same cell outline, every minute for ten minutes in a row. Scale bar is 20μm.

Next, once all the cells have been delineated and their shape tracked over time, we can start the classification by Machine Learning (ML) methods. Machine Learning (ML) is the process by which a computer is trained using data. In our case, we present the computer with images of cells and we classify these in various categories. Each image is linked to the data associated, i.e. the measurements made by image analysis stored in a database. For example, it is the same principle used by Optical Character Softwares (OCR). After enough samples have been presented to the computer, it has enough "experience" to construct a set of rules that will classify the cells without the help of the user, and we can let the computer score (i.e. put each image in a specific category) the rest of the image collection.

To do this, we used CellProfiler Analyst software[37,38] to classify the cells into five different morphological categories: *round, oblong, protrusive, amoeboid and blebbing*. To train the computer to recognize and classify these cell shapes, we use a training sample from the collected images, similar to the ones shown in **Figure 3.3**. Namely, a few hundred cell captions are classified. For each caption, we indicate which category it falls into, and once we have manually scored enough cells, we let the computer run Support Vector Machine (SVM) algorithms so as to establish classification rules. At this point, we can let the computer score the rest of the cell captions. The output is a table, where each entry is the caption of a specific cell, at a certain time point. A typical entry looks like:

$$\boldsymbol{S}_{ij} = [cell_i, time = j, s_{ij,1}, s_{ij,2}, s_{ij,3}, s_{ij,4}, s_{ij,5}]$$

where $s_{ij,k}$ represents the score of cell i at time j for the phenotype k, and only takes 0 (negative match) or 1 (positive match) as values. Here k can designate 'round', 'oblong', 'protrusive', 'amoeboid' or 'blebbing'.

Once all the images have been scored, for each cell, we count the number of times the cell has been classified into each category. This is done by averaging the S_{ij} over time (indexed by j). We get the fractions for each phenotype, represented by a fraction vector $\boldsymbol{f}_i$ of cell *i*:

$$\boldsymbol{f}_i = \left[r_{round}, r_{oblong}, r_{protrusive}, r_{amoeboid}, r_{blebbing}\right].$$

Also, by construction, we have the following relationship between the fractions:

$$\sum_{k \in \{phenotype\}} r_k = 1$$

The mathematical representation of the results for the whole experiment is a matrix with values between 0 and 1. Each column designates one of the phenotypes used to classify the cell, and each row represents a cell. Smaller values of fractions are represented by lighter colors, and larger values by darker colors.

The results are visualized in a *heatmap*, which shows how often a cell scores into a particular category. In this *heatmap*, each line represents a cell, and each colored unit (rectangle) represents the fraction of time the cell spends as round, protrusive, amoeboid, oblong or blebbing, respectively. Cells scoring as "*round*" almost half the time are the most prevalent cells in the population. However, we see that some cells almost always score in the same category, showing a very stable phenotype (dark red units in **Figure 3.4**). On the other hand, it is interesting to note that cells do not necessarily stick to one phenotype. This might be explained in several ways. First, the algorithm in itself and the rules established after ML are, by construction, not perfect. The same cell, while slightly changing its shape over time, can fall into different categories over time. For instance, a cell showing an *amoeboid* behavior will change its shape very quickly, and in ways that are not predictable. From the heatmap, we can see that some

groups of cells have very similar behavior. Can we find a way to measure this similarity, i.e., can we group the cells based on this measure and visualize the results?
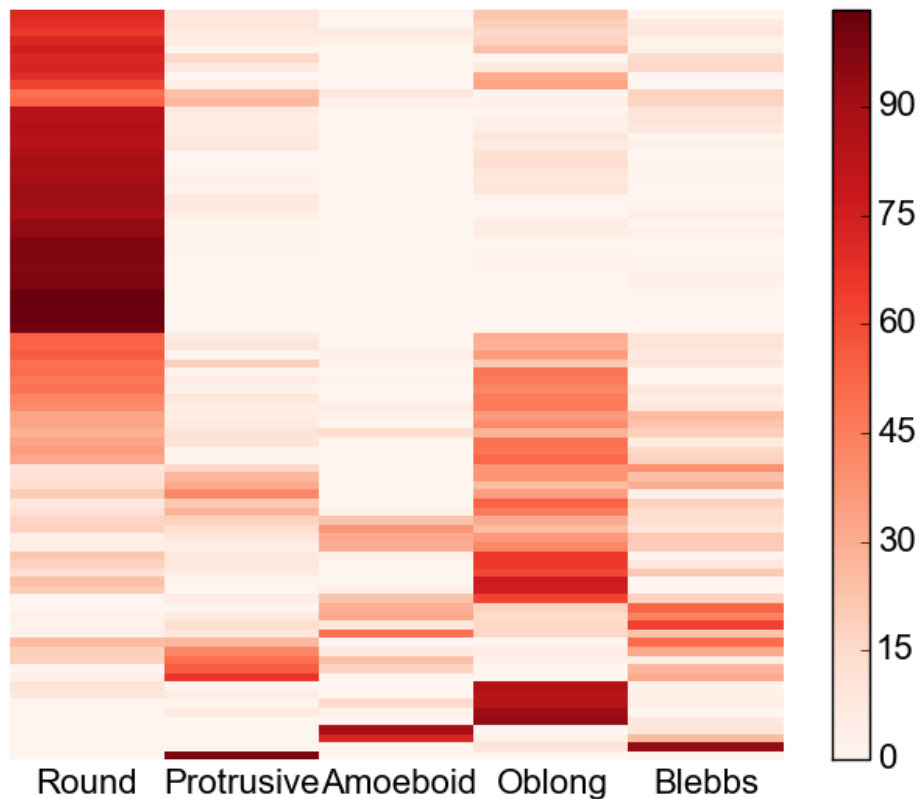
Figure 3.4: Cluster analysis using unsupervised clustering methods. This *heatmap* represents scoring fractions per cell per phenotype. Each column represents a different phenotype. Cell images (90 cells, around 7800 images) are scored according to the rules established by CPA, and the *heatmap* shows the amount of times a particular cell is scored into one of the phenotypes (round, oblong, amoeboid, protrusive and blebbing). The Color-bar shows the fraction's scale in percentage.

We can now look into pairwise relationships between cells. To do so, we represent the cell population with a graph. Unlike graphs representing mathematical functions, the graphs we use here are constructed with nodes connected with each other by edges, and each edge has a

weight representing the strength of the relationship between the two nodes (if the relationship exists).

To build a graph representing the similarities between the cells, several steps are required. We first model each fraction vector $\boldsymbol{f}_i$ as a point in a 5-dimensional Euclidian space. More accurately, since the sum of all the fractions is equal to one, these variables are bound, and the $\boldsymbol{f}_i$s are varying in a hyperplane of dimension 4, and the coordinates of each data point are the previously computed ratios $[r_{round}, r_{oblong}, r_{protrusive}, r_{amoeboid}, r_{blebbing}]$. By construction, all the data points are found inside the sphere of radius one centered at the origin (since a ratio is at most equal to one, in the case of cell always classified with the same phenotype all along the experiment). For every pair of cells, say cell i and cell j, we can associate a distance between the two cells, which is the Euclidian distance:

$$d_{ij} = \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2 = \sqrt{\sum_{k \in \{phenotype\}} (r_{i,k} - r_{j,k})^2}$$

All the distances are gathered to form a distance matrix D (with n the number of cells):

$$D = [d_{ij}]_{\substack{i=1..n \\ j=1..n}} = \begin{bmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{bmatrix}$$

Let us note that the diagonal of D is filled with 0's. Since the closer the points in space, the more similar the cells are, and we chose to model the similarity between two cells (i.e. how closely they behave) by a weight equal to the square of the inverse of the distance:

$$\forall (i,j) \in [\![1; n]\!]^2 / d_{ij} \neq 0, w_{ij} = \frac{1}{d_{ij}^2}$$

Once all the relationships between the nodes (cells) have been characterized and weighted, we can move on to the visualization part of the graph. By construction, the graph is undirected, meaning that there is no direction (hierarchy) associated with an edge. If two nodes, A and B, are connected to each other, an undirected relationship means that there is no hierarchy between A and B (such as precedence, regulation, etc…), and A and B are equally interacting. To position cells in space and get a 2-dimensional representation of the cells based on their similarity, we use a force directed graph drawing algorithm[39,40] (see **Figure 3.5**), and the open-source visualization software Gephi[41].

68
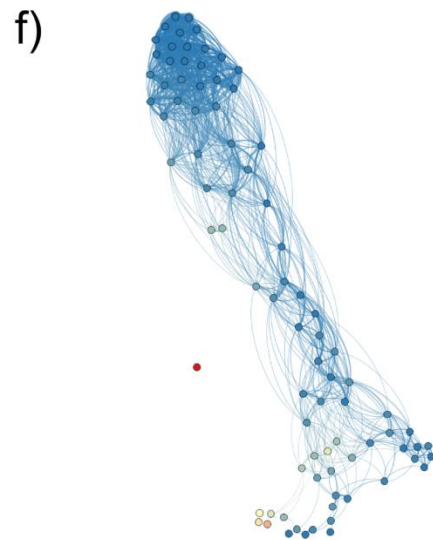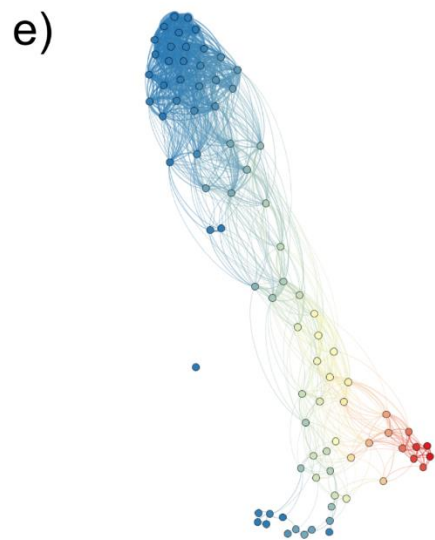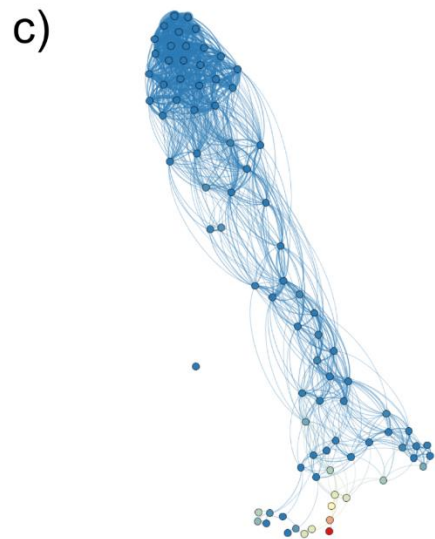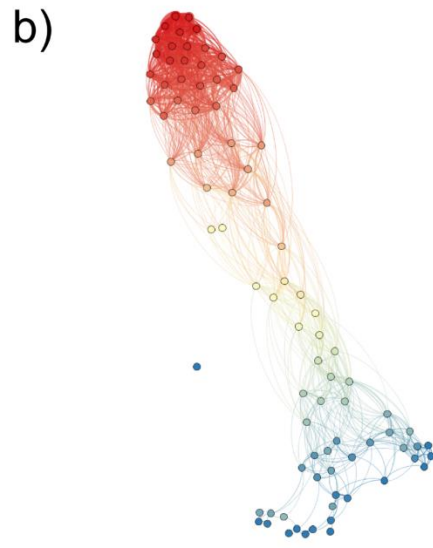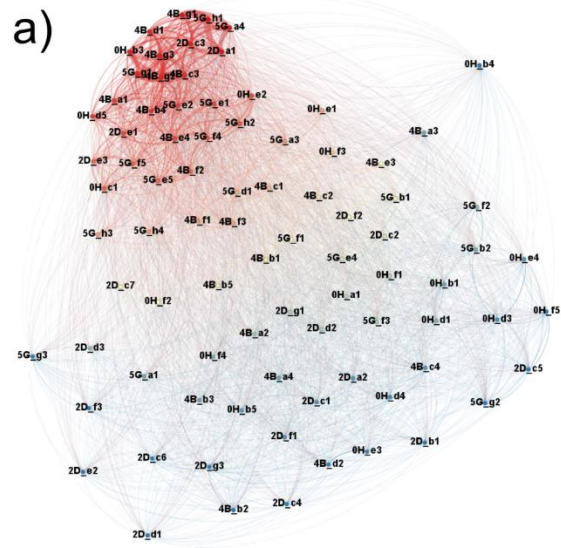
Figure 3.5: Graph representation of the cell clustering using the pairwise distances calculated with the dendrograms. Color tags correspond to clustering according to the cells' modularity class. **a)** Round fractions, with cell labels and unfiltered edges, **b)** round fractions with filtered edges, **c)** amoeboid fractions, **d)** blebbing fractions, **e)** oblong fractions and **f)** protrusive fractions. Red color designates a fraction closer to one, and blue a fraction closer to 0, with yellow at 0.5. Edge thickness is proportional to the weight of the edge.

Simply put, positioning the cells is equivalent to an *N*-body problem, and the similarity between cells is simulated by a set of attractive and repulsive forces, with the algorithm's loop ending when the energy of the n-body system is minimized. The nodes repel each other, while the edges produce a spring force that attracts the nodes toward each other. This model is called "spring-electric" model, because the attractive force is modeled on a spring force, and the repulsive one is modeled after an electric force model[42,43]. It is important here to remind the reader that these simulated interactions are not actual forces between the cells in the microfluidic chamber. However, once the positions of the nodes are set, the different figures represent the same graph but showing different properties of the nodes. As such, the spatial organization of the nodes is only weight dependent, and not property dependent, i.e., it does not matter if the cell is blue or red or belongs to this or that category; only the calculated weight matters towards drawing the edge between two nodes (see **Figure 3.5a**). When we color the nodes according to the fraction of *"round"* scores, we can clearly see a core of mostly round (red) cells, and further in the periphery, more abnormal cells. To help with the clarity of the visualization and the esthetics, we also applied a filter removing edges with a very low weight, which leaves only a quarter of the visible edges (**Figure 3.5b**). The shape is determined by an energy minimization process, and therefore the shape is not decided in advance. Extremely 'distant' cells have an edge with very low weight, so we filtered out the edges with low weights (below 3). This means that cells that are too different (more than a third of the time they belong to different categories), are cut from each other, i.e., their connection is cut.

This type of graph allows us to immediately grasp how each cell, in its behavior, relates to one another. As expected, the more round cells, which behave similarly, and supposedly are the least aggressive, are all clustered really close to each other (**Figure 3.5b**). They form the core

of the population, from which other cells distance themselves. Looking at the spread of the other phenotypes in the population, we notice continuous transitions between the different phenotypes. For instance, it is interesting to note that cells are getting to be *oblong* before becoming *amoeboid*, *blebbing* or having *protrusive* shapes. Regarding these three last phenotypes, we notice that blebbing ensures a transition between the two others. This graphic representation provides us with a very clear way for isolating the cells whose behavior deviates the most from the norm of the tumor cells. Again, since we are looking for cells with very specific patterns, these should stand out from the rest of the population (Following the Anna Karenina principle "happy families are all alike, but every unhappy family is unhappy in its own way"). Now that we have mapped the population and extracted the relationships between the cells, can we identify the most 'relevant' cells in the population?

More than identifying the phenotype, which gives us insight into what a specific cell is actually doing, we can also identify the cells that sit in between two phenotypes, and represent the transition. These cells are important because they can provide the explanation to the way cells shift from one behavior to the other. We therefore looked at the connections between the cells, and especially at the mathematical representation of the importance of the cell. To do so, we used graph centrality metrics such as the defined below *betweenness centrality* (see **Figure 3.6**) which serves as a measure of the centrality of a node in a graph. Simply put, they help identify the nodes that connect the other nodes together. An important aspect of the relationship between two distant nodes (not directly connected) is the shortest path between these two nodes, and the length of this shortest path. The *betweenness centrality* of a node is the number of shortest paths between two other nodes that pass through it. We also identify groups of cells that are very isolated from the rest of the population, because their *betweenness centrality* is very low. When matched with the phenotypes, they correspond to *protrusive, blebbing* and *amoeboid* behaviors. However, more importantly, we successfully identified the cells that connect these extreme behaviors with the rest of the normal population (see red and light red colored cells in **Figure 3.6a**).
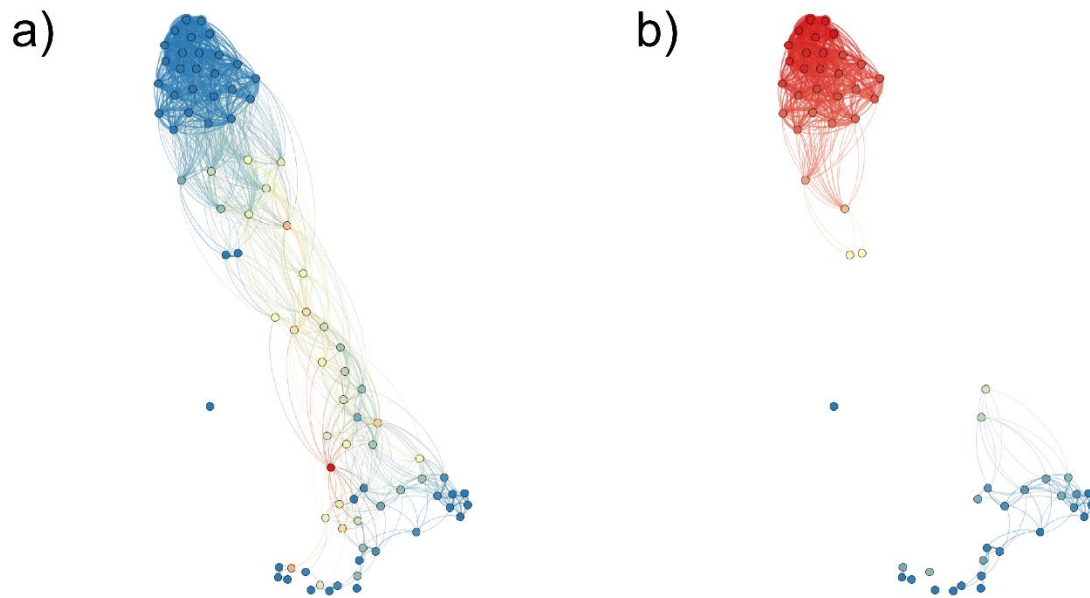
Figure 3.6: **a)** Betweenness centrality and **b)** cells in the 67th percentile in betweenness centrality. The values of both metrics are normalized to fit between 0 and 1. Red is closer to 1, blue is closer to 0. Cells with the highest values are colored in red.

Measuring centrality helps us identify the nodes that can be defined as the most important in the graph, i.e. its backbone. If we remove these most "central" cells of the graph, we end up with a disjointed graph (see **Figure 3.6b**). In one part, we find cells that are almost always round, while in the other part, we find a less structured portion of the graph, comprised of the cells that show a high morphological activity. In a way, since the cells with higher morphological activity are strongly associated with a more malignant profile, it is as if we had "cut" the ties between the general population of the tumor and its most metastatic members. Now, when it comes to a cell population, and to a tumorous population in particular, this can help us identify not only the cells that are very different from the rest of the population (i.e., the ones that could be the vectors of metastasis), as well as the cells that represent the crucial link between mildly offensive cancer cells and the very malignant ones. Identifying these "central" cells and neutralizing them could potentially lead to a way of controlling the metastatic process, by preventing the tumor cells from acquiring the ability to migrate through tissues and settle in new, distant ones.

Discussion

During the metastatic process, cells escape the original tumor by intravasation. During this phase, separated single cells or multi-cellular clusters migrate through the extra-cellular matrix (ECM) surrounding the tumor, and find their way through the endothelium into the bloodstream. Cells can also migrate via the lymphatic system. In the case of the bloodstream, cancer cells end up circulating in a fluid, without attachment to any surface until they are arrested, either by another cluster of cells (cancerous or not) that block their downward flow movement, or by getting stuck in a micro-vessel with dimension similar to that of the cell.. In both cases, prior to arrest, cells undergo mechanical shear stress on their surface. In order to recreate conditions that are close to what cells encounter after intravasation, it is thus necessary to recreate a shear stress, while keeping the cells unattached to a surface. This is what CMR does. The range of rotation speed experienced by the single cells creates levels of shear stress similar to what is found in the bloodstream[33]. In addition, the cells are confined within a triangular well, and even though such confinement has not yet been studied in terms of its biological relevance, it might also play a role in affecting the cell's behavior, by luring the cell into sensing a wall that it might be able to attach to. However, it must be stressed that the microchip is far from being a reproduction of the bloodstream and vasculature. Simply, there is an important rarity of knowledge regarding what happens to cancer cells when they get into the bloodstream. Do they attach to the endothelium as soon as they can, or do they wander in a somewhat mechanically challenging environment and 'decide' suddenly to attach to a new tissue? Do the cells change shape according to their genotypes, or more in response to mechanical and chemical stimuli? For a long time, the impact of mechanical cues from the circulating cells' microenvironment has been, if not simply ignored, strongly downplayed[44]. Therefore, we believe that the CMR method can be used as a way to study the fate of cancer cells, i.e. their changes in morphology, when subjected to a specific mechanical shear stress.

It has been widely reported that there are clear morphological differences between metastatic and much less malignant cancer cells, when they are plated. However, this distinction can become difficult to make with real samples. First, because of the small amount of cells available, especially in the case of Circulating Tumor Cells (CTCs), which are rare, difficult to capture and most difficult to culture. Second, plating cells right after capture/extraction can severely affect their phenotypes[45–47]. In that regard, we believe that keeping the cells *suspended and rotating* allows for faster and more detailed morphological events to be tracked. Even more importantly, we can track behaviors and morphological changes that are intrinsic to the floating or circulating cell *label free*, without relying on any biomarker or *a priori* genotype. Given that heterogeneity affects every cell population, it is important to go after the cells that have a metastatic potential, regardless of their markers. Given that the formation of protrusions, blebs and amoeboid behavior have all been strongly correlated with an increase in malignancy[4,48–50], it follows that detecting such morphological phenotypes could greatly help in the identification of the cells that are most responsible for the metastatic process.

In past decades, the advances in genetic profiling and knock-down experiments (where one or more genes are silenced) have allowed researchers to isolate the impact of motility on cancer metastasis[4]. Beyond motility, it has been discovered that cancer cells could metastasize using different migratory behaviors, such as the ones leading to amoeboid or protrusive shapes. Both these phenotypes have been shown to significantly increase the metastatic potential of cancer cells[4,48–50]. This increased potential is not only due to the higher plasticity of the cells, but due to their ability to switch from one type of motility to another as well, in response to the environment (and, in particular, in response to attempts to knock down or limit the transition towards one of the phenotypes). In addition, cells that show a greater plasticity also show a higher ability to invade tissues and form new tumor colonies, as has been shown when these cells are directly injected into rat tails[4]. Therefore, it is crucial to identify these highly motile and "nimble" cells. Using the presented method, combined with sampling cells over time during therapy, could potentially lead to an assessment of the response to personalized therapy. For instance, one could run the CMR method on samples collected at different time points, and

compare the properties of the different resulting graphs. A more compact graph means that cells are more homogeneous, and if mostly *round*, it would probably mean a better prognosis (and success of therapy) than with a looser graph, with many disconnected nodes, showing a higher extent of heterogeneity.

In our study, we have been able to start from a purely biological perspective, i.e. the shape of rotating cells, and formed a relationship graph based on the similarities between cells. All along the analysis, our method maintains the ability to follow single cells individually. They are all identified with the area code and the position of the trap where they landed in the microfluidic chip. As such, this system lends itself well as a single cell analysis assay. More than just distinguishing "good" from "bad" cells, we can also identify the cells that take part in the transition between such phenotypes. In our future work, our goal is to establish the relationship between the morphology analysis and the extravasation and seeding behavior *in vivo*. We will focus on correlating the information extracted from graph analysis with what happens at the molecular level in the cell by using unsupervised clustering algorithms, and analyzing chemically the types of cells by looking into their DNA expression. We believe that this project has the potential to help identify new sets of genes or proteins involved in the transition between benign cancer cells to highly malignant cancer cells, such as occurs during an Epithelial to Mesenchymal Transition (EMT).

In addition, we have shown here the important degree of heterogeneity characterizing a population of cells from a cultured cell line. We are very hopeful that such heterogeneity can also be shown and exploited when using cancer cells from patients. Indeed, potentially, one could follow the response to treatment by forming this kind of similarity graphs at different time points during therapy. More than the specific properties of the graphs (diameter –length between the two most distant points-, the number of clusters, centrality measures, etc...) the way these properties change in time might help in evaluating and adjusting the therapeutic strategy (especially in its ability to reduce malignancy).

Conclusion

In summary, we have here successfully demonstrated a new *dynamic histology* method, based on combining the Cell Magneto-Rotation (CMR) technique with Machine-Learning methods, so as to identify and classify highly plastic cells from a metastatic population of breast cancer cells, based on the evolution of their morphology. Our massively parallel single cell analysis assay looks into the similarities and dissimilarities of cells' behaviors over time and, using graph theory, we could identify cells whose phenotype may be associated with a more, or most, malignant potential, *without the use of any biomarker*. This technique lends itself well to mapping the heterogeneity characterizing a tumorous cell population, and identifying shifts in behavior and morphology within this population. We also demonstrated how we could dissect a heterogeneous population of cells into groups of cells with behaviors strongly correlated with their metastatic potential. We plan to concentrate our work on extracting the cell, after characterization by image analysis, and confronting its behavior with the analysis of its gene expression. In addition, we plan to increase the potential of this method by building an automated prototype, so as to increase the quantity of cells observed, as well as the ability to construct a predictive model for classifying cells without the intervention of a trained operator.

Materials and Methods

Microfluidic trapping system and cell loading

The microfluidic trapping device is made of Polydymethylsiloxane, according to the protocol used by Park et al.[34] (Micro and Nano, 2009). Each well has a triangular shape, with a side size of 40μm and a depth of 35μm. The dimensions of the triangle can be adapted as function of the size of the cell population being used. To avoid potential cell adhesion, the devices are

dipped into a 3% solution of Pluronic F68 for 24 hours, and then rinsed with Phosphate Buffered Saline (PBS). The chip has two ports: An inlet port and an outlet port. Cells are loaded with a 100µL pipetter into the inlet, and gently suctioned above the traps by pipetting from the outlet. Once positioned, the device is put on top of a rare earth magnet to pull the cells down. We repeat these steps several times, until we get a satisfying loading ratio (above 60% of the traps occupied by single cells). This loading steps take around 3 minutes, and no more than 5 minutes. Finally, cells are washed with fresh media by gently pipetting fresh media into the device (fresh media is placed at the inlet port and pipetted from the outlet port).

Cell Culture and Magnetization

Breast cancer MDA-MB-231 cells transfected with Green Fluorescent Protein (GFP) were obtained from Dr Kleer's lab at the University of Michigan and cultured in RPMI 1640 supplemented with 10% fetal bovine serum (FBS) and 1% Penicillin-Streptomycin-Glutamine (PSG). Media and supplements were all purchased from Life Technologies©), in a cell incubator at 37°C, with 5% CO_2 and 100% humidity.

Amine Coated Magnetic nanoparticles (Ocean Nanotech©) with a diameter of 30nm, are prepared in a 1mL stock solution of 200µg/mL in cell culture media. We then add 15µL of Poly-L-Lysine at 0.1%w/v (Sigma-Aldrich©), and the solution is left for an hour on a rotator at room temperature.

To magnetize the cells, magnetic nanoparticles (MNPs) are added to the cells and cell media at a final concentration of 8.5µg/mL. Cells' confluency before addition of the MNPs is of around 20-30%. Cells are incubated overnight in this media. Cell confluency or incubation period can vary in function of the desired cell density.

When suitably confluent, cells are washed three times using Hank's Balanced Salt Solution (HBSS, Life Technologies©), and gently detached using a cell scraper (Fischer Scientific©). Cells are incubated for 24 hours with cell culture media to which are added 20mg/mL of amine-coated magnetic nanoparticles (30 nm, Ocean Nanotech). Before being exposed to fluorescence exciting

light, cells are washed with HBSS three times to remove traces of phenol red contained in the cell culture media, and then incubated for an hour with regular cell culture media, but without Phenol Red, and supplemented with an oxygen radical scavenger, Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma-Aldrich) at 0.25 nM.

After an hour, cells are washed with HBSS, and gently detached using a cell scraper. Cell density is then adjusted by the help of a magnetic separator. Cells are then gently pipetted into the microfluidic device, and using a magnet, cells are pulled into the traps. We repeat this operation between three and five times, until reaching a good trapping efficiency (around 2/3 of occupied traps) and a single cell occupancy ratio of around 60%.

Cell Imaging

Cells are imaged on an Olympus© IX71™ microscope, equipped with an arc-mercury lamp (U-RX-T™) and a digital camera (Olympus© Q-Color 3™). Images are captured with the Olympus© Q-Capture™ software. To reduce cell exposure to fluorescent light, we use a homemade shutter that opens every 60s for 900ms.

Stage Environment

Temperature and humidity are controlled using a homemade, on-stage system that keeps the cells at 37°C with 100% humidity. Cell media is supplemented with HEPES in order to limit the effects of the absence of CO_2 at 5%.

Acknowledgments

References

1. Park S, Ang RR, Duffy SP, Bazov J, Chi KN, et al. (2014) Morphological Differences between Circulating Tumor Cells from Prostate Cancer Patients and Cultured Prostate Cancer Cells. PLoS ONE 9: e85264. doi:10.1371/journal.pone.0085264.

2. Sottoriva A, Verhoeff JJ, Borovski T, McWeeney SK, Naumov L, et al. (2010) Cancer stem cell tumor model reveals invasive morphology and increased phenotypical heterogeneity. Cancer Res 70: 46–56.

3. Sanz-Moreno V, Gadea G, Ahn J, Paterson H, Marra P, et al. (2008) Rac Activation and Inactivation Control Plasticity of Tumor Cell Movement. Cell 135: 510–523. doi:10.1016/j.cell.2008.09.043.

4. Kosla J, Paňková D, Plachý J, Tolde O, Bicanová K, et al. (2013) Metastasis of aggressive amoeboid sarcoma cells is dependent on Rho/ROCK/MLC signaling. Cell Commun Signal 11: 51. doi:10.1186/1478-811X-11-51.

5. Shen Q, Rahn JJ, Zhang J, Gunasekera N, Sun X, et al. (2008) MUC1 Initiates Src-CrkL-Rac1/Cdc42–Mediated Actin Cytoskeletal Protrusive Motility after Ligating Intercellular Adhesion Molecule-1. Mol Cancer Res 6: 555–567. doi:10.1158/1541-7786.MCR-07-2033.

6. Turner N, Pestrin M, Galardi F, De Luca F, Malorni L, et al. (2014) Can biomarker assessment on circulating tumor cells help direct therapy in metastatic breast cancer? Cancers 6: 684–707.

7. Justice BA, Badr NA, Felder RA (2009) 3D cell culture opens new dimensions in cell-based assays. Drug Discov Today 14: 102–107.

8. Abbott A (2003) Cell culture: Biology's new dimension. Nature 424: 870–872. doi:10.1038/424870a.

9. Trumper L, Brady G, Bagg A, Gray D, Loke S, et al. (1993) Single-cell analysis of Hodgkin and Reed-Sternberg cells: molecular heterogeneity of gene expression and p53 mutations. Blood 81: 3097–3115.

10. Dexter DL, Leith JT (1986) Tumor heterogeneity and drug resistance. J Clin Oncol Off J Am Soc Clin Oncol 4: 244–257.

11. Hoey T (2010) Drug resistance, epigenetics, and tumor cell heterogeneity. Sci Transl Med 2: 28ps19. doi:10.1126/scitranslmed.3001056.

12. Meacham CE, Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. Nature 501: 328–337. doi:10.1038/nature12624.

13. Satija R, Shalek AK (2014) Heterogeneity in immune responses: from populations to single cells. Trends Immunol 35: 219–229. doi:10.1016/j.it.2014.03.004.

14. Carding SR, Egan PJ (2002) γδ T cells: functional plasticity and heterogeneity. Nat Rev Immunol 2: 336–345. doi:10.1038/nri797.

15. Powell AA, Talasaz AH, Zhang H, Coram MA, Reddy A, et al. (2012) Single Cell Profiling of Circulating Tumor Cells: Transcriptional Heterogeneity and Diversity from Breast Cancer Cell Lines. PLoS ONE 7: e33788. doi:10.1371/journal.pone.0033788.

16. Boman BM, Wicha MS (2008) Cancer Stem Cells: A Step Toward the Cure. J Clin Oncol 26: 2795–2799. doi:10.1200/JCO.2008.17.7436.

17. Beier D, Hau P, Proescholdt M, Lohmeier A, Wischhusen J, et al. (2007) CD133+ and CD133− Glioblastoma-Derived Cancer Stem Cells Show Differential Growth Characteristics and Molecular Profiles. Cancer Res 67: 4010 –4015. doi:10.1158/0008-5472.CAN-06-4180.

18. Clevers H (2011) The cancer stem cell: premises, promises and challenges. Nat Med: 313–319. doi:10.1038/nm.2304.

19. El-Heliebi A, Kroneis T, Wagner K, Meditz K, Kolb D, et al. (2014) Resolving Tumor Heterogeneity: Genes Involved in Chordoma Cell Development Identified by Low-Template Analysis of Morphologically Distinct Cells. PLoS ONE 9: e87663. doi:10.1371/journal.pone.0087663.

20. Ludwig JA, Weinstein JN (2005) Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer 5: 845–856.

21. Wagle N, Berger MF, Davis MJ, Blumenstiel B, DeFelice M, et al. (2012) High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov 2: 82–93.

22. Ruike Y, Imanaka Y, Sato F, Shimizu K, Tsujimoto G (2010) Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. BMC Genomics 11: 137.

23. Nagarajan V, Varma S, Hurt D, Guo Y, Huyen Y (2010) Microarray Meta-Miner (MMM): An integrated method and a web tool to identify genes with similar expression profile. 2010 2nd International Conference on Computer Technology and Development (ICCTD). pp. 70–74. doi:10.1109/ICCTD.2010.5646068.

24. Mason O, Verwoerd M (2007) Graph theory and networks in biology. Syst Biol IET 1: 89–119.

25. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, et al. (2011) Using graph theory to analyze biological networks. BioData Min 4: 10. doi:10.1186/1756-0381-4-10.

26. Christakis NA, Fowler JH (2010) Social network sensors for early detection of contagious outbreaks. PloS One 5: e12948.

27. Christakis NA, Fowler JH (2013) Social contagion theory: examining dynamic social networks and human behavior. Stat Med 32: 556–577.

28. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, et al. (2013) Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS ONE 8: e61318. doi:10.1371/journal.pone.0061318.

29. Bednarz-Knoll N, Alix-Panabières C, Pantel K (2011) Clinical relevance and biology of circulating tumor cells. Breast Cancer Res 13: 228. doi:10.1186/bcr2940.

30. Hughes AD, Mattison J, Powderly JD, Greene BT, King MR (2012) Rapid isolation of viable circulating tumor cells from patient blood samples. J Vis Exp JoVE: e4248. doi:10.3791/4248.

31. Desitter I, Guerrouahen BS, Benali-Furet N, Wechsler J, Jänne PA, et al. (2011) A new device for rapid isolation by size and characterization of rare circulating tumor cells. Anticancer Res 31: 427–441.

32. Vona G, Sabile A, Louha M, Sitruk V, Romana S, et al. (2000) Isolation by size of epithelial tumor cells : a new method for the immunomorphological and molecular characterization of circulatingtumor cells. Am J Pathol 156: 57–63. doi:10.1016/S0002-9440(10)64706-2.

33. Elbez R, McNaughton BH, Patel L, Pienta KJ, Kopelman R (2011) Nanoparticle Induced Cell Magneto-Rotation: Monitoring Morphology, Stress and Drug Sensitivity of a Suspended Single Cancer Cell. PLoS ONE 6: e28475. doi:10.1371/journal.pone.0028475.

34. Park JY, Morgan M, Sachs AN, Samorezov J, Teller R, et al. (2009) Single cell trapping in larger microwells capable of supporting cell spreading and proliferation. Microfluid Nanofluidics 8: 263–268. doi:10.1007/s10404-009-0503-9.

35. Kessel D, Luo Y (1996) Delayed oxidative photodamage induced by photodynamic therapy. Photochem Photobiol 64: 601–604.

36. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, et al. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 7: R100. doi:10.1186/gb-2006-7-10-r100.

37. Jones TR, Carpenter AE, Lamprecht MR, Moffat J, Silver SJ, et al. (2009) Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. Proc Natl Acad Sci 106: 1826.

38. Kamentsky L, Jones TR, Fraser A, Bray M-A, Logan DJ, et al. (2011) Improved structure, function, and compatibility for CellProfiler: modular high-throughput image analysis software. Bioinformatics: btr095. doi:10.1093/bioinformatics/btr095.

39. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Inf Process Lett 31: 7–15. doi:10.1016/0020-0190(89)90102-6.

40. Kobourov SG (2012) Spring Embedders and Force Directed Graph Drawing Algorithms. ArXiv12013011 Cs. Available: http://arxiv.org/abs/1201.3011. Accessed 18 July 2014.

41. Bastian M, Heymann S, Jacomy M, others (2009) Gephi: an open source software for exploring and manipulating networks. ICWSM 8: 361–362.

42. Jacomy M, Venturini T, Heymann S, Bastian M (2014) ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 9: e98679. doi:10.1371/journal.pone.0098679.

43. Noack A (n.d.) Unified quality measures for clusterings, layouts, and orderings of graphs, and their application as software design criteria, Einheitliche Gütemaße für Clusterings, Layouts und Orderings von Graphen, und deren Anwendung als Software-Entwurfskriterien. Available: http://opus4.kobv.de/opus4-btu/frontdoor/index/index/docId/377. Accessed 18 July 2014.

44. Wirtz D, Konstantopoulos K, Searson PC (2011) The physics of cancer: the role of physical interactions and mechanical forces in metastasis. Nat Rev Cancer 11: 512–522. doi:10.1038/nrc3080.

45. Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, et al. (2008) Dynamic Proteomics of Individual Cancer Cells in Response to a Drug. Science 322: 1511–1516. doi:10.1126/science.1160165.

46. Tay S, Hughey JJ, Lee TK, Lipniacki T, Quake SR, et al. (2010) Single-cell NF-[kgr]B dynamics reveal digital activation and analogue information processing. Nature 466: 267–271. doi:10.1038/nature09145.

47. Lahav G, Rosenfeld N, Sigal A, Geva-Zatorsky N, Levine AJ, et al. (2004) Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat Genet 36: 147–150. doi:10.1038/ng1293.

48. Giri A, Bajpai S, Trenton N, Jayatilaka H, Longmore GD, et al. (2013) The Arp2/3 complex mediates multigeneration dendritic protrusions for efficient 3-dimensional cancer cell migration. FASEB J 27: 4089–4099. doi:10.1096/fj.12-224352.

49. Wolf K, Mazo I, Leung H, Engelke K, von Andrian UH, et al. (2003) Compensation mechanism in tumor cell migration: mesenchymal-amoeboid transition after blocking of pericellular proteolysis. J Cell Biol 160: 267–277. doi:10.1083/jcb.200209006.

50. Roca H, Hernandez J, Weidner S, McEachin RC, Fuller D, et al. (2013) Transcription Factors OVOL1 and OVOL2 Induce the Mesenchymal to Epithelial Transition in Human Cancer. PLoS ONE 8: e76773. doi:10.1371/journal.pone.0076773.

Chapter 4 :
Magnetorotation based Dynamic Histology with an Unsupervised Machine Learning Method for Unbiased Identification of Mesenchymal Prostate Carcinoma Cells (HR-14) *vs*. Epithelial Prostate Carcinoma Cells (PC-3E)

Introduction

With the wide recent acceptance of tumor heterogeneity at the cellular level comes recognition of the importance of identifying the cells that lead the metastatic process. One of the landmarks of this process is the Epithelial to Mesenchymal Transition (EMT), where cancer cells of epithelial phenotype acquire superior abilities to migrate, reorganize their skeleton (plasticity) and resist the relatively high shocks encountered when entering the bloodstream. Even though the precursors of EMT are not yet completely deciphered, there has been a general agreement, and scientific evidence supporting it, on its pivotal role in and contribution to cancer aggressiveness [1–5]. Typically, during EMT, epithelial cancer cells lose their cell to cell adhesion markers and acquire abilities that give them, among others, a higher ability to migrate through cytoskeletal changes, squeeze in tight spaces thanks to increased synthesis of the Extra-Cellular Matrix (ECM), and resist the shocks of circulating in the blood stream. By acquiring the characteristics of mesenchymal cells, cancer cells become more invasive, and are enabled to travel to distant tissues, so as to set up new colonies.

Recent research has shown that EMT is not only a pivotal step of metastasis, but quantifying its spread could also be used as a prognosis and could greatly refine our ability to assess more accurate survival chances of patients[1]. As expected, a higher spread of EMT is correlated with a poorer outcome. However, a quantification of EMT is still a bottleneck and challenge [5].

As we can see, characterizing EMT at a molecular, genetic, motile, morphological or protein expression level is the key to identifying the cells involved in the migration and spread of cancer metastasis [6–9]. The EMT process has been well characterized regarding its effect on the morphology of cells, especially when it comes to plated cell lines. The elongated and spindle-like

shapes of mesenchymal cells strongly contrast with the egg-like shapes of epithelial cells. Along with morphology, protein expression is commonly used to characterize the EMT. Several biomarkers, highly correlated with EMT, have already been found [10], and even though these biomarkers are not expressed by all the cells that underwent EMT, combining them can help increase the rate of detection [11].

Even though there are numerous characteristics of EMT, no standard method able to strictly discriminate EMT cells has been found yet. Instead, it is necessary to combine many of the above features to increase the sensitivity of detection methods. This comes at the expense of the simplicity of the decision making process used, and the increase in the amount of data and variables has naturally led to the use of Machine Learning (ML) processes to help in this task.

ML methods are processes based on various statistical models able to learn from the data in order to cluster and classify objects. The principle of an ML algorithm is thus to build a set of rules to predict outcomes based on inputs. Though already widely used to predict cancer outcomes [12–14], the use of Machine Learning to study the EMT process from a morphological point of view is still in its infancy, with most of the effort focused on the genomic and protein expression levels [11,15–17]. In parallel, when it comes to the detection part, using cell morphology and its features as indicators of EMT could be a powerful method as well, and can constitute an interesting alternative or addition to biomarkers, so as to detect EMT, as mutations might impact the biomarkers expressed at the surface of cells. For instance, Verdone and colleagues have shown that it was possible to separate the epithelial and mesenchymal phenotypes based on nuclear structure and texture of cells imaged in stained tissue sections [18]. However, as suggested by Verdone et al., an important step toward better prognosis would be the identification of EMT cells among CTCs captured from the blood of patients. One of the limitations of using tissue sections is the need to fix the cells, rendering impossible the extraction and study of target cells, and the tissue environment in which cells are found, which contrasts with the freely floating state of CTCs. Histological analysis is also strongly influenced by the specialists in charge of them. As such, transposing an ML algorithm trained by one specialist to another one is almost impossible [19]. Moreover, the algorithm used by Verdone and colleagues, like many other instances, is a *supervised* one, meaning that one needs to know *a priori* the

number and descriptive features of each of the phenotypes expected to be clustered and classified, thus severely limiting potential discovery of new rare subpopulations. However, with *unsupervised* algorithms, there is no need for human operators, and thus limiting possible biases or the overlooking of rare events that can have a definite impact on the progress of the disease, as well as a negative impact on its therapy.

In addition, most of the methods used to capture CTCs are biomarker based, and the biomarkers used, such as the Epithelial Cell Adhesion Molecule (EpCAM) used in CellSearch, only target epithelial cells and leave out the mesenchymal ones. This can be solved by using label-free filters, a technique that also has the advantage of avoiding the issue of mutations of biomarkers during therapy. In this case, the CTCs that are captured more broadly reflect the actual CTC population in the bloodstream, but there is still a need to identify the EMT cells after they have been captured. Plating captured cells has so far had a very low success rate, and cells phenotype can be altered during this process[7], and this calls for the necessity of identifying the cells as soon as they are captured.

To this end, we developed a method to identify EMT cells based on their morphology in suspension. We make use of a large-scale, multiplexed, dynamical morphology based analysis, relying on the Cell Magneto-Rotation (CM) method. Cancer cells, rendered magnetized with the uptake of magnetic nanoparticles, are individually trapped in a microfluidic device, which was specifically designed for this purpose[20]. Because migrating cells often end up floating in the bloodstream, we hypothesized that their behavior in a non-adherent environment would be significantly different from the one shown by the rest of the tumor population. More particularly, since the cells before and after EMT are known to show very different properties in terms of migration abilities and plasticity, our hope has been to accentuate and widen the expression of this difference between the two phenotypes, so as to facilitate their identification.

Epithelial PC-3E and mesenchymal HR-14 Prostate Carcinoma cells were first transfected with Green Fluorescent Protein (GFP) so as to accurately track their morphology using fluorescence microscopy (cf. **figure 4.1**).

Cells are flowed in a microfluidic device as designed by Park et al. [20], and are trapped in individual wells. Owing to the use of a very high resolution monochromatic camera (QImaging

86

Retiga 6000, QImgaing), and of a programmable motorized microscope stage in the X, Y and Z directions, we were able to increase the number of cells that are monitored, simultaneously. All of the different pieces of equipment of the setup are controlled using the Micro-Manager software [21] and a custom made script in Java to control the camera, the stage and the shutter.
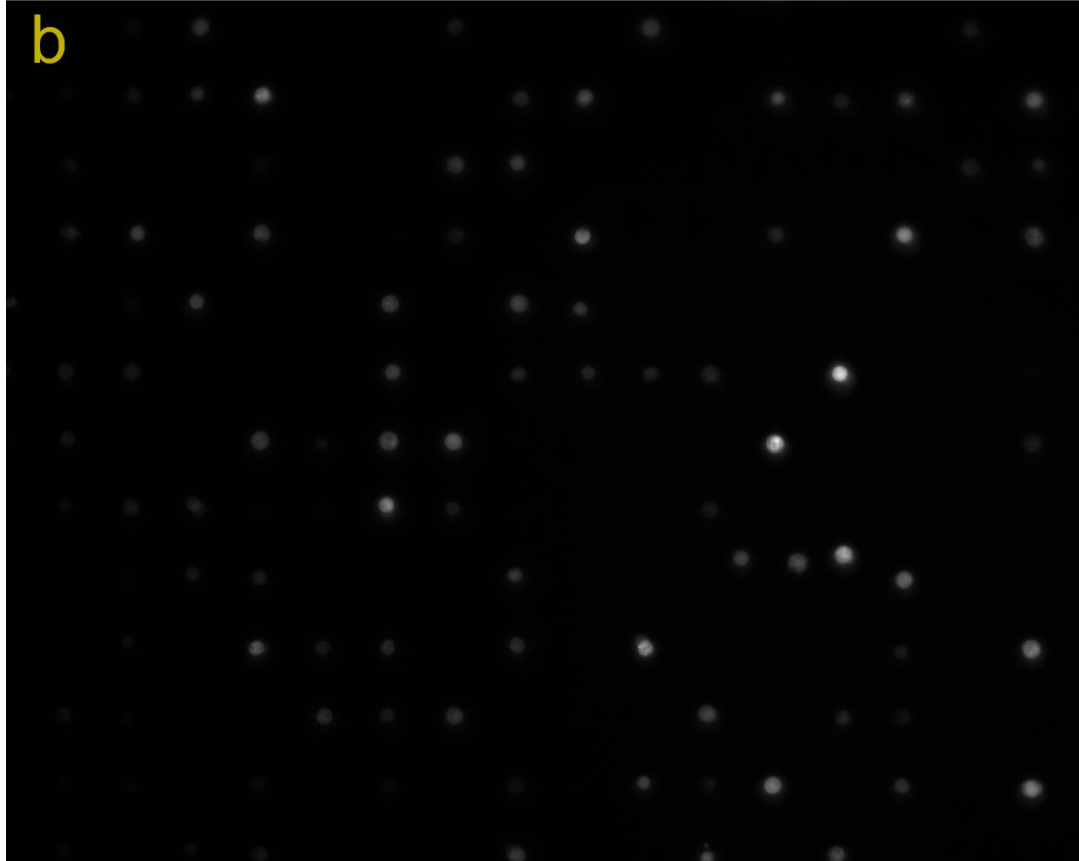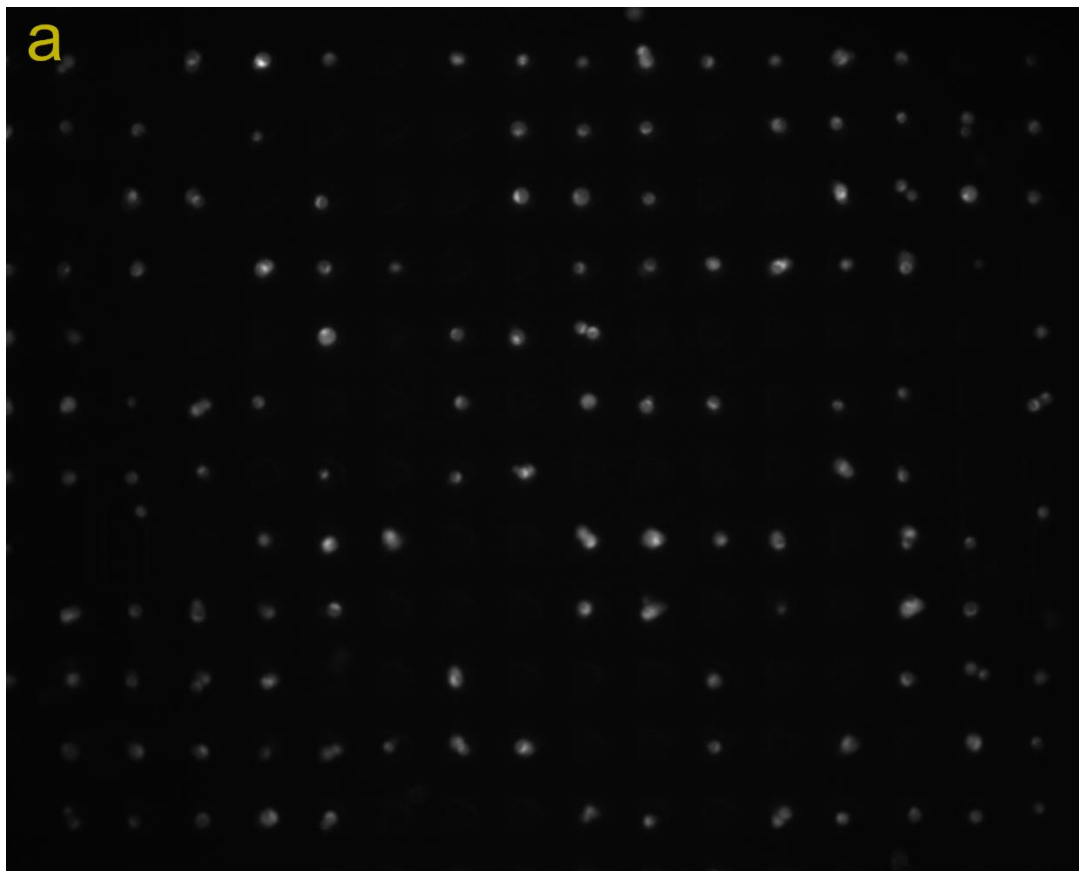
Figure 4.1: Monochromatic captions of **a)** PC-3E (epithelial) cells and **c)** HR-14 (mesenchymal) cells trapped in the microwells of a subsection of the microfluidic device.
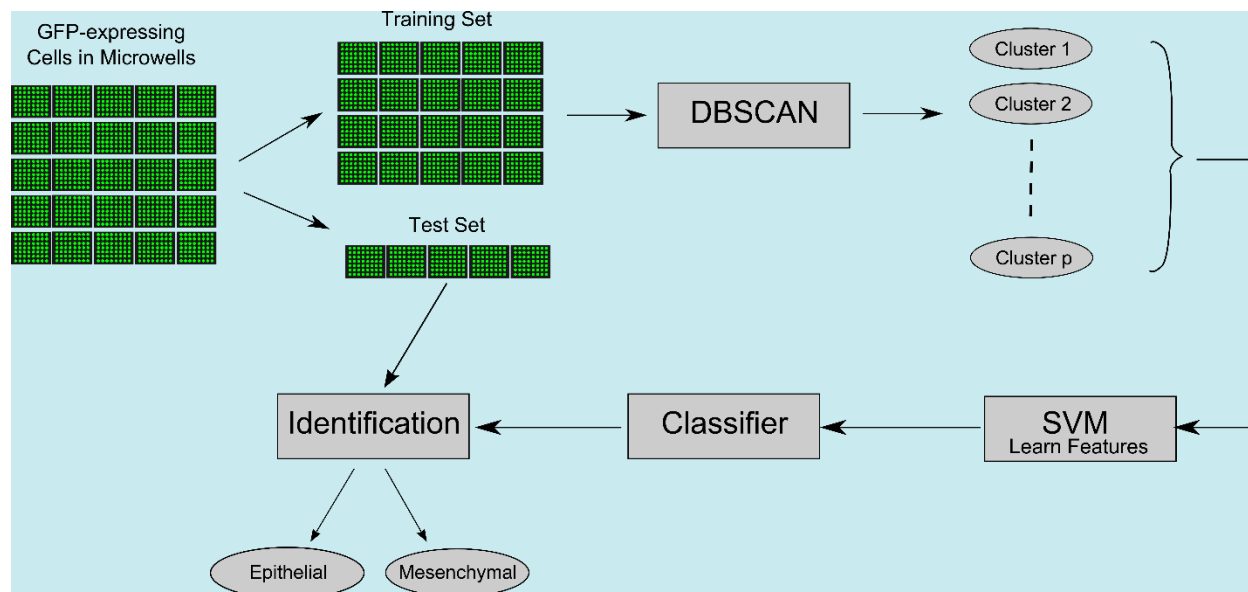
Figure 4.2: Flow chart of the analytical process. Grid of GFP expressing cells are imaged sequentially, the microscope stage moving from one section to the other and then loops again. The same section is visited and imaged once every 60 seconds. Cell images are then sent to analysis using CellProfiler software on a High Performance Computing cloud, and the data set is then split into a training set and a test set. For clarity reasons, the random shuffling step was omitted.

Before we get into the details of the various steps, let us describe the general process (see flow chart in **figure 4.2**). Once cells are imaged and analyzed, we split the dataset into two randomized subsets: a training set (composed of 70% of the cells), and a test set (the reminding 30%). The training is then processed with the clustering (DBSCAN) algorithm, which detects subpopulations in the dataset. It is important to note here that every occurrence of the DBSCAN algorithm results in a different clustering because of the randomization of the training set (to avoid overfitting). The number of found clusters can also vary, but we found (see below) that overall the number of clusters was quite stable. In addition, the clustering is *completely unsupervised*, meaning that *no* specific instructions regarding phenotypic specificities are used as input for the algorithm. Cells are thus clustered in a purely mathematical way, and the biological relevance of each cluster is analyzed in a post-processing manner. The benefit of such an approach is the possibility to find subpopulations of cells that would be difficult to classify or

predict *a priori*. Each data point is thus labeled with the class it was assigned to by the clustering algorithm. We can then use the dataset and the label to a Support Vector Machine (SVM) algorithm, which computes boundaries in between the different classes, and builds a decision function. At this point, the phenotype of the cell does not appear anymore in the datasets, but we can associate each class with a specific phenotype. The decision function F, in our case, will have polynomial dependencies on the variables (features). If fed a data point $\vec{x} = (x_1, x_2, \dots, x_n)$, where *n* is the number of features, $F(\vec{x})$ returns the predicted class of $\vec{x}$. Once the classifier is ready, we apply the decision function F to each of the data points in the test set, and classify them among the different clusters. With the correspondence between clusters and phenotypes found earlier, we can then assign a phenotype to each test data point. We finally compare the predicted phenotypes with the actual phenotypes of the test set, so as to evaluate the performance of our algorithm.

The ML algorithms can be split into two categories: supervised and unsupervised learning algorithms. In supervised learning, an operator has to indicate the number of expected classes in which the observations should fit, while in unsupervised learning, it is the task of the algorithm to find patterns in the observations and find the number of classes. One of the major drawbacks of a supervised ML is that it requires the intervention of a human operator, who might be prone to bias. To overcome this issue, we utilized an *unsupervised* ML algorithm, and more specifically the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) [22]. Broadly speaking, the algorithm takes two arguments: a distance parameter ε and a density parameter *minPoints (number of near neighbors):* Starting at a random data point P, the algorithm looks at the number of other data points that are at a distance from P that is smaller than ε. If this number is higher than minPoints, then it considers P to be a cluster, otherwise, P is considered as noise (so far in the algorithm loop). Then, it visits a new, yet unvisited data point Q, and applies the same ruling. However, if Q and P are distant by less than ε, and if Q has more than minPoints neighbors and P is already in a cluster, it adds Q to the corresponding cluster.

To compare PC-3E and HR-14 cells, we imaged a wide number of cells (1417 PC-3E and 649 HR-14 cells), each imaged for 60 mins, rendering a total of around 196,000 data points. However, unfortunately, not all the datapoints could be used. For instance, we filtered the data

points where the object recognition software captured two objects instead of one, still resulting in around 131,000 final data points, representing around 2500 distinct cells. For each data point (a cell image at a specific time point), we measure 112 morphological features (area, shape, angular distribution, texture, etc…), using the image analysis software CellProfiler. To be able to compare both data sets (epithelial and mesenchymal), we standardized them to have, for each feature, an average of 0 and a standard deviation of 1. Simply put, for each data point X and feature i, we replace X[i] by $\frac{X[i]-E(feature_i)}{\sigma}$, where E(feature-i) is the average of the i-th feature over all the data points. Next, of all the features we measure, some of them might be redundant, i.e. their high level of correlation does not give more information, and this redundancy just uses more memory and slows down the algorithm. To cope with this issue, we performed Principal Component Analysis (PCA), and reduced the number of features to only 14. However, it has to be said that the new features that we get (the principal components), are not actual features anymore, but the result of an orthogonal linear transformation that makes the new components, called principal components (or eigenvectors in more general cases), linearly independent from each other.

Results

We split the data into a training data set (70% of the whole data), and a test set. In ML, one of the major issues is the problem of overfitting, in which the computer learns "too well", and closely fits the training set, leading to very good classification results on the training set but poor results on new data. For this reason, the test set and the training set are made of completely distinct sets of cells. Therefore, the computer learns on a set of data points from one set of cells, and is then tested on another set, with new cells. We then proceeded in three steps. First, using unsupervised clustering we find the different noticeable clusters (subgroups), then build a classifier based on these subgroups, assign each of these subgroups to a specific phenotype, and finally test the classifier on the testing set.

The first step is to regroup data points in clusters that share a lot of similarities. At this point, we need a bit more information to evaluate the outcome, namely the phenotypic

composition of the clusters. Simply put, we need to find out how phenotypes are spread among clusters, what really happens to the outliers and "who" they are. To do so, we can use two different metrics called *homogeneity measure* and *completeness measure*. The homogeneity of the clustering measures if each cluster contains only members of a single class (i.e. phenotype), while the completeness measure relates to the question whether all the members of a given class were assigned to the same cluster. Each measure is between 0 and 1, where 1 means a perfect clustering and classification. Let $A = \{A_1, A_2, \ldots, A_n\}$ be the true classes of data points that we have ("the ground truth"), and $C = \{C_1, C_2, \ldots, C_l\}$, the classes obtained after clustering operations. We will set N to be the total number of data points. Let $a_m = |A_m|$ be the number of objects (i.e. cells) belonging to the m-th class, $c_k = |C_k|$ be the number of objects classified in the k-th cluster by the algorithm, and n_{mk} the number of objects that belong to both A_m and C_k. We can then define the homogeneity measure by:

$$homo = \begin{cases} 1 \; if \; E(A, C) = 0 \\ 1 - E(A/C)/E(A) \end{cases}$$

where $E(A/C) = -\sum_m \sum_k \frac{n_{mk}}{N} \log(\frac{n_{mk}}{a_k})$ and $E(A) = -\sum_k \frac{a_k}{N} \log\left(\frac{a_k}{N}\right)$,

which is the entropy of set A.

Similarly, the completeness is defined by:

$$comp = \begin{cases} 1 \; if \; E(C, A) = 0 \\ 1 - E(C/A)/E(C) \end{cases}$$

where $E(C/A) = -\sum_m \sum_k \frac{n_{mk}}{N} \log(\frac{n_{mk}}{c_m})$ and $E(C) = -\sum_m \frac{c_m}{N} \log\left(\frac{c_m}{N}\right)$,

which is the entropy of set C.

After we have done that, we can run the DBSCAN algorithm. To train the computer, we feed the algorithm with a training set made of 70% of the data set. Once again, in order to avoid issues with overfitting, every time we train the computer, we feed the DBSCAN algorithm (and hence the classifier) a random set of cells. To do so, cells are randomly assigned as part of the training set or as part of the test set. Namely, from the list of cells of each phenotype, we randomly pick 70% of them as training set, and the reminder as a testing set. The shuffling

algorithm can be found in **Appendix C**. This ensures that the results we find are not just artifacts of one particular training session. Clustering results vary from one pick to another, but the number of found clusters is relatively stable for set values of ε and minPts. In order to choose the value of these two parameters, we first performed a grid search. Basically, we ran the DBSCAN algorithm with different values for ε (from 2 to 4 by increments of 0.1) and for minPts (from 75 to 140 by increments of 5), and used the couple (ε, minPts) that gave the highest homogeneity on average (a homogeneity of around 0.45 and a number of clusters on average at 7).

With values of ε = 2.7 and minPts = 95, we end up with 7 different clusters plus the group of outliers, as shown in **figure 4.3** (where we removed the outliers, i.e. the datapoints that did not belong to any cluster).
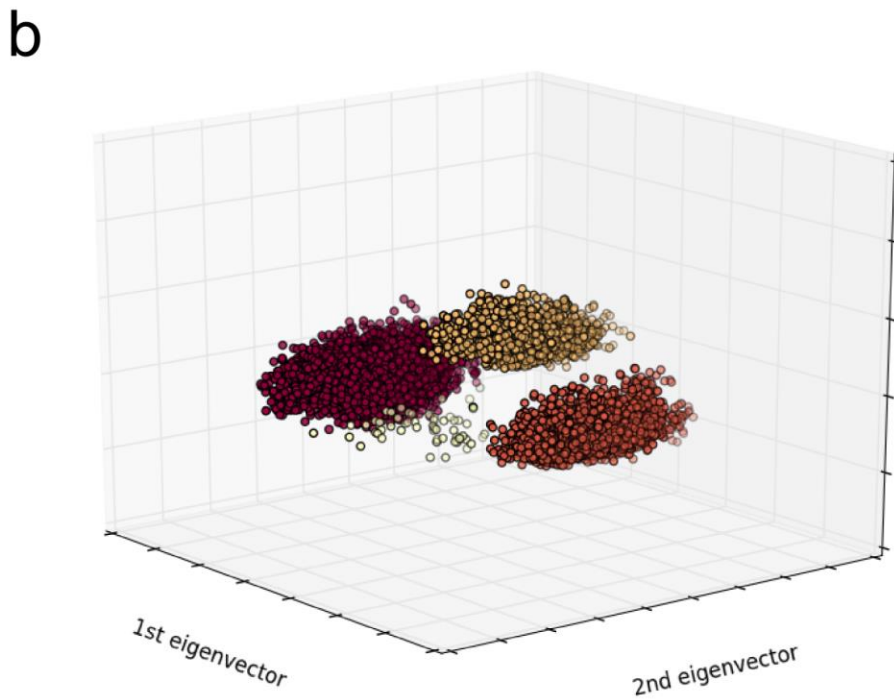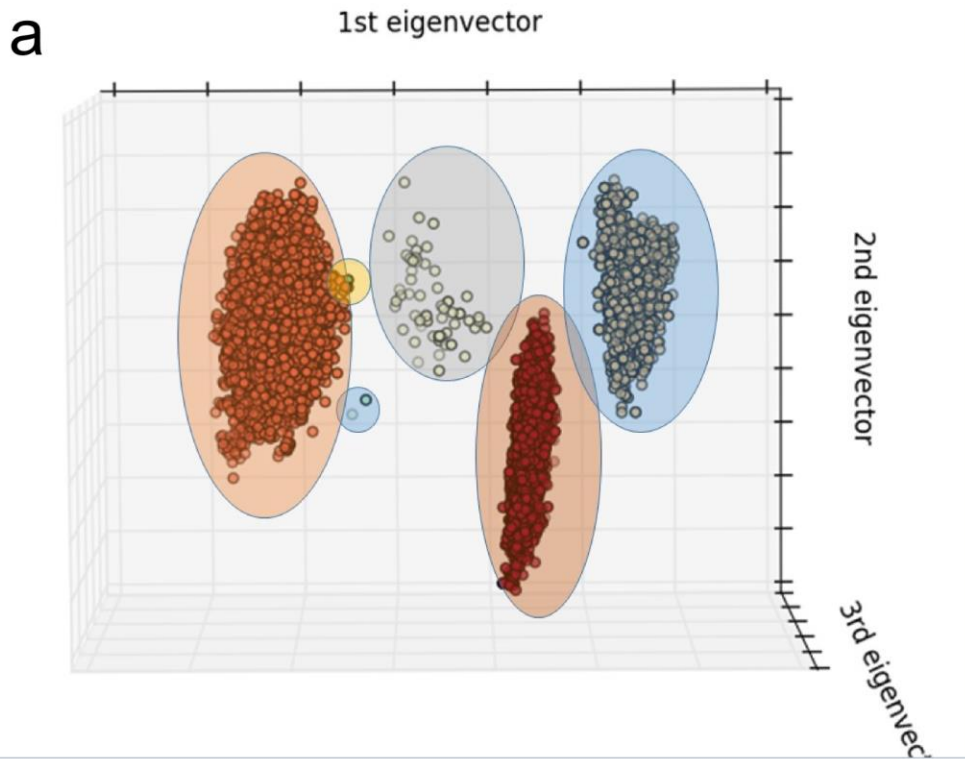
Figure 4.3: DBSCAN clustering results plotted against the first 3 eigenvectors. In **a)** clusters are haloed, while **b)** show a different angle of the same data points. For clarity, data points belonging classified as outliers were omitted.

When we plot the data points in the three dimensional space of the first three eigenvectors (i.e. the three principal components that account for the highest variance), we notice 3 big clusters and 3 smaller ones. For the sake of clarity, we only displayed classified datapoints, and not the ones considered as outliers. Later on however, we will use the *class of outliers* as a separate cluster when it comes to recognize cell phenotypes. Even though we can clearly see well defined clusters, we need to look deeper into other metrics to know if they can be used to train a recognition algorithm. Basically, the more homogeneous each cluster is in terms of phenotype composition, the better for ulterior training and classification in terms of accuracy. When we compute the metrics of our clustering, we find an average homogeneity of 0.413, and an average completeness ratio of only 0.272, which is not very high (see **table 4.1**). To recall, the completeness metric measures if all the cells of the same phenotype are put together in the same cluster, while the homogeneity measures if in each cluster, we find only cells of the same phenotype. Since we end up with more clusters than we have phenotypes (two in total – epithelial and mesenchymal), this explains the fact that we have a relatively low completeness coefficient, and we need to look into the composition of each cluster to find out if we are successfully clustering phenotypes.

Epsilon	minPoints	Number of data points	Number of Clusters	Homogeneity	Completeness	Number of Outliers	Epithelial outliers (%)	Mesenchy-mal outliers (%)
2.7	95	92488	8	0.413 (+/- 0.035)	0.272 (+/- 0.016)	57131 (+/- 631)	82.8%	17.2% (+/- 0.51)

Table 4-1: Classification report from the DBCAN algorithm.

To do so, for each cluster, we computed the ratio of data points belonging to the mesenchymal type to the total number of data points in each cluster, outliers included. At a closer look at the repartition, we can see that the outliers are overwhelmingly from the epithelial phenotype, with 82.8% of all the outliers being epithelial, while in the whole dataset (after filtering), the epithelial data points represent around half of all the data points. More

95

importantly, the clusters that we obtained using the DBSCAN algorithm are composed of almost pure populations. Given how the phenotypes are spread among clusters and outliers, it is striking that the epithelial population is more diverse and diffuse, which is indicated by the difficulty to find big clusters of PC-3E cell data points, while the mesenchymal cells seem to be very homogeneous in the way they are different from their epithelial counterparts, as suggested by the three clusters that comprise almost all the mesenchymal data points (see **Table 4-2**). Since we seem to have very distinct features between the two phenotypes, can we train a computer to recognize each phenotype?

cluster	Population (average)	Mesenchymal Ratio (average)
-1	57131	0.172
0	13148	0.935
1	14174	0.904
2	6951	0.965
3	609	0.535
4	177	0.267
5	177	0.154
6	122	0.222
7	122	0.000934579

Table 4-2: Average cluster composition after DBSCAN clustering (averaged over 15 runs).

Classifier's results

To do so, we compared several classifiers: k-nearest neighbors, decision tree, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), a naïve Bayes classifier, a random forest classifier and two support vector machine (SVM) based classifiers. Simply put, SVMs are supervised machine learning algorithms used to classify objects into two categories. Intuitively, in the simplest case, the two classes are linearly separated (i.e. by a plane in 3D), and for any new data point fed to the classification algorithm, its classification is decided based on which side of the plane it falls. Even though it seems rather simple, SVMs are surprisingly very powerful. The model can be extended to multiple classes, and this is what we do here, since our cells are clustered in distinct and numerous clusters.

From the classes (clusters) obtained above, we use a classifier on the same training set to build rules to classify the data points. Once this is done, we can use this newly made classifier on our test-set to validate or improve the classification if need be.

Each data point from the test-set is classified as a member of one of the seven classes obtained from the earlier clustering on the training set (7 clusters and 1 group of outliers). Given their high purity, for each cluster, we assign it to the dominant phenotype represented in its population. Therefore, we reduce the number of classes to two, namely "epithelial" and "mesenchymal", which will allow us to evaluate the performance of our classifier.

We tested two SVM different classifiers. A linear SVM based classifier, and a non-linear SVM classifier with a polynomial (degree 3) kernel. Kernels are widely used in classification problems where non-linear boundary decisions are needed, because of their simplicity, computational cost and the increase in classification accuracy. Basically, if $\{x_i\}_{i=1..n}$ is our set of attributes (variables) of our problem (i.e. area, perimeter, angular distribution etc...), we can use a mapping of the attributes instead of the attributes themselves in the non-linear case. And the mapping that we use here is in the form of:

$$k(x, x') = (\gamma \langle x | x' \rangle)^d$$

Where we used d = 3 and γ = 1/14 (the inverse of the number of features after PCA transformation).

Even using a linear model in a multi-class scale, our algorithms successfully classify between 85 and 92% of the new data points (at the exception of Adaboost, which is a much simpler classifier and does not perform as well). It is worth noting that all the classifiers perform much better than what a random classifier would do, namely 50%. The best performing classifier is the LDA one, with a f1-score of 0.922, followed by a Naïve Bayes classifier recording a f1-score of 0.89.

	precision	recall	f1-score
Adaboost	0.78	0.758	0.743
Decision Tree	0.85	0.848	0.847
K-NN	0.87	0.871	0.871
LDA	0.92	0.922	0.922
Naïve Bayes	0.89	0.89	0.89

SVM Polynomial	0.86	0.857	0.854
QDA	0.86	0.853	0.853
Random Forrest	0.89	0.880	0.877
SVC Linear	0.86	0.856	0.853

Table 4-3: Comparison of different classifiers (each classifier was run 30 times, polynomial SVM 15 times). Each result is the average value obtained after the runs.

The classification reports for our classifiers are shown in **table 4.3**. *Precision*, or positive predictive value, is the probability that a random data point is correctly classified. Namely, given a class C, precision is the ratio of data points correctly labeled, over the total number of data points truly belonging to the class. On the other hand, the *recall* of a class C, also known as sensitivity, is the number of correctly classified data points, divided by the total number of data points classified in C. The number *1 - Precision* gives the *ratio of false negatives*, while *1 – Recall* gives the *ratio of false positives* (see **Appendix** at the end).

Interestingly, the linear and non-linear SVM perform very similarly. Given its simplicity and its speed, a linear classifier is surprisingly robust. To recall, a random classifier that would have been trained on the same set would assign a probability $p_{epithelial} = \frac{N_{epithelial}}{N_{total}} = 0.56$ for a cell to be epithelial (the number of epithelial cells that we monitored was slightly higher than the number of mesenchymal ones), and $1 - p_{epithelial}$ for it to be mesenchymal, regardless of the cell's features. This would give an average precision and a recall of both 0.50.

In view of the above, a classifier using a linear discriminant analysis is capable of correctly classifying a mesenchymal cell with a probability of 92.3%, while only 10.6% of the mesenchymal cells are classified as false negatives (see **Appendix D** for detailed results for each phenotype). A way to measure the efficiency of a classification is by measuring its *precision* and *recall* on each category to be classified. *recall* and *precision* are estimators of the actual probabilities to pick a cell correctly (*recall*), and to pick all the correct cells. Because we do not have access to the true values of these probabilities, we use these estimators instead and use them as if they were the true probabilities.

On the other hand, epithelial cells are classified with a precision of 0.92 and a recall of 0.94, and this is also a very high success rate.

98
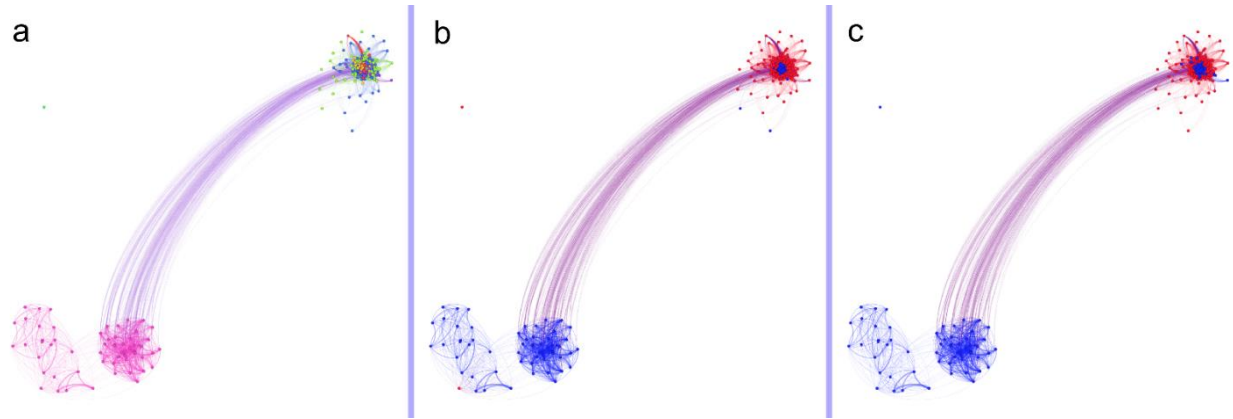
Figure 4.4: OpenOrd layout of the relationship graph of PC-3 epithelial and mesenchymal cells **a)** with modularity classes shown as colors, **b)** with true phenotypes shown and **c)** with predicted phenotypes shown. In **b)** and **c)**, epithelial cells are represented in red, mesenchymal in blue.

The classifier having been trained and showing high accuracy, can we identify the phenotype of cells with only one picture, and how accurate will be the outcome? Namely, how can we use it in a practical way to identify individual cells that might show a very different profile than the rest of the population? One particular way that we envision to use our method would be to feed the classifier with a new set of cells, from epithelial and mesenchymal populations, and visualized the results using relationship graph visualization tools, such as the one we used, the open source software Gephi [23].

To do so, we loaded into the classifier data points from one single time point, in a way to simulate a snapshot of a cell population. Therefore, each cell is imaged only once, and we can assimilate a data point to a "cell". As we did previously, we only utilized data points that were not used to train the classifier. In one step, we classified the cells and recorded the prediction result. In parallel, we computed a distance matrix with the data set. From the distances between each data point, we built a relationship data base. For each couple of cells, we assigned a weighed to their relationship proportional to the square of the inverse of the mutual distance, and graphically, this relationship is represented by an edge linking two points. In order to help with clarity and to lower the computational burden, we only considered the 30th percentile in terms of weight, and then visualized the relationship network using the OpenOrd algorithm, a Force-Directed algorithm [24]. The results are shown in **figure 4.4**, and just as with DBSCAN, clear

clusters appear. However, as shown in **figure 4.4a**, the clusters are not exactly the same, and, in particular, we find an aggregate of cells comprising a mixture of epithelial and mesenchymal cells. Still, we find several groups of cells that are clearly different than the majority of the population, as pointed out by DBSCAN as well. What can we say about the groups of cells? To get a better discrimination power, we compute the modularity of each node. The modularity of a node is the difference between the number of edges within a group and the expected number of edges within this group (i.e. the number of edges within this group if edges were randomly distributed). Computing the modularity of a graph is a maximization process, by varying the number of groups and the members of each groups. We find 8 different modularity classes, and it can discriminate between groups of nodes that seem very close on the graph. As a control, we also looked at how each modularity class was spread in between the two phenotypes, and compared to what the classifier would predict. Surprisingly, the modularity classes manage to separate rather well the phenotypes, and we can now identify the groups of cells leading to classification errors. For instance, class 6 seems to account for most of the misclassified epithelial cells. However, the proximity of the cells in the graph suggest that the cells do look very similar.

Modularity Class	Population	Pct Total Population	Pct Epithelial (True)	Pct Epithelial (Predicted)	Pct Mesenchymal (True)
0	182	27.29	99.45	95.05	0.55
1	156	23.39	98.72	98.08	1.28
2	116	17.39	26.72	29.31	73.28
3	74	11.09	2.70	6.76	97.30
4	61	9.15	0.00	1.64	100.00
5	30	4.50	20.00	16.67	80.00
6	47	7.05	8.51	40.43	91.49
7	1	0.15	0.00	100.00	100.00

Table 4-4: Phenotype repartition by modularity class, with actual values and values predicted by classifier.

Now that we can confidently discriminate cells of a more aggressive phenotype from the epithelial ones, can we identify the features that contributed to differentiate the two phenotypes?

To answer this question, we computed the features' importance, given the classification that we obtained, and filtered the ones that cumulatively accounted for more than 50% of the

variance between the different clusters, using the original dataset, i.e. before dimensionality reduction with PCA (where we lose the link to the original features). **Table 4.5** shows the list of selected features, along with their importance. Even though some of the features are purely morphological ones (such as the area), we notice an overwhelming share of texture related features.

Measured feature	Contribution to variance (%)
Cell_Texture_SumEntropy _10_0	2.38
Cell_RadialDistribution_MeanFrac_ 7of 8	2.44
Cell_Texture_SumVariance_ 10_0	2.47
Cell_Texture_SumAverage_ 20_0	2.74
Cell_Texture_SumEntropy_ 20_0	2.84
Cell_Texture_InfoMeas2_ 10_0	3.20
Cell_RadialDistribution_RadialCV_ 5of8	3.62
Cell_Texture_InverseDifferenceMoment_10_0	3.72
Cell_Texture_AngularSecondMoment_10_0	3.89
Cell_Texture_SumAverage_10_0	4.50
Cell_Texture_AngularSecondMoment_3_0	4.78
Cell_Texture_SumEntropy_3_0	5.94
Cell_Texture_InverseDifferenceMoment_3_0	7.83

Table 4-5: List of features with a cumulative importance above 50%.

Discussion

DBSCAN is one of the most used clustering algorithms to date [22]. Compared to other clustering methods, it has the advantages of being unsupervised (the number of clusters do not have to be known a priori). Also, the clusters do not have to be linearly separable, and they can have very different sizes. Considering the problem of classifying cancer cells, human operators usually face an issue when having to choose the number of morphological categories first, and how to classify each cell in a second time. As DSCAN is unsupervised, we can easily cope with the first step, and this makes the method presented here very flexible and adaptable to new types of cancer cells, or cell populations in general. In addition, contrary to some other clustering techniques, such as k-Mean clustering [25,26], the clusters can have very different sizes. This is a significant advantage, to identify the smaller subsets of the population that might have the biggest impact on tumor regrowth and metastasis. However, one of the drawbacks of DBSCAN is

that it is density dependent. Depending on the minimum number of points in a cluster (minPoints) and the distance ε that we assign, the results of the algorithms (number of clusters, population of each cluster, etc...) can vary significantly, regarding clusters that are more or less dense. More complex implementations of DBSCAN (such as OPTICS – Ordering Points for the Identification of the Cluster Structure) exist [27], and they solve this problem. However, this latter implementation is not publicly available on diverse platforms, and as a proof of principle for a new method, the end point being the successful classification of cancer cells of different phenotypes, DBSCAN was very effective at providing the baseline clusters that we used to train the computer.

One of the reasons for the low impact of classical morphological features can be the redundancy of many of them. They can be put into a restricted number of subgroups where all the features are highly correlated. As such, as a whole, they do not individually impact much on the variance, especially after dimensionality reduction. In addition, texture and angular distribution can capture a lot of information regarding the structure of the cell, in contrast to typically geometric measures, such as the area. In particular, it was noted that the nucleus had a very distinct architecture between the epithelial and mesenchymal forms of PC-3 cells. Even though we do not directly measure the morphology of the nucleus, this is still taken into account by our measurements of the texture and other pixel intensity distributions of the cytoplasm. In addition, the cytoskeleton is playing an important role in EMT (such as the loss of polarity for instance); it greatly affects the internal organization of the cell, which can be partially captured by the texture related measurements. Obviously, tagging the different parts of the cell (nucleus, cytoskeleton, membrane) with markers of different wavelength would probably give an even greater power of discrimination, and even allow us a better understanding, not just of *inter*-phenotype differences, but also of *intra*-phenotype differences. However, it is worth mentioning that even with a much simpler setup that do not use specific biomarkers and only one fluorescent wavelength filter, we were still able to capture the heterogeneity of the mesenchymal cell population. Regarding their epithelial counterparts, the heterogeneity could also be captured, but a more sophisticated system would allow us to find the cells or group of cells at the origin of this heterogeneity.

Regarding the area, it has to be said that a very small fraction of cells, less than a handful per thousand cells, are super-cells, i.e. multinuclear cells. These cells have a radius of around 100um, cannot settle in our wells, and are therefore filtered out. Nonetheless, these cells, regardless of their phenotype, are easily identified.

Conclusion

In conclusion, we have developed a method that can distinguish EMT cells, the driving force of cancer metastasis, from the rest of the cancer cell population, by analyzing the morphology of cells when magnetically rotated and in suspension. Our results show that we can cluster cells together by analyzing morphological features, in an *unsupervised* way, thus bypassing the need for a bias and error-prone human operator, for detecting rare subpopulations of cells that could have a critical impact on therapy development. These features can then be used to build rules for a Machine Learning based classifier. In particular, we can identify cells of different phenotypes that share the same properties, while isolating the most different ones, with an accuracy of around 90%. We believe that this method could be applied beyond the scope of cancer metastasis, and to the analysis of mutations that create cytoskeletal changes in other types of cellular populations, and thus lead to the discovery of atypical cells that are part of the cancer spread phenomena.

Materials and Methods

Microfluidic trapping system and cell loading

The microfluidic trapping device is made of Polydymethylsiloxane, according to the protocol used by Park et al.[20] (Micro and Nano, 2009). Each well has a triangular shape, with a side size of 40μm and a depth of 35μm. The dimensions of the triangle can be adapted in function of the size of the cell population being used. To avoid potential cell adhesion, the devices are

dipped into a 3% solution of Pluronic F68 for 24 hours, and then rinsed with Phosphate Buffered Saline (PBS). The chip has two ports: An inlet port and an outlet port. Cells are loaded with a 100µL pipetter into the inlet, and gently suctioned above the traps by pipetting from the outlet. Once positioned, the device is put on top of a rare earth magnet to pull the cells down. We repeat these steps several times, until we get a satisfying loading ratio (above 60% of the traps occupied by single cells). This loading steps take around 3 minutes, and no more than 5 minutes. Finally, cells are washed with fresh media by gently pipetting fresh media into the device (fresh media is placed at the inlet port and pipetted from the outlet port).

Cell Culture and Magnetization

Epithelial Prostate Carcinoma PC-3 cells were transfected with Green Fluorescent Protein (GFP) and cultured in RPMI 1640 supplemented with 10% fetal bovine serum (FBS) and 1% Penicillin-Streptomycin-Glutamine (PSG). Media and supplements were all purchased from Life Technologies©), in a cell incubator at 37°C, with 5% CO_2 and 100% humidity. Mesenchymal Prostate Carcinoma HR-14 cells were obtained following the protocol described by Roca et al.[4].

Amine Coated Magnetic nanoparticles (Ocean Nanotech©) with a diameter of 30nm, are prepared in a 1mL stock solution of 200µg/mL in cell culture media. We then add 15µL of Poly-L-Lysine at 0.1%w/v (Sigma-Aldrich©), and the solution is left for an hour on a rotator at room temperature.

To magnetize the cells, magnetic nanoparticles (MNPs) are added to the cells and cell media at a final concentration of 8.5µg/mL. Cells' confluency before addition of the MNPs is of around 20-30%. Cells are incubated overnight in this media. Cell confluency or incubation period can vary in function of the desired cell density.

When suitably confluent, cells are washed three times using Hank's Balanced Salt Solution (HBSS, Life Technologies©), and gently detached using a cell scraper (Fischer Scientific©).

Cells are incubated for 24 hours with cell culture media to which are added 20mg/mL of amine-coated magnetic nanoparticles (30 nm, Ocean Nanotech). Before being exposed to fluorescence exciting light, cells are washed with HBSS three times to remove traces of phenol red contained in the cell culture media, and then incubated for an hour with regular cell culture media, but without Phenol Red, and supplemented with an oxygen radical scavenger, Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma-Aldrich) at 0.25 nM.

After an hour, cells are washed with HBSS, and gently detached using a cell scraper. Cell density is then adjusted by the help of a magnetic separator. Cells are then gently pipetted into the microfluidic device, and using a magnet, cells are pulled into the traps. We repeat this operation between three and five times, until reaching a good trapping efficiency (around 2/3 of occupied traps) and a single cell occupancy ratio of around 60%.

Cell Imaging

Cells are imaged on an Olympus© IX71™ microscope, equipped with an arc-mercury lamp (U-RX-T™) and a digital camera (Olympus© QImaging Retiga™ 6000). To image simultaneously multiple positions of the device, the microscope stage is replaced with a motorized stage (ASI MS-4400 XYZ Automated Stage). Images are captured with the software package Micro-Manager (extension of ImageJ), while the stage is programmed and controlled via a custom made script in Micro-Manager.

Stage Environment

Temperature and humidity are controlled using a homemade, on-stage system that keeps the cells at 37°C with 100% humidity. Cell media is supplemented with HEPES in order to limit the effects of the absence of CO_2 at 5%.

Computer codes

All the programs and scripts used for analysis were coded in Python 2.7 using the Machine Learning Scikit-Learn[28] and Pandas packages, while the motorized stage and camera setup was programmed in JavaScript.

References

1. Fuchs IB, Lichtenegger W, Buehler H, Henrich W, Stein H, et al. (2002) The prognostic significance of epithelial-mesenchymal transition in breast cancer. Anticancer Res 22: 3415–3419.

2. Mani SA, Guo W, Liao M-J, Eaton EN, Ayyanan A, et al. (2008) The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. Cell 133: 704–715. doi:10.1016/j.cell.2008.03.027.

3. Polyak K, Weinberg RA (2009) Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat Rev Cancer 9: 265–273. doi:10.1038/nrc2620.

4. Roca H, Hernandez J, Weidner S, McEachin RC, Fuller D, et al. (2013) Transcription Factors OVOL1 and OVOL2 Induce the Mesenchymal to Epithelial Transition in Human Cancer. PLoS ONE 8: e76773. doi:10.1371/journal.pone.0076773.

5. Thiery JP, Acloque H, Huang RYJ, Nieto MA (2009) Epithelial-Mesenchymal Transitions in Development and Disease. Cell 139: 871–890. doi:10.1016/j.cell.2009.11.007.

6. Tse JR, Engler AJ (2011) Stiffness Gradients Mimicking In Vivo Tissue Variation Regulate Mesenchymal Stem Cell Fate. PLoS ONE 6: e15978. doi:10.1371/journal.pone.0015978.

7. Wong IY, Javaid S, Wong EA, Perk S, Haber DA, et al. (2014) Collective and individual migration following the epithelial–mesenchymal transition. Nat Mater 13: 1063–1071. doi:10.1038/nmat4062.

8. Sanz-Moreno V, Gadea G, Ahn J, Paterson H, Marra P, et al. (2008) Rac Activation and Inactivation Control Plasticity of Tumor Cell Movement. Cell 135: 510–523. doi:10.1016/j.cell.2008.09.043.

9. Wolf K, Mazo I, Leung H, Engelke K, von Andrian UH, et al. (2003) Compensation mechanism in tumor cell migration: mesenchymal-amoeboid transition after blocking of pericellular proteolysis. J Cell Biol 160: 267–277. doi:10.1083/jcb.200209006.

10. Scanlon CS, Tubergen EAV, Inglehart RC, D'Silva NJ (2013) Biomarkers of Epithelial-Mesenchymal Transition in Squamous Cell Carcinoma. J Dent Res 92: 114–121. doi:10.1177/0022034512467352.

11. Fan X-J, Wan X-B, Huang Y, Cai H-M, Fu X-H, et al. (2012) Epithelial–mesenchymal transition biomarkers and support vector machine guided model in preoperatively

predicting regional lymph node metastasis for rectal cancer. Br J Cancer 106: 1735–1741. doi:10.1038/bjc.2012.82.

12. Gupta S, Tran T, Luo W, Phung D, Kennedy RL, et al. (2014) Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ Open 4: e004007. doi:10.1136/bmjopen-2013-004007.

13. Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, Ahmad LG (2013) Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inf 4: 124.

14. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, et al. (2013) Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS ONE 8: e61318. doi:10.1371/journal.pone.0061318.

15. Drier Y, Domany E (2011) Do Two Machine-Learning Based Prognostic Signatures for Breast Cancer Capture the Same Biological Processes? PLoS ONE 6: e17795. doi:10.1371/journal.pone.0017795.

16. Percio S, Coltella N, Grisanti S, Bernardi R, Pattini L (2014) A HIF-1 network reveals characteristics of epithelial-mesenchymal transition in acute promyelocytic leukemia. Genome Med 6: 84. doi:10.1186/s13073-014-0084-4.

17. Ramaswamy S, Ross KN, Lander ES, Golub TR (2003) A molecular signature of metastasis in primary solid tumors. Nat Genet 33: 49–54. doi:10.1038/ng1060.

18. Verdone JE, Parsana P, Veltri RW, Pienta KJ (2014) Epithelial–mesenchymal transition in prostate cancer is associated with quantifiable changes in nuclear structure. The Prostate: n/a–n/a. doi:10.1002/pros.22908.

19. Cruz JA, Wishart DS (2007) Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform 2: 59–77.

20. Park JY, Morgan M, Sachs AN, Samorezov J, Teller R, et al. (2009) Single cell trapping in larger microwells capable of supporting cell spreading and proliferation. Microfluid Nanofluidics 8: 263–268. doi:10.1007/s10404-009-0503-9.

21. Edelstein A, Amodaj N, Hoover K, Vale R, Stuurman N (2001) Computer Control of Microscopes Using µManager. Current Protocols in Molecular Biology. John Wiley & Sons, Inc. Available: http://onlinelibrary.wiley.com/doi/10.1002/0471142727.mb1420s92/abstract. Accessed 16 December 2014.

22. Ester M, Kriegel H, S J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise AAAI Press. pp. 226–231.

23. Bastian M, Heymann S, Jacomy M, others (2009) Gephi: an open source software for exploring and manipulating networks. ICWSM 8: 361–362.

24. Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Softw Pract Exp 21: 1129–1164. doi:10.1002/spe.4380211102.

25. Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28: 129–137. doi:10.1109/TIT.1982.1056489.

26. Forgy E (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21: 768–769.

27. Ankerst M, Breunig MM, Kriegel H, Sander J (1999) OPTICS: Ordering Points To Identify the Clustering Structure ACM Press. pp. 49–60.

28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12: 2825–2830.

Chapter 5 :
Summary and future directions

Summary

This thesis demonstrated the feasibility of a high-throughput, adaptable and cost effective system to analyze the morphological changes of single cancer cells in suspension. From the proof of concept of the magneto-rotation of single cells towards the possibility to classify, discriminate and identify cells of different metastatic potential, we have laid the basis for the conception of a fully automated device dedicated to assess the phenotypic heterogeneity of tumor cells by way of their morphological signatures.

In chapter 2, the principles of the CM method were exposed: magnetization of biological cells via the uptake of superparamagnetic nanoparticles and spinning motion of single cells by the application of a low intensity rotating magnetic field. With this protocol, HeLa cells are rendered magnetic, and we showed that under specific conditions (a rotating magnetic field with a frequency above 10Hz or so), cells behave similarly as magnetic beads. This chapter focuses on the ability to transform a biological cell into its very own bio-sensing actuator, free of any biomarker. For our method to be successfully used as a biosensor, we showed that it did not harm cells, be it regarding cell viability or even more constraining, cell division. We then successfully applied ourselves to test the biosensor as a reliable and fast way to detect minuscule morphological changes by the changes of the cell's rotation period. As such, chapter 1 proved that CM can be used as a method to perform live/dead cell assay, with earlier detection than with PI.

Chapter 3 introduces a breakthrough in the morphological analysis of single cells using the CM method. First, it validates the multiplexing of the method, by trapping cells in the individual wells of a microfluidic device. Then, while we still rotated the cells, from that point on, we relied on an object detection system to extract morphological features of single cells. Even though we reduced the magnification from 40X to 10X, we can accurately describe the shape of the cells. Because of the multiplication of cells observed simultaneously and the increase of measured features for each cell image (by two orders of magnitude each), chapter 3 exposed the

development of an innovative method to analyze and classify breast cancer cells MDA-MB-231, tagged with Green Fluorescent Protein (GFP). With the impossibility to use a classical statistical toolbox to extract information from the multidimensional tables that we treated, we relied on supervised Machine Learning (ML) classification algorithms to classify cells over long periods of time. We observed that cells could then be separated in between different classes, based on how and how often they would change shape over time. Finally, chapter 3 demonstrated how to use relationship graph networks to visualize and isolate cells that could potentially be at the crossroads between different phenotypes. This could be used as a tool to detect the minority of cells that could be targeted to reduce the malignancy of a cancer.

Finally, in chapter 4, we demonstrated the ability to distinguish between cancer cells of different metastatic potential by magnetically rotating them, extracting morphological features and training a computer to recognize and classify cells. To do so, we first modified the setup we used, adding a programmable motorized stage (in the three directions, X, Y and Z). Added to a high angle and high definition monochromatic camera, we managed to increase the number of cells observed by a factor of 20 compared to chapter 2 (where we already had a hundred fold increase in the multiplexing level). While we use the same method to extract the morphological features than in chapter 2, we introduced a different clustering method. Indeed, chapter 3 showed that using the unsupervised clustering algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) followed with Support Vector Machines (SVM) based ML algorithms, Prostate Carcinoma mesenchymal cells could be distinguished from their epithelial counterparts.

Future directions

There are several ways to go from the point we are at. Improvements and challenges that are ahead can be split between the technical and the more biological ones.

When it comes to the technical aspect of this method, the first aspect to improve would be the data treatment step in order to analyze images, cluster data points and extract information on the spot. This requires setting up a computer with high parallel computing power (say using GPU units) that will be used simultaneously, capturing cell images and treating the data on the spot. For the moment, it is not feasible to use distant computer clusters to do so, because of the

time it takes to transfer the data back and forth to a server. In addition, this solution would prove more flexible when it comes to adding new modules to the analytical process. In any case, this would significantly reduce the time spent in post-treatment, allowing one to get classification and heterogeneity results much faster and in real-time.

Following, the *in-situ* processing of the data opens up the possibility to recapture the cells of interest right after imaging, since they are still alive and individually tracked (which is not the case of other existing high-throughput technologies). The challenge here is to redesign the trapping device to allow for micro-pipetting, or use magnetic pulls to pick cells up (which would cause much less strain on the cells than suction). This way, with the progress in single cell DNA and RNA analysis techniques, we will be able to correlate morphological changes and behaviors with individual cells' genomes and protein expressions, and help identify the cells that are most critical to the diagnosis.

The extraction of morphological features could also be improved by tagging various parts of the cell with different fluorescent dyes/proteins. For instance, we could prepare the cells so that the actin filaments are tagged in green, the membrane in blue and the nucleus in red. This way, we can greatly refine the information we can extract from each cell, and improve our discrimination power. However, this requires to change parts of the imaging setup, using an LED source for fluorescence, with multiple filters. Such setup already exists, such as the X-Cite® (Excelitas Technologies®), but it is quite expensive and will make the technology less available to other laboratories looking into using the CM method for their own needs.

The second part of potential improvements concerns the biological use of the method described in this thesis. One of the limitations of our method is that even though it can be easily adapted and used by fellow researchers, it is still far from a bench-top device that could be used by oncologists with fresh cancer cells just extracted from a patient's tumor. It is well known that actual cancer cells are less resistant than their cell lines counterparts, and it is far more difficult to tag them, whether with fluorescent dyes or with magnetic particles. On the other hand, these cells present a much higher degree of heterogeneity than plated cell lines, and would be the ideal target population to study with our system. A way to get around this problem could be to use a laminar flow to rotate the cells (but this will not necessarily rotate the cells around their own

axis), and to use brightfield images. For the latter part, this would pose challenges to the object recognition and feature extraction parts, but once cell parts have been well characterized and correlated with the metastatic potential, it might be possible to only focus on specific domains of the cells (such as, say, the cytoskeleton or the nucleus).

Another challenge will be to correlate the way we map the heterogeneity of a cell population (by using graph analysis) with the evolution of patients' health during therapy. If successfully done, the CM method could become a very powerful tool to help design personalized therapeutic strategies for patients, strategies that limit the spread of cancer. To do this, we need to first test the CM method in xenografts (collecting cells grown in an animal), and then with cells directly from patients.

Finally, outside of the cancer scope, it could be interesting to use the CM method in order to study cell-cell interactions (by trapping two different cells in each well).

In conclusion, we presented a new method to track over time and analyze cellular morphology changes in populations of single cancer cells in suspension. Used in a multiplexed fashion and combined with powerful ML algorithms and relationship graph visualization and analysis, such a tool will improve the quality of diagnostics and prognostics for patients, and will allow a quick and accurate assessment of the evolution of the cancer spread and the relative success of a therapy. This will also reduce the number/dose of drugs used, and therefore help improve the quality of life of patients by reducing the side effects associated with many of the cancer fighting therapies available today.

Appendices

Appendix A

Features measured with CellProfiler

      CellProfiler [1–3] first delineates an object using an intensity threshold. Once the region of interest has been identified, several features are computed. These features can be split into three different families. First, purely geometrical features (i.e. area, perimeter, etc…), then radial distributions, and finally texture related features.

Area/shape features

      The *area*, as expected, is the total number of pixels inside the boundaries of the region of interest, while the perimeter is the number of pixels constituting the border of the region of interest.

      Form factor:

$$F = \frac{4\pi S}{P^2}$$

where S is the area and P is the perimeter of the region of interest.

      Extent is the proportion of pixels in the bounding box that are also in the region of interest.

      Euler number is the number of objects in the region of interest minus the number of holes in these objects.

      Orientation is the angle between the x-axis and the major axis length of the object.

      The *center coordinates* (CENTER_X, CENTER_Y) are the average of each coordinate in the object.

      Minor and major axis length are the minor and major axis length of the ellipse that has the same second moments as the region of interest. The second moments of an object are given by the covariance matrix of all the points in this same object.

      Eccentricity of the object is calculated as the eccentricity of the ellipse that has the same second moments (major axis length and minor axis length) as the object, and $e = \frac{c}{a}$ where c is the distance from the center to a focus of the ellipse and a is the distance from that focus to a vertex.

Compactness is the variance of the radial pixels of an object divided by the area.

The *Feret diameter* of an object is the distance between two parallel lines that are tangent to each side of an object. By rotating the lines, we get a set of diameters, and the minimum and maximum of these diameters are respectively called the *Minimum Feret diameter* and the *Maximum Feret diameter*.

Zernike features

To calculate the *Zernike features* of an object[4], the region of interest is mapped into a unit disc, centered on the center of the object. All the coordinates (x, y) of the points belonging to the object are then transformed in polar coordinates (r, θ), defined as follow:

$r = \sqrt{x^2 + y^2}$ and $\theta = \arctan(\frac{y}{x})$. The points for where r > 1 are excluded from the calculation of the Zernike moments. Namely, the Zernike moments are calculated on the largest disc fitting inside the object.

Zernike moments are defined as follow, with m and n being two integers, and I(x,y) the intensity of the pixel at the (x,y) coordinate. For $m - n$ even, we have:

$$A_{mn} = \frac{m + 1}{\pi} \int_x \int_y I(x, y) [Z_{mn}(x, y)]^* dx dy$$

Where

$$Z_{mn}(x, y) = R_{mn}(r) e^{jn\theta}$$

And

$$R_{mn}(r) = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k (n - k)!}{k! \left(\frac{m + n}{2} - k\right)! \left(\left(\frac{m - n}{2}\right) - k\right)!} r^{n-2k}$$

If $m - n$ is odd, then the polynomial function R_{mn} is zero. In our case, only the 10 first Zernike features are measured ($0 \leq n \leq 9$), rendering a total of 30 morphological features.

To measure the various *radial distributions*, the region of interest is split in concentric circular bins, the center of each bin being the center of the object. FractAtD is the fraction of total stain of the object within a given bin. *MeanFrac* is the mean fractional intensity at a given radius, and finally, *RadialCV* is the coefficient of variation of intensity within a ring.

116

Haralick features

The last set of features are called *Haralick features* [5]. Before we get into the detailed formulas of each of the features, let us introduce first the *gray-level co-occurrence matrix **G***, a square matrix of dimension *N* where *N* is the number of levels of gray. In our case, *N* is either 2^8 or 2^{12} (8 or 12bits images). The element G(i,j) is defined as the probability that a *pixel* of value i is adjacent to a *pixel* of value j. Pixels in diagonal directions are also considered to be adjacent pixels.

$$G = \begin{bmatrix} p(1,1) & \cdots & p(1,N) \\ \vdots & \ddots & \vdots \\ p(N,1) & \cdots & p(N,N) \end{bmatrix}$$

Given the coefficients of the co-occurrence matrix, we can then calculate the following features:

Angular second moment: $\sum_i \sum_j p(i,j)^2$

Contrast:
$$\sum_{n=0}^{N-1} n^2 \sum_{i=1}^{N} \sum_{j=1}^{N} p(i,j), |i-j| = n$$

Correlation:
$$\frac{\sum_i \sum_j ijp(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

where μ_x, μ_y, σ_x and σ_y are the means and standard deviations of p_x and p_y, the partial probability density functions.

Sum of squares: *Variance*: $\sum_i \sum_j (i-\mu)^2 p(i,j)$

Inverse difference moment: $\sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$

117

Sum average: $\sum_{i=2}^{2N} i p_{x+y}(i)$ where *x* and *y* are the coordinates (row and column) of an entry in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to *x+y*.

Sum variance:

$$\sum_{i=2}^{2N}(i - f_8)^2 p_{x+y}(i)$$

Where f_8 is the sum entropy:

$$f_8 = -\sum_{i=2}^{2N} p_{x+y}(i)\log(p_{x+y}(i))$$

Entropy:

$$-\sum_i \sum_j p(i,j)\log(p(i,j))$$

Difference Variance:

$$\sum_{i=0}^{N-1} p_{x-y}(i)\log(p_{x-y}(i))$$

Information Measure of Correlation 1:

$$\frac{HXY - HXY_1}{\max\{HX, HY\}}$$

Information Measure of Correlation 2:

$$(1 - \exp[-2(HXY_2 - HXY)])^{1/2}$$

Where $HXY = -\sum_i \sum_j p(i,j)\log(p(i,j))$, HX and HY are the entropies of p_x and p_y, $HXY_1 = -\sum_i \sum_j p(i,j)\log(p_x(i)p_y(j))$ and $HXY_2 = -\sum_i \sum_j p_x(i)p_y(j)\log(p_x(i)p_y(j))$

Complete list of measured parameters
Cell_AreaShape_Area
Cell_AreaShape_Center_X

Cell_AreaShape_Center_Y
Cell_AreaShape_Compactness
Cell_AreaShape_Eccentricity
Cell_AreaShape_EulerNumber
Cell_AreaShape_Extent
Cell_AreaShape_FormFactor
Cell_AreaShape_MajorAxisLength
Cell_AreaShape_MaxFeretDiameter
Cell_AreaShape_MaximumRadius
Cell_AreaShape_MeanRadius
Cell_AreaShape_MedianRadius
Cell_AreaShape_MinFeretDiameter
Cell_AreaShape_MinorAxisLength
Cell_AreaShape_Orientation
Cell_AreaShape_Perimeter
Cell_AreaShape_Solidity
Cell_AreaShape_Zernike_0_0
Cell_AreaShape_Zernike_1_1
Cell_AreaShape_Zernike_2_0
Cell_AreaShape_Zernike_2_2
Cell_AreaShape_Zernike_3_1
Cell_AreaShape_Zernike_3_3
Cell_AreaShape_Zernike_4_0
Cell_AreaShape_Zernike_4_2
Cell_AreaShape_Zernike_4_4
Cell_AreaShape_Zernike_5_1
Cell_AreaShape_Zernike_5_3
Cell_AreaShape_Zernike_5_5
Cell_AreaShape_Zernike_6_0
Cell_AreaShape_Zernike_6_2
Cell_AreaShape_Zernike_6_4
Cell_AreaShape_Zernike_6_6
Cell_AreaShape_Zernike_7_1
Cell_AreaShape_Zernike_7_3
Cell_AreaShape_Zernike_7_5
Cell_AreaShape_Zernike_7_7
Cell_AreaShape_Zernike_8_0
Cell_AreaShape_Zernike_8_2
Cell_AreaShape_Zernike_8_4
Cell_AreaShape_Zernike_8_6
Cell_AreaShape_Zernike_8_8
Cell_AreaShape_Zernike_9_1
Cell_AreaShape_Zernike_9_3
Cell_AreaShape_Zernike_9_5

Cell_AreaShape_Zernike_9_7
Cell_AreaShape_Zernike_9_9
Cell_RadialDistribution_FracAtD_OrigGray_1of8
Cell_RadialDistribution_FracAtD_OrigGray_2of8
Cell_RadialDistribution_FracAtD_OrigGray_3of8
Cell_RadialDistribution_FracAtD_OrigGray_4of8
Cell_RadialDistribution_FracAtD_OrigGray_5of8
Cell_RadialDistribution_FracAtD_OrigGray_6of8
Cell_RadialDistribution_FracAtD_OrigGray_7of8
Cell_RadialDistribution_FracAtD_OrigGray_8of8
Cell_RadialDistribution_MeanFrac_OrigGray_1of8
Cell_RadialDistribution_MeanFrac_OrigGray_2of8
Cell_RadialDistribution_MeanFrac_OrigGray_3of8
Cell_RadialDistribution_MeanFrac_OrigGray_4of8
Cell_RadialDistribution_MeanFrac_OrigGray_5of8
Cell_RadialDistribution_MeanFrac_OrigGray_6of8
Cell_RadialDistribution_MeanFrac_OrigGray_7of8
Cell_RadialDistribution_MeanFrac_OrigGray_8of8
Cell_RadialDistribution_RadialCV_OrigGray_1of8
Cell_RadialDistribution_RadialCV_OrigGray_2of8
Cell_RadialDistribution_RadialCV_OrigGray_3of8
Cell_RadialDistribution_RadialCV_OrigGray_4of8
Cell_RadialDistribution_RadialCV_OrigGray_5of8
Cell_RadialDistribution_RadialCV_OrigGray_6of8
Cell_RadialDistribution_RadialCV_OrigGray_7of8
Cell_RadialDistribution_RadialCV_OrigGray_8of8
Cell_Texture_AngularSecondMoment_OrigGray_3_0
Cell_Texture_Contrast_OrigGray_3_0
Cell_Texture_Correlation_OrigGray_3_0
Cell_Texture_DifferenceEntropy_OrigGray_3_0
Cell_Texture_DifferenceVariance_OrigGray_3_0
Cell_Texture_Entropy_OrigGray_3_0
Cell_Texture_Gabor_OrigGray_3
Cell_Texture_InfoMeas1_OrigGray_3_0
Cell_Texture_InfoMeas2_OrigGray_3_0
Cell_Texture_InverseDifferenceMoment_OrigGray_3_0
Cell_Texture_SumAverage_OrigGray_3_0
Cell_Texture_SumEntropy_OrigGray_3_0
Cell_Texture_SumVariance_OrigGray_3_0
Cell_Texture_Variance_OrigGray_3_0

References

1. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, et al. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 7: R100. doi:10.1186/gb-2006-7-10-r100.

2. Lamprecht MR, Sabatini DM, Carpenter AE, others (2007) CellProfiler^{TM}: free, versatile software for automated biological image analysis. Biotechniques 42: 71.

3. Kamentsky L, Jones TR, Fraser A, Bray M-A, Logan DJ, et al. (2011) Improved structure, function, and compatibility for CellProfiler: modular high-throughput image analysis software. Bioinformatics: btr095. doi:10.1093/bioinformatics/btr095.

4. Rocha L, Velho L, Carvalho PCP (2002) Image moments-based structuring and tracking of objects. XV Brazilian Symposium on Computer Graphics and Image Processing, 2002. Proceedings. pp. 99–105. doi:10.1109/SIBGRA.2002.1167130.

5. Haralick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67: 786–804.

DBSCAN Algorithm

The Density Based Spatial Clustering of Applications with Noise (DBSCAN)[1], is a density based clustering algorithm. To form a cluster around a point in space, we first have to set a *distance limit* (ε - epsilon) and a *density minPts*, the minimum number of points in the sphere of radius ε centered on the point of interest. If *P* and *Q* are two points in space, if P is surrounded by a sufficient number of points and if the distance between the two points P and Q is smaller than ε, then P and Q are considered to be part of the same cluster. Q is considered to be *directly density-reachable* from P. If there exist a sequel of point $\{P_i\}_{i=0..n}$ such that $P_0 = P$ and $P_n = Q$, and P_i and P_{i+1} are density reachable for any index *i* between 0 and n, then Q is said to be *density reachable* from P.

The notion of *density reachability* being asymmetrical (Q can be density-reachable from P but P might not be density reachable from Q, for instance in the case where Q lies on the edge of a cluster). However, if we define P and Q to be *density connected* if there exist a point O such that P and Q are both density reachable from O, then *density connectedness* is symmetric, and this is what is used to build the DBSCAN algorithm.

With ε and *minPts* defined, the algorithm starts at a random point that has not been visited yet. If the ε-neighborhood (the set of points at a distance smaller than ε) has a *cardinality* higher or equal to *minPts*, a cluster is started. Otherwise, the point is considered as *noise*.

If a cluster was started, the whole ε-neighborhood is considered part of the cluster, and all the points belonging to the ε-neighborhood are subsequently added to the cluster, as well as their own ε-neighborhood, if found to be dense enough. This process goes on until the entire the density-connected cluster has been found. The algorithm then starts at a new unvisited point and repeats   the   same   process.

Pseudo-code
The pseudo-code of the algorithm is the following:

```
DBSCAN(D, eps, MinPts)
   C = 0
   for each unvisited point P in dataset D
      mark P as visited
      NeighborPts = regionQuery(P, eps)
      if sizeof(NeighborPts) < MinPts
         mark P as NOISE
      else
         C = next cluster
```

```
expandCluster(P, NeighborPts, C, eps, MinPts)
```

Where expandCluster is defined as:

```
expandCluster(P, NeighborPts, C, eps, MinPts)
   add P to cluster C
   for each point P' in NeighborPts
      if P' is not visited
         mark P' as visited
         NeighborPts' = regionQuery(P', eps)
         if sizeof(NeighborPts') >= MinPts
            NeighborPts = NeighborPts joined with NeighborPts'
      if P' is not yet member of any cluster
         add P' to cluster C
```

```
regionQuery(P, eps)
```

   return all points within P's eps-neighborhood (including P)

References

1.   Ester M, Kriegel H, S J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise AAAI Press. pp. 226–231.

Appendix C

Python code for the DBSCAN Algorithm and the random shuffling of data points prior to training a classifier:

```
# -*- coding: utf-8 -*-

import numpy as np

from sklearn.cluster import DBSCAN
from sklearn import metrics
from sklearn.datasets.samples_generator import make_blobs
from sklearn.preprocessing import StandardScaler
import pandas as pd
from sklearn import cluster, covariance, manifold
from sklearn import decomposition
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import psutil
import math
import time
import random

from sklearn import svm
from sklearn.externals import joblib
from sklearn.cross_validation import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_moons, make_circles, make_classification
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.lda import LDA
from sklearn.qda import QDA

from sklearn.externals import joblib
from sklearn.metrics import classification_report

## Tarining set ratio
```

```
n = 0.7
time_limit = 60

## Parameters for DBSCAN

epsilon = 2.7
minPoints = 95

### Create a function that randomly picks an element from a list or an array

def random_element_iterator( iterator ):
   N = 0

   for item in iterator:
      N += 1
      # 1/N chance
      if random.random() * N < 1:
         element = item

   return element

### This function returns a list of randomly selected elements from a list

def random_subset( iterator, K ):
   result = []
   N = 0

   for item in iterator:
      N += 1
      if len( result ) < K:
         result.append( item )
      else:
         s = int(random.random() * N)
         if s < K:
            result[ s ] = item

   return result

### This function takes into argument the predited cluster of data points and
### returns the corresponding predicted phenotype

def getPheno(prediction, clus2pheno):
   pheno_predict = []
   for i in range(len(prediction)):
```

```python
        pheno_predict.append(clus2pheno[prediction[i]])
    return pheno_predict


### This function takes as argument the data points, the cell names and the
### ratio of training set and returns a randomly shuffled training set and
### testing set, with the respective lables for the phenotypes

def shuffle_cells(df_data, df_names, ratio):
    #index = [x for x in range(len(list_names))]
    list_names = list(set(df_names['Image_Metadata_Cell']))
    print "full ", len(df_data)
    selected = random_subset(list_names, ratio*len(list_names))
    train_set = selected
    test_set = list(set(list_names)-set(selected))

    criterion_train = df_names['Image_Metadata_Cell'].map(lambda x: x in train_set)
    criterion_test = df_names['Image_Metadata_Cell'].map(lambda x: x in test_set)
    df_names_train = df_names[criterion_train]
    df_names_test = df_names[criterion_test]

    index_train = df_names_train.index.values
    index_test = df_names_test.index.values
    df_data_train = df_data.loc[index_train]
    print "train ", len(df_data_train)
    df_data_test = df_data.loc[index_test]
    print "test ", len(df_data_test)
    for feature in ['Image_Metadata_Time','Image_Metadata_Cell', 'Image_Count_Cell']:
        df_data_train = df_data_train.drop(feature, 1)
        df_data_test = df_data_test.drop(feature, 1)

    return df_names_train, df_data_train, df_names_test, df_data_test

def shuffle(df_data, df_names, ratio):
    list_names = list(set(df_names['Image_Metadata_Cell']))
    selected = random_subset(list_names, ratio*len(list_names))
    train_set = selected
    test_set = list(set(list_names)-set(selected))

    df_train = df_data.map((lambda x: x in train_set))

    return train_set, test_set
```

This function simply splits the data into a training set and testing set,
without shuffling the cells.

```python
def classic_split(df_data, df_names, ratio):
    list_names = list(set(df_names['Image_Metadata_Cell']))
    print "full ", len(df_data)
    #selected = random_subset(list_names, ratio*len(list_names))
    selected = list_names[:int(ratio*len(list_names))]
    train_set = selected
    test_set = list(set(list_names)-set(selected))

    criterion_train = df_names['Image_Metadata_Cell'].map(lambda x: x in train_set)
    criterion_test = df_names['Image_Metadata_Cell'].map(lambda x: x in test_set)
    df_names_train = df_names[criterion_train]
    df_names_test = df_names[criterion_test]

    index_train = df_names_train.index.values
    index_test = df_names_test.index.values
    df_data_train = df_data.loc[index_train]
    print "train ", len(df_data_train)
    df_data_test = df_data.loc[index_test]
    print "test ", len(df_data_test)
    for feature in ['Image_Metadata_Time','Image_Metadata_Cell', 'Image_Count_Cell']:
        df_data_train = df_data_train.drop(feature, 1)
        df_data_test = df_data_test.drop(feature, 1)

    return df_names_train, df_data_train, df_names_test, df_data_test


def visualize_clusters(X, n_clusters_, core_samples_mask, labels, outliers = False):
    ### To getter a better understanding of interaction of the dimensions
    # plot the first three PCA dimensions
    fig = plt.figure(1, figsize=(8, 6))
    ax = Axes3D(fig, elev=-150, azim=110)
    ##
    ###np.savetxt("X_reduced.csv", X[class_member_mask & core_samples_mask], delimiter =
',')
    ##

    ##ratio of points to display
    m = 3
    unique_labels = set(labels)
    colors = plt.cm.Spectral(np.linspace(0, 1, len(unique_labels)))
    for k, col in zip(unique_labels, colors):
        if k == -1:
            # Black used for noise.
```

```
        col = 'k'
    class_member_mask = (labels == k)
    X_reduced = X_train[class_member_mask & core_samples_mask]
    ax.scatter(X_reduced[::m, 0], X_reduced[::m, 1], X_reduced[::m, 2], c=col,
            cmap=plt.cm.Paired)
    if (outliers):
        X_reduced = X_train[class_member_mask & ~core_samples_mask]
        ax.scatter(X_reduced[::m, 0], X_reduced[::m, 1], X_reduced[::m, 2], c=col,
                cmap=plt.cm.Paired)
    #ax.set_title("First three PCA directions")
    ax.set_xlabel("1st eigenvector")
    ax.w_xaxis.set_ticklabels([])
    ax.set_ylabel("2nd eigenvector")
    ax.w_yaxis.set_ticklabels([])
    ax.set_zlabel("3rd eigenvector")
    ax.w_zaxis.set_ticklabels([])



    plt.show()

### This function implements the DBSCAN algorithm, print the clustering metrics
### such as homogeneity and completeness, and returns the labels, the number of
### clusters, the homogeneity and the data points in each clusters

def dbscan_analysis(epsilon, minPoints, X, labels_true):
    print("\n")
    print("####################################")
    print "Epsilon: " + str(epsilon) +'\t minPoint:' + str(minPoints)
    db = DBSCAN(eps=epsilon, min_samples=minPoints).fit(X)
    core_samples = db.core_sample_indices_
    core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
    core_samples_mask[db.core_sample_indices_] = True
#   np.savetxt("db_labels_n_%s_t_%s.csv" % (n, time_limit), db.labels_, delimiter = ",")
#   np.savetxt("core_samples_n_%s_t_%s.csv" % (n, time_limit), core_samples, delimiter = ',')
#   np.savetxt("core_samples_mask_n_%s_t_%s.csv" % (n, time_limit), core_samples_mask,
delimiter = ',')
    labels = db.labels_

    # Number of clusters in labels, ignoring noise if present.
    n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
    #

#      ##print core_samples
#
```

```python
        print('Estimated number of clusters: %d' % n_clusters_)
        print("Homogeneity: %0.3f" % metrics.homogeneity_score(labels_true, labels))
        print("Completeness: %0.3f" % metrics.completeness_score(labels_true, labels))
        print("V-measure: %0.3f" % metrics.v_measure_score(labels_true, labels))
        print("Adjusted Rand Index: %0.3f"
            % metrics.adjusted_rand_score(labels_true, labels))
        print("Adjusted Mutual Information: %0.3f"
            % metrics.adjusted_mutual_info_score(labels_true, labels))
        ##print proc, mem, swap
        ##print("Silhouette Coefficient: %0.3f"
            ##% metrics.silhouette_score(X, labels))
        lab_true =[]
        lab_pred = []
        for x in range(len(labels)):
            if labels[x] > -1:
                lab_true.append(labels_true_train[x])
                lab_pred.append(labels[x])

        outliers_p = []
        for x in range(len(pc3_train)):
            if labels[x] == -1:
                outliers_p.append(x)

        print "homogeneity test"
        homo = metrics.homogeneity_score(lab_true, lab_pred)
        print homo
        print "without outliers" + str(metrics.homogeneity_score(labels_true_train, labels))

        n_outliers = len(labels) - len(lab_true)

        print len(lab_true)
        print len(labels)
        print "number of outliers = " + str(len(labels) - len(lab_true))
        print "PC3E outliers ratio = " + str(float(len(outliers_p))/len(pc3_train))
        print "HR-14 outliers ratio = " + str(float(n_outliers-len(outliers_p))/len(hr14_train))
        return n_clusters_, labels, core_samples_mask, homo

def cluster_analysis(n_clusters_, labels_true, labels):
    clusters = []
    labels_to_split = []
    for k in range(-1, n_clusters_):
        population = labels_true[[labels == k]]
        clusters.append(population)
        p = float(sum(population))/len(population)
```

130

```python
        if (0.1< p < 0.9):
            labels_to_split.append(k)
        print "Cluster " +str(k) + " population: " + str(len(population)) + " ratio: " + str(p)
    return clusters, labels_to_split


def convert_class_report(class_report, name, parameters):
    titles = class_report.split("\n")[0].split()
    titles.insert(0, "phenotype")
    titles.append("parameters")
    class_p = class_report.split("\n")[2].split()
    class_p.append(parameters)
    class_h = class_report.split("\n")[3].split()
    class_h.append(parameters)
    class_total = class_report.split("\n")[5].split()
    class_total = class_total[2:]
    class_total.append(parameters)
    return pd.DataFrame([class_p, class_h, class_total], columns = titles)



##Load phenotypes data separately

name_h = pd.DataFrame.from_csv('name_list_hr14.csv', index_col = 'ImageNumber')
name_p = pd.DataFrame.from_csv('name_list_pc3.csv', index_col = 'ImageNumber')
name_p2 = pd.DataFrame.from_csv('name_list_pc23.csv', index_col = 'ImageNumber')

name_p['Time'] = name_p['Time']-min(name_p['Time'])
name_p2['Time'] = name_p2['Time']-min(name_p2['Time'])

name_h = name_h[(name_h['Time']<time_limit) & (name_h['Image_Count_Cell'] == 1)]
name_p = name_p[(name_p['Time']<time_limit) & (name_p['Image_Count_Cell'] == 1)]
name_p2 = name_p2[(name_p2['Time']<time_limit) & (name_p2['Image_Count_Cell'] == 1)]

### Load data points
ddf_h = pd.DataFrame.from_csv('data_image_all_hr14.csv', index_col = 'ImageNumber')
ddf_p = pd.DataFrame.from_csv('data_image_all_pc3.csv', index_col = 'ImageNumber')
ddf_p2 = pd.DataFrame.from_csv('data_image_all_pc23.csv', index_col = 'ImageNumber')

print "loaded"

### Form an index using the common indices of names list and data points

index_h = list(set(list(name_h.index.values)) & set(list(ddf_h.index.values)))
index_p = list(set(list(name_p.index.values)) & set(list(ddf_p.index.values)))
index_p2 = list(set(list(name_p2.index.values)) & set(list(ddf_p2.index.values)))
```

131

```
### To make sure we do not have doublons, we remove points where two objects are
measured
### They have the same name

ind_h = []
ind_p = []
ind_p2 = []

for i in index_h:
    if len(ddf_h.ix[i]) ==113:
        ind_h.append(i)

for i in index_p:
    if len(ddf_p.ix[i]) ==113:
        ind_p.append(i)

for i in index_p2:
    if len(ddf_p2.ix[i]) ==113:
        ind_p2.append(i)




df_h = ddf_h[(ddf_h['Image_Count_Cell']==1)&(ddf_h['Image_Metadata_Time']<time_limit)]
df_p = ddf_p[(ddf_p['Image_Count_Cell']==1)&(ddf_p['Image_Metadata_Time']<time_limit)]
df_p2 =
ddf_p2[(ddf_p2['Image_Count_Cell']==1)&(ddf_p2['Image_Metadata_Time']<time_limit)]

labels_true_p = [0 for x in range(len(df_p)+len(df_p2))]
labels_true_h = [1 for x in range(len(df_h))]
labels_true = np.concatenate((labels_true_p, labels_true_h))

print "df_p"
print len(df_p)

print "df_p2 "+ str(len(df_p2))



df_full = pd.concat((df_p, df_p2, df_h))
df_full_names = pd.DataFrame(index = df_full.index)
df_full_names['Image_Metadata_Cell'] = pd.Series(df_full['Image_Metadata_Cell'])
df_full_names['Image_Metadata_Time'] = pd.Series(df_full['Image_Metadata_Time'])
df_full_names.to_csv("df_full_names_t_%s.csv" % time_limit)
```

```python
print "ready"

epi1_names = df_full_names[df_full_names['Image_Metadata_Cell'].str.contains("102614")]
epi2_names = df_full_names[df_full_names['Image_Metadata_Cell'].str.contains("102314")]
mes_names = df_full_names[df_full_names['Image_Metadata_Cell'].str.contains("H")]


def main():


###############################################################################
    # Compute DBSCAN

    print "start DBSCAN"
    start = time.time()
    n_clusters_, labels, core_samples_mask = dbscan_analysis(epsilon, minPoints, X_train,
labels_true_train)

    end = time.time()

    print end - start
    #visualize_clusters(X_train,n_clusters_, core_samples_mask, labels)
    clusters, labels_to_split = cluster_analysis(n_clusters_, labels_true_train, labels)




if __name__ == '__main__':


    report_list = []
    homogeneity_list = []
    cluster_num = []
    ### Loop the algorithm to average the results

    for k in range(20):

        epi1_names_train, epi1_data_train, epi1_names_test, epi1_data_test = shuffle_cells(df_p,
epi1_names, n)
        epi2_names_train, epi2_data_train, epi2_names_test, epi2_data_test =
shuffle_cells(df_p2, epi2_names, n)
        mes_names_train, mes_data_train, mes_names_test, mes_data_test = shuffle_cells(df_h,
mes_names, n)
```

```python
    #epi1_names_train, epi1_data_train, epi1_names_test, epi1_data_test = classic_split(df_p,
epi1_names, n)
    #epi2_names_train, epi2_data_train, epi2_names_test, epi2_data_test =
classic_split(df_p2, epi2_names, n)
    #mes_names_train, mes_data_train, mes_names_test, mes_data_test = classic_split(df_h,
mes_names, n)

    X_H_train = mes_data_train.as_matrix()
    X_P_train = epi1_data_train.as_matrix()
    X_P2_train = epi2_data_train.as_matrix()

    scaler_p = StandardScaler()
    scaler_h = StandardScaler()

    X_PC_train = np.concatenate((X_P_train, X_P2_train))

    scaler = StandardScaler()

    X_train = np.concatenate((X_PC_train, X_H_train))

    pc3_train = [0 for x in range(len(X_PC_train))]
    hr14_train = [1 for x in range(len(X_H_train))]
    labels_true_train = np.concatenate((np.array(pc3_train), np.array(hr14_train)))

    np.savetxt("labels_true_train_n_%s_t_%s.csv" % (n, time_limit), labels_true_train,
delimiter = ',')

    #### Prepare testing set

    X_H_test = mes_data_test.as_matrix()
    X_P_test = epi1_data_test.as_matrix()
    X_P2_test = epi2_data_test.as_matrix()

    X_PC_test = np.concatenate((X_P_test, X_P2_test))

    X_test = np.concatenate((X_PC_test, X_H_test))



    pc3_test = [0 for x in range(len(X_PC_test))]
    hr14_test = [1 for x in range(len(X_H_test))]
    labels_true_test = np.concatenate((np.array(pc3_test), np.array(hr14_test)))
```

```python
    np.savetxt("labels_true_test_n_%s_t_%s.csv" % (n, time_limit), labels_true_test, delimiter
= ',')

    #### Standardization and PCA
    ##Perform PCA
    pca = decomposition.PCA(n_components=14)
    #X_train = StandardScaler().fit_transform(X_train)

    X = np.concatenate((X_train, X_test))
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.fit_transform(X_test)

    pca.fit(X_train)
    X_train = pca.transform(X_train)
    X_test = pca.transform(X_test)

    X_train = np.array(X_train)

    print pca.explained_variance_ratio_

    X_test = np.array(X_test)


  ##X_train = X[:len(X_train)]
  ##X_test = X[len(X_train):]

  #np.savetxt("X_train_n_%s_t_%s.csv" % (n,time_limit) , X_train, delimiter = ',')
  #np.savetxt("X_test_n_%s_t_%s.csv" % (n,time_limit) , X_test, delimiter = ',')


    print "start DBSCAN"
    start = time.time()
    n_clusters_, labels, core_samples_mask, homo = dbscan_analysis(epsilon, minPoints,
X_train, labels_true_train)
    cluster_num.append(n_clusters_)
    homogeneity_list.append(homo)
    end = time.time()

    print end - start
    #visualize_clusters(X_train,n_clusters_, core_samples_mask, labels)
    clusters, labels_to_split = cluster_analysis(n_clusters_, labels_true_train, labels)

### Store the results in files
```

```python
np.savetxt("cluster_num.csv", np.array(cluster_num), delimiter = ',')
np.savetxt("homogeneity.csv", np.array(homogeneity_list), delimiter = ',')
```

Appendix D

Classification results by phenotype.

	precision	recall	f1-score	support
adaboost	0.765	0.860667	0.798	22503.17
dec_tree	0.837	0.909333	0.872	22491.17
knn	0.868667	0.913333	0.888667	22518.9
lda	0.922667	0.944	0.932333	22502
naive_bayes	0.89	0.91	0.9	22520
poly	0.83	0.941333	0.880667	22493.8
qda	0.901667	0.833333	0.864333	22501.97
randomForrest	0.847333	0.965333	0.902	22529.23
svc_10	0.831	0.9375	0.881	22500.75

Table D-1: Classification results for epithelial cells

	precision	recall	f1-score	support
adaboost	0.807	0.621	0.670333	17011.97
dec_tree	0.867667	0.766333	0.814	17030.8
knn	0.877667	0.817667	0.845	16992.93
lda	0.923667	0.894667	0.908	17017
naive_bayes	0.88	0.85	0.87	17011
poly	0.908667	0.746	0.818667	17165.27
qda	0.807	0.879333	0.838333	16938.63
randomForrest	0.943667	0.767667	0.845333	16995.73
svc_10	0.9025	0.7455	0.817	16949.9

Table D-1: Classification results for mesenchymal cells.