Label-free Detection of Cell-cycling Polyploid Cells in Osteosarcoma

A thesis submitted in fulfilment of the partial requirement for the degree of Doctor of Philosophy

by

Basmah Abdullah Almagwashi

School of Physics and Astronomy - Cardiff University

UK

June 2022

Abstract

The current work investigates the prospect of detecting live cell-cycling polyploid cells (in osteosarcoma through endocycling) based on their label-free scatter and images. Human polyploid cells are cells that carry multiples of their diploid DNA content. In many types of tumours, this DNA irregularity makes these cells one of the triggers behind genome instability, particularly when such cells commit to a ploidy cell cycle. Studies conducted on viable polyploid cells are often limited by classic fluorescent stoichiometric DNA labels for detection. These labels are only appropriate for end-point assays; thus prompts the need for label-free detection methods. In the current work, polyploidy in osteosarcoma populations has been induced via drug treatment with the reversible topoisomerase inhibitor ICRF-193, where the generation of a rare endoreduplicating population can be enriched through different drug treatment doses and regimen. Different flow cytometer platforms are then used to detect and measure this cell cohort using label-free parameters of interest: the forward and side scatter on conventional flow cytometry (CFC), and brightfield and darkfield images on imaging flow cytometry (IFC). The current work showed that both the forward and side scatter intensity (area) measurements can be used in the detection of the target polyploid cells with low diploid contamination (approximately 0.4% of the original diploidy population) through automated clustering analysis on the data collected on the CFC platforms. For label-free image data collected on IFC, six supervised machine learning models were able to classify up to 75% of the original target cells accurately, where as low as 0.3% of the diploidy population were mislabelled as target cells. The work also recommended the detection specifications of the target cells, as well as protocols for sample handling, sample controls, and data analysis methods to facilitate the reproducibility of the findings on any flow cytometric platform.

Acknowledgements

I thank God for giving me the energy to complete this work, and for blessing me with all the amazing individuals who have supported me throughout this very long journey.

First, I would like to thank my supervisors, Peter Smowton and Rachel Errington, for their continuous support since the very start until the very end. Their helpful feedback, guidance, and intriguing discussions have made my journey both possible and fun. Thank you so much.

I would like to also thank all my colleagues in Cardiff University from the school of physics and astronomy, and the Tissue Microenvironment Group (TMEG) in the school of medicine, for all the helpful discussions and the interesting science they shared. Special thanks to Marie Wiltshire from TMEG, whose generous assistance and valuable advice made it possible for me to handle many of the challenges.

I would like to acknowledge my sponsor in Saudi Arabia, King Abdulaziz University (KAU), for making my PhD studies in Cardiff University possible. I am especially grateful to my colleagues in the physics department in KAU for their continued support and understanding. I extend my gratitude to the staff in the Saudi Arabian Cultural Bureau in UK, and the people of Cardiff, for their friendliness and kindness.

To all of my dear friends, the old and the new, and the precious twin stars, I can not thank you enough for always cheering me up, and bringing me smiles and joy.

Lastly, to my dear and most precious family. Thank you very much. I am afraid my words are too lacking to express how important your presence is to me. Mom, Dad, Haya, Mimi, Ali, Al-Hasan, Sumayah, Sawsan, Bushra, Abdulrahman, Mohammad and Tota, the precious and most lovely Tota, thank you all for being part of my life. Thank you for your unconditional love and kindness. From the bottom of my heart, thank you so much for being here with me throughout the fun and difficult times. I pray that you all be blessed with great health and happy times full of dreams coming true, and furry little ones who do not bring you allergies. Cats are happiness after all.

Contents

Contents

List of Figures

List of Tables

List of Abbreviations

24T48R	24-hour drug treated cell sample with 48-hour recovery
48T24R	48-hour drug treated cell sample with 24-hour recovery
A	area
ADC	analogue to digital convertor
APD	avalanche photodiode
BF	brightfield
BP	bandpass
CCD	charged coupled device
CFC	conventional flow cytometry/flow cytometer
CIF	compensated image file (a file format)
CON	drug-free control cell sample
Cyclin B1-eGFP	enhanced green fluorescent protein fused to cyclin B1 protein
CV	coefficient of variance
DAFi	directed automated filtering and identification of cell populations
DF	darkfield
DNA	deoxyribonucleic acid
DRAQ5	1,5-bis (2-(di-methylamino)ethylamino)-4,8-dihydroxyanthracene-9,10-dione
FC	flow cytometry/flow cytometer
FCS	flow cytometry standard (a file format)
FITC	fluorescein isothiocyanate
FL	fluorescence/fluorescent
FSC	forward scatter channel
G0	cells at the resting phase of the cell cycle
G0/G1	cells at the resting or growth/gap 1 phases of the cell cycle
G1	cells at the growth/gap 1 phase of the cell cycle
G2	cells at the growth/gap 2 phase of the cell cycle
G2/M	cells at the growth/gap 2 or mitotic phases of the cell cycle
G2/M+ polyploid	cells at the growth/gap 2 or mitotic phases of the polyploid cell cycle or committed to a second polyploid cell cycle
GFP	green fluorescent protein
H	height
ICRF-193	meso-4,4'-(3,2-Butanediyl)-bis(2,6-piperazinedione)
LED	light-emitting diode
IFC	imaging flow cytometry/flow cytometer
K	(after numbers) thousand
LASER	light amplification by stimulated emission of radiation
Lin	linear
Log	logarithmic
LP	longpass
M	cells at the mitotic phase of the cell cycle
ML	machine learning
mW	milliwatts
nm	nanometre
p-value	calculated probability value
PerCP	peridinin-chlorophyll-protein

PMT	photomultiplier tube
QC	quality control
RCV	robust coefficient of variance
RIF	raw image file (a file format)
RSD	robust standard deviation
RT	room temperature
S	cells at the DNA synthesis phase of the cell cycle
SD	standard deviation
SSC	side scatter channel
SQL	structured query language
TIFF	tag image file format (a file format)
U-2 OS	human bone osteosarcoma epithelial cell line
W	width
μg	microgram
μL	microliter
μm	micrometre
μM	micromolar

Chapter 1 :

Introduction

1.1   Introduction

This thesis investigates non-invasive detection methods of cell-cycling polyploid cells in human osteosarcoma (bone cancer) via flow cytometry. Polyploid human cells are cells that possess multiples of their normal diploid (i.e. two set of chromosomes or 2N) DNA set, and their presence can promote cancer relapses. Flow cytometry is a high throughput technique for single cell analysis via laser light scatter and fluorescence emission. The label-free parameters it measures on its conventional instruments are the small angle forward scatter and the large angle (perpendicular to the forward direction) side scatter, as well as the brightfield and darkfield images of the cells on imaging flow cytometry. In this chapter, we first present the research motivation, where we define polyploid cells and discuss their impact on cancer development. Next, the research problem section discusses the current limitations in polyploid cell detection, where the intended research approach is also presented. This is then followed by a brief background on light scatter and the potential scatter sources found in cells. Next, the research hypotheses and objectives are presented. Lastly, the chapter concludes by presenting the overall structure for the rest of the thesis chapters.

1.2   Research motivation

The genomic heterogeneity observed in human osteosarcoma, and in fact many cancers, is one of the obstacles affecting its treatment advances [1]–[3]. The presence of cancer cells with highly irregular DNA content, such as polyploid cells, is considered one of the factors that promote further genome complexity. Polyploidy can occur naturally in mammalian cells for different purposes such as organ development, as well as to form cell barriers between the embryo and the mother's blood supply [4], [5]. When they are observed in tumours, however, these cells can show characteristics that contribute to cancer treatment evasion as well as tumor development and relapses, through delayed growth or active irregular cell cycling [4], [6], [7]. Specifically, polyploid cells that show further resistance, via continuing the cell cycle through a ploidy mitotic phase, are suspected of increasing genome instability in cancers [1], [8]. Therefore, this work will examine and target polyploid cells in osteosarcoma cell line U-2 OS (ATCC HTB-96), which is a human cell line commonly used to understand chromosomal instability due to its high aneuploidy (irregular chromosome count) rate.

For example, osteosarcoma U2-OS cell line has been observed to mostly resist cell death via developing a polyploid phenotype when exposed in vitro to ICRF-193 [9], a drug that reversely inhibits DNA Topoisomerase II (thus not allowing cells to divide or progress from the second growth [G2] phase to the mitotic [M] phase during cell cycle). Investigating the long-term impact of drug-resistant polyploid cells within tissues may not be easily feasible with the current standard detection methods for high DNA content in live cells, as will be discussed in the next section.

1.3   Research problem and approach

Studies on drug-induced alterations in cell cycle distributions, such as the generation of polyploidy cohorts, employ stoichiometric DNA binding fluorescent labels, which enables the quantification of DNA at the single level. However, such methods are only appropriate for end-point assays contrary to long term continuous studies. For long experiments on live cells, fluorescence labelling choices can be limited when it comes to colour compatibility with other labels emitting in the visible light, or due to a potential phototoxicity [10] or cytotoxicity [11]. This extends to recently developed DNA labels that have not been researched enough to assess their impact on cells, e.g. some have been found to introduce degrees of cytotoxicity when they first claimed none [12], [13]. Overall, it is believed that any labels designed for DNA quantification could inevitably negatively impact the cells [13]. Consequently, a label-free method is highly desirable to distinguish polyploid cells from diploid cells, especially in live cell sorting applications intended for long term culture. Moreover, such methods would need to be accessible to different flow cytometric platforms the way common DNA fluorescent labels are.

When it comes to label-free parameters, there is an observed tendency of polyploid cells to be larger than their diploid counterparts [1], [7], but it is yet to be assessed for cell detection or sorting. Moreover, evidence suggests that the internal structures of polyploid cells could be different from diploid cells beyond the possible cell content duplication [1]. General light scatter of cells (reviewed in the next section) could provide insight on such potential differences between the two cell classes, including size. Therefore, this work aims to experimentally investigate the prospect of identifying the drug resistant and actively cell cycling polyploid cells in osteosarcoma,

i.e. cells at the G2/M phase of the polyploidy cell cycle, through label-free cell scatter and imaging. This includes estimating the potential of said parameters in terms of cell purity or yield for cell sorting applications. For this purpose, a high throughput single cell analysis technique is employed, where different benchtop flow cytometers will be used for the cell characterisation and to assess the reproducibility of the findings.

1.4   Background and literature review

In the following subsections, we present a general background of light scatter in matter, and a review of previous works that had investigated light scatter sources in cells.

1.4.1   Light scatter in matter

Light travelling through particles within a medium undergoes extinction via absorption or scatter. This process is affected by the main elements in the scatter process, which are the incident light (its frequency and polarisation), the particles causing the scatter (their numbers, size, shape, composition, and orientation), and the medium around them [14].

When electromagnetic waves interact with matter, the electric field oscillates the electric charges within it in the same frequency as the incident wave. The original radiation is quenched, and secondary radiations of the same incident wavelength are generated from the resulting dipole moments to make what we observe as the scattered light. Any energy alterations introduced are regarded as an absorption. In the case of scatter by single particles, one could describe said particles as consisting of small segments, e.g. molecules or atoms, whose oscillations result in secondary emissions. These emissions superpose to produce the scattered light, which varies in intensity from one direction to another [14], [15]. The variations come from the phase shift introduced by the different positions of the small segments relative to each other and the overall shape of the particle. The superpositions of all the phase-shifted waves produce the direction-dependent maxima and minima commonly seen in scatter profile graphs. These phase shifts are reduced when the size of the particle becomes very small relative to the incident wavelength, producing an isotropic scatter profile. On the other

hand, they become apparent when the particle is comparatively larger, producing a varying scatter profile with constructive superpositions mainly towards the forward scatter as the size increases. In the case of scatter by many particles, each one would become a source of incident radiation to the nearby particles beside the original radiation, so in scatter problems they are treated as coupled particles. However, if their number density is very small and they are spaced far enough, the particles are treated as individual scatter sources, and the overall scattered light is the summation of their intensities [14], [15].

In light scatter problems, indirect problems are those where one attempts to identify the shape and characteristics of a particle based on its scatter. This type of problems is more difficult to solve than forward problems, which approach scatter analysis the other way around (i.e. observing the resulting light scatter of a known source). This is due to the amount of missing information or the physically challenging data to be collected such as the amplitude and phase of the scattered light [14]. Therefore, simulations of cell scatter often assume models of possible relevant scatter sources in an object, while analysis of experimental scatter results is often compared to such simulations or overlapped with the positions of suspected features in the respective darkfield images, as will be reviewed in the next section.

1.4.2   Light scatter in cells
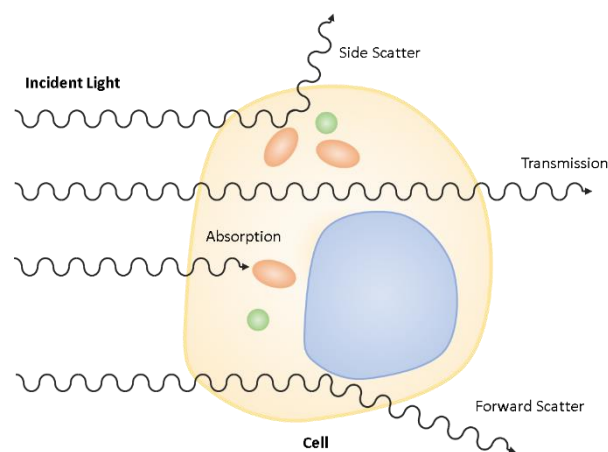
Figure 1-1: incident light on a cell undergoing extension via scatter or absorption

In flow cytometry experiments (see Chapter 2), small-angle forward scatter is perceived to correlate to cell size while the large-angle side scatter correlates to cell granularity (see Figure 1-1). For cells with diameters of 3-30 µm, the detected side scatter intensity is observed to be lower than that of the forward (with a factor of approximately 10^3) [16]. Different cell features or organelles have been explored in literature as potential light scattering centres, such as the cell membrane, cytosol, nucleus, mitochondria, and lysosomes [17]–[20]. This was performed on isolated organelles or within the cell, and conclusions were made from zero/one-dimensional scatter measurements or two-dimensional scatter images, which are then compared to simulated ones or to the positions of the investigated organelles' fluorescence [17]–[21]. The forward scatter of modelled live and fixed (i.e. preserved but dead) cells was shown to be affected by the nuclear-to-cytoplasmic diameter ratios (N/C) of the cells only when said cells were larger than approximately 14 µm [16]. On the other hand, the same work showed that the side scatter direction was more sensitive than forward scatter detecting changes in the relative refractive index of N/C rather than changes in its size ratio. In a study on suspensions of isolated nuclei and mitochondria, the first were observed to scatter more towards the forward direction while the latter towards the side direction. In addition, the experiments concluded that size or shape of cells contributes less towards their side scatter [17]. This was also observed in another work [22], where the forward scatter was affected by changes in the relative refractive index between the cell's cytosol and the surrounding medium. The same work had also experimented on suspensions of isolated nuclei. A positive correlation was observed between the nuclear side scatter and DNA content (or other possible nuclear structures) compared to simple increase in the size of the nuclei (which were modelled as homogenous spheres). In a recent work, a supervised machine-learning technique was used to classify the mitotic cell cycle phases for Jurkat cells using brightfield (which consists of transmitted light and some forward scatter) and darkfield (side scatter) images [23]. The work used Random Forest and Gradient Boosting algorithms in the classification, where Gradient Boosting, which performed better, ranked the granularity of the overall darkfield images as a top classifying feature for the different mitotic cell phases (different DNA content). It is worth mentioning that when it comes to the contribution of the overall side scatter of cells, it has been remarked by [22] that

the nuclei do not seem to be the only contributor within cells (estimated to be approximately less than 40%).

On the other hand, contribution of approximately 90% of cell side scatter was attributed to the mitochondria in another study [18]. There, the light scatter of mitochondria was investigated via two-dimensional darkfield imaging of single normal and cancerous liver cells, and this large-angle scatter was observed to spatially correlate with the positions of the mitochondria within the cell. Side scatter images showed random distributions of intensity spots for the liver cancer cell. In comparison, these spots aggregated nearby the nuclei for the normal ones, and both of which were highly correlated to where the mitochondria fluorescence was disturbed [18]. This shows that variation in organelles' aggregations seem to affect the side scatter as anticipated from the light scatter theory [14], [15], [18], [20]. This result was also shown via a three-dimensional simulation [20], where two different mitochondria distributions in the cytoplasm were modelled: a random one (associated with cancer cells), and another where they aggregated nearby the nucleus (as seen in normal cells). The results showed that the simulated three-dimensional side scatter intensity patterns were distinguishable between the two cases. Moreover, their forward scatter also seemed to be affected, suggesting that the presence of diverse aggregation networks may affect cell size-related interpretations from forward scatter data [20]. The near submicron mitochondria being considerably large to exhibit a Mie-like (forward dominant) scatter was also mentioned in reference [18], and was suggested to correlate to the morphology of the organelle's network.

On another note, imaging experiments on human cervical carcinoma cells SiHa and rat embryo fibroblast cells MR1 showed that lysosomes and internuclear structures, (beside other small structures) seem to be as significant as mitochondria to large-angle scatter [24]. They also showed how the polarisation of the incident light played a role in determining the significance of an organelle in terms of its contribution to the large-angle scatter. This was observed in how the lysosomes' contribution to the scatter changed compared to the mitochondria when incident light polarisation was changed from parallel to orthogonal (relative to the scatter plane). The orthogonally polarised light increased the scatter of the lysosomes. In either case, the lysosomes had larger side scatter than the mitochondria [24].

Evidence in the discussed literature suggests different variables that could potentially affect cell light scatter, including differences in the cell lines of normal or cancer cells, their malignancy, and their viability. Factors that in hindsight could provide an understanding of the studied cells without perturbing their functionality with harmful markers or labels. For the osteosarcoma polyploid cells, the subject of this work, not only could the increase in DNA content lead to larger cell sizes or nuclei, but it could also increase their metabolism [25], which could lead to increase in mitochondria numbers [18] as well as other organelles. Some parts of the nuclear structures could be missing as well in polyploid cells compared to their diploid cells [1]. All of these suggest that not only could label free scatter of polyploid cells assist in identifying these cells apart from their diploid ones, but it could also provide feedback on the inner cellular structures causing any potential difference in scatter.

1.5   Research Hypotheses

Based on the presented literature review of observed cell scatter sources we propose the following hypotheses for the current work:

1. Cell-cycling-polyploid cohort populations in osteosarcoma can be detected via their high forward and side scatter intensity measurements relative to the diploid cells within heterogeneous samples.
2. Cell classification with the brightfield and darkfield images can segregate cell-cycling-polyploid cohort populations in osteosarcoma from the diploid cells within heterogeneous samples.

The first hypothesis is tested in Chapter 4 while the second in Chapter 5.

1.6   Research Objectives

1.6.1   Primary research objectives

Throughout the findings of this project, we aim to meet the following objectives:

1. Characterise the forward and side scatter of the ICRF-193 drug induced polyploid cells in osteosarcoma against the diploid cells in the sample (both as viable cells suspended in media) using benchtop flow cytometers, and state the optimum detection thresholds for the least diploid cell contamination.

2. Utilise machine learning to apply cell classification of two-dimensional brightfield or darkfield images, collected on imaging flow cytometer, for the target polyploid, non-target polyploid, and diploid cells, as well as extract meaningful statistical readouts from the images, such as cell sizes.

Both of which are addressed in Chapter 4 and Chapter 5, respectively.

1.6.2   Secondary research objectives

The approach of the current work involves handling the studied cells on different platforms. Therefore, the following secondary objectives aim to observe the quality and reproducibility of the results, as well as the analysis methods.

1. Propose a protocol for the handling of the investigated biological samples that ensures reproducibility of the biology and stability of the scatter measurements across different platforms.

2. Utilise recent automated cell population gating tools and optimise them as needed to unify the cell gating method (see section 2.5.3.2 in Chapter 2 for more) in the analysis for different platforms.

3. Identify biology or scatter controls that can be referenced on various platforms to replicate the findings and/or perform a label-free sort of the target cells.

4. Construct the label-free detection specifications for the target cells that could assist any commercial or non-commercial flow cytometer instrument or microdevice in the realisation of the thesis's findings.

The first and second objectives are addressed in Chapter 3. The third and fourth objectives are discussed mainly in Chapter 4 and Chapter 5, but also addressed throughout the rest of the thesis. All objectives are readdressed and discussed at the conclusion (Chapter 6) of the thesis.

1.7   Thesis structure

Following this chapter, Chapter 2 gives an overview on flow cytometry, the methodology of the current work. The overview intends to provide a foundation to understand the results, through discussing the different used instruments, their limitations with a reference to the thesis's results, and the standard analysis approaches. This is followed by the first results chapter, Chapter 3, which is concerned with establishing the results reproducibility and standardisation protocols for the thesis experiments. Chapter 4 shows the results addressing the validation of the first hypothesis. Chapter 5 follows suit with the second hypothesis. Finally, the thesis overall discussion and conclusions are presented in Chapter 6, where we readdress the research hypotheses, objectives and findings, as well as the future work.

1.8   References

[1] Z. Storchova and D. Pellman, 'From polyploidy to aneuploidy, genome instability and cancer', *Nat. Rev. Mol. Cell Biol.*, vol. 5, no. 1, pp. 45–54, Jan. 2004, doi: 10.1038/nrm1276.

[2] H. K. Brown, M. Tellez-Gabriel, and D. Heymann, 'Cancer stem cells in osteosarcoma', *Cancer Lett.*, vol. 386, pp. 189–195, Feb. 2017, doi: 10.1016/j.canlet.2016.11.019.

[3] A. J. Saraf, J. M. Fenger, and R. D. Roberts, 'Osteosarcoma: Accelerating Progress Makes for a Hopeful Future', *Front. Oncol.*, vol. 8, Jan. 2018, doi: 10.3389/fonc.2018.00004.

[4] Z. Shu, S. Row, and W.-M. Deng, 'Endoreplication: The Good, the Bad, and the Ugly', *Trends Cell Biol.*, vol. 28, no. 6, pp. 465–474, Jun. 2018, doi: 10.1016/j.tcb.2018.02.006.

[5] T. L. Orr-Weaver, 'When bigger is better: the role of polyploidy in organogenesis', *Trends Genet.*, vol. 31, no. 6, pp. 307–315, Jun. 2015, doi: 10.1016/j.tig.2015.03.011.

[6] D. T. Fox and R. J. Duronio, 'Endoreplication and polyploidy: insights into development and disease', *Development*, vol. 140, no. 1, pp. 3–12, Jan. 2013, doi: 10.1242/dev.080531.

[7] J. Coward and A. Harding, 'Size Does Matter: Why Polyploid Tumor Cells are Critical Drug Targets in the War on Cancer', *Front. Oncol.*, vol. 4, May 2014, doi: 10.3389/fonc.2014.00123.

[8] M. Jemaà *et al.*, 'Heterogeneity in sarcoma cell lines reveals enhanced motility of tetraploid versus diploid cells', *Oncotarget*, vol. 8, no. 10, Mar. 2017, doi: 10.18632/oncotarget.14291.

[9] P. J. Smith *et al.*, 'Mitotic Bypass Via An Occult Cell Cycle Phase Following DNA Topoisomerase II Inhibition In p53 Functional Human Tumor Cells', *Cell Cycle*, vol. 6, no. 16, pp. 2071–2081, Aug. 2007, doi: 10.4161/cc.6.16.4585.

[10] M. Purschke, N. Rubio, K. D. Held, and R. W. Redmond, 'Phototoxicity of Hoechst 33342 in time-lapse fluorescence microscopy', *Photochem. Photobiol. Sci.*, vol. 9, no. 12, p. 1634, 2010, doi: 10.1039/c0pp00234h.

[11] P. J. Smith, M. Wiltshire, and R. J. Errington, 'DRAQ 5 Labeling of Nuclear DNA in Live and Fixed Cells', *Curr. Protoc. Cytom.*, vol. 28, no. 1, Apr. 2004, doi: 10.1002/0471142956.cy0725s28.

[12] O. Sen, A. T. Saurin, and J. M. G. Higgins, 'The live cell DNA stain SiR-Hoechst induces DNA damage responses and impairs cell cycle progression', *Sci. Rep.*, vol. 8, no. 1, p. 7898, Dec. 2018, doi: 10.1038/s41598-018-26307-6.

[13] D. Wlodkowic, J. Skommer, and Z. Darzynkiewicz, 'Cytometry in cell necrobiology revisited. Recent advances and new vistas', *Cytometry A*, vol. 77A, no. 7, pp. 591–606, Mar. 2010, doi: 10.1002/cyto.a.20889.

[14] C. F. Bohren and D. R. Huffman, *Absorption and Scattering of Light by Small Particles*. Weinheim, Germany: Wiley-VCH Verlag GmbH, 1998. doi: 10.1002/9783527618156.

[15] M. P. Keating, *Geometric, physical, and visual optics*, 2nd ed. Boston: Butterworth-Heinemann, 2002.

[16] N. Catsimpoolas, Ed., *Cell analysis*. New York: Plenum Press, 1982.

[17] J. R. Mourant, J. P. Freyer, A. H. Hielscher, A. A. Eick, D. Shen, and T. M. Johnson, 'Mechanisms of light scattering from biological cells relevant to noninvasive optical-tissue diagnostics', *Appl. Opt.*, vol. 37, no. 16, p. 3586, Jun. 1998, doi: 10.1364/AO.37.003586.

[18] P. L. Gourley *et al.*, 'Ultrafast Nanolaser Flow Device for Detecting Cancer in Single Cells', *Biomed. Microdevices*, vol. 7, no. 4, pp. 331–339, Dec. 2005, doi: 10.1007/s10544-005-6075-x.

[19] X.-T. Su, C. Capjack, W. Rozmus, and C. Backhouse, '2D light scattering patterns of mitochondria in single cells', *Opt. Express*, vol. 15, no. 17, p. 10562, 2007, doi: 10.1364/OE.15.010562.

[20] X.-T. Su, K. Singh, W. Rozmus, C. Backhouse, and C. Capjack, 'Light scattering characterization of mitochondrial aggregation in single cells', *Opt. Express*, vol. 17, no. 16, p. 13381, Aug. 2009, doi: 10.1364/OE.17.013381.

[21] M. Lin, X. Qiao, Q. Liu, C. Shao, and X. Su, 'Light-sheet-based 2D light scattering cytometry for label-free characterization of senescent cells', *Biomed. Opt. Express*, vol. 7, no. 12, p. 5170, Dec. 2016, doi: 10.1364/BOE.7.005170.

[22] J. R. Mourant *et al.*, 'Light scattering from cells: the contribution of the nucleus and the effects of proliferative status', *J. Biomed. Opt.*, vol. 5, no. 2, p. 131, 2000, doi: 10.1117/1.429979.

[23] H. Hennig *et al.*, 'An open-source solution for advanced imaging flow cytometry data analysis using machine learning', *Methods*, vol. 112, pp. 201–210, Jan. 2017, doi: 10.1016/j.ymeth.2016.08.018.

[24] O. C. Marina, C. K. Sanders, and J. R. Mourant, 'Correlating light scattering with internal cellular structures', *Biomed. Opt. Express*, vol. 3, no. 2, p. 296, Feb. 2012, doi: 10.1364/BOE.3.000296.

[25] N. Sher, J. R. Von Stetina, G. W. Bell, S. Matsuura, K. Ravid, and T. L. Orr-Weaver, 'Fundamental differences in endoreplication in mammals and *Drosophila* revealed by analysis of endocycling and endomitotic cells', *Proc. Natl. Acad. Sci.*, vol. 110, no. 23, pp. 9368–9373, Jun. 2013, doi: 10.1073/pnas.1304889110.

Chapter 2 :

Flow Cytometry: methodology overview

2.1   Introduction

The hypotheses of the thesis are concerned with investigating the label-free parameters of the studied cells through laser light scatter as well as brightfield and darkfield images. This is conducted via flow cytometry (FC), a high throughput technique in which a light source illuminates or excites single particles within a focused stream. The resulting scattered or emitted photons are then collected and directed into detectors, where intensity pulses are converted into electrical signals to be measured and analysed for each passing particle. Optical flow cytometry can be categorised based on the detection method into two types, conventional flow cytometry (CFC), and imaging flow cytometry (IFC). In CFC, zero/one-dimensional detections are performed, where the total intensity of the scattered or emitted light is focused into a single point on a photon detector. In IFC, two-dimensional detections are achieved through brightfield, darkfield or fluorescence imaging on a high-resolution camera.

On either type of platform, the success of experiments relies on good experimental design and well considered control measures. Data reproducibility and usability can be realised by standardising as many steps as possible across the different platforms. These include sample preparation protocols, acquisition settings, and data analysis methods. In this chapter, we discuss aspects of the flow cytometry instrumentation (section 2.2), experiment design (section 2.3 and 2.4), and data analysis (section 2.4), which highlight key similarities or differences regarding the used instruments in the current work. This understanding guides our data collection and analysis methods throughout the thesis. It also illustrates the shared potential of these instruments regarding our research objectives, as well as the expected limitations. For further background and insight on flow cytometry, resources such as Practical Flow Cytometry by Shapiro [1] are recommended.

2.2   Instrumentation

In this work, the investigation on cell scatter is performed on data collected via two CFC instruments, BD FACSCalibur and the digital BD FACSVerse (Becton Dickinson Inc., UK)., and via one IFC instrument, Amnis® ImageStream®X Mark II (Merck Millipore) (see Figure 2-1). In terms of hardware circuitry, FACSCalibur is an

example of analogue CFC instruments that are in use today, while FACSVerse is that of digital CFC instruments. CFC data could be said to give better statistical intensity readouts per cell and with a less complex system. On the other hand, the visualisation of the detected cells in IFC gives a better insight regarding the sources of the collected light as well as providing more parameters for analysis.
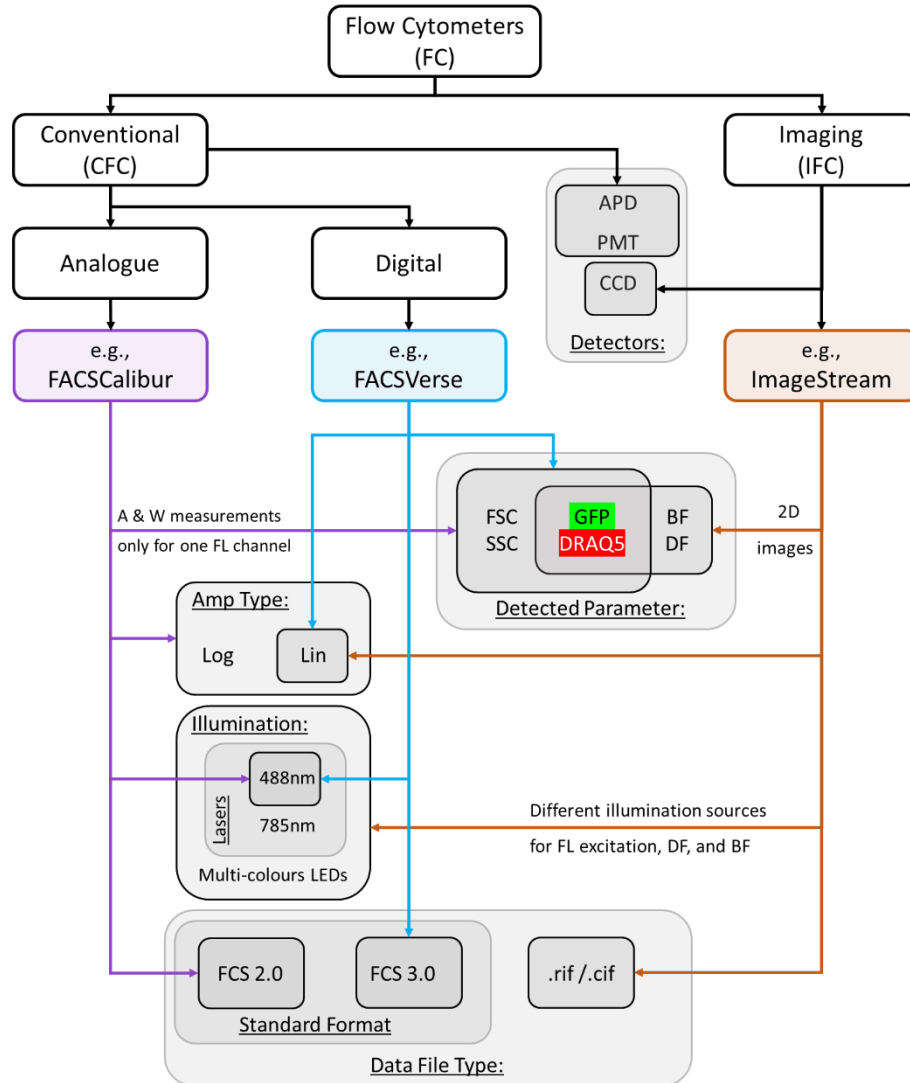
Figure 2-1: A flow chart summarising the flow cytometry instruments in the current work in terms of detectors, operated illumination sources, signal amplification type, the detected parameters of interest (where label-free ones are in black), and the data file types (please refer to the relevant sections in this chapter for further details).

The working principle of this method is mostly shared between all commercial flow cytometers. Nonetheless, consideration should be given to the inherent instrumental differences. As has been stated above, such understanding helps in the standardisation

steps of the experimental data collection and analysis across the different instruments. It also helps in recognising oddly behaving samples, anticipating differences, and correcting for errors. In this section, we present a brief description of the main parts of flow cytometers while referencing the used instruments as examples. This starts from when cells enter the analysers through their fluidic system, interact with the optics, and data conversion within their electronics.

2.2.1 Fluidics

The instrument's fluidics is the first stage with which a sample interacts in flow cytometers, and it serves the purpose of shaping the sample into a focused stream before the illumination source for single particle analysis. This subsection briefly explains the process, then discusses how the relevant experimental settings are handled on the different instruments in the current work.

Within the instrument, the fluidic sample is first pressurised then injected into the sheath fluid within the flow cell or flow chamber (see Figure 2-2). The relative speed and pressure between the sample and the sheath fluid is controlled to achieve hydrodynamic focusing, where the sample core flows in a single stream that does not mix with the surrounding fluid following laminar flow principles. Both fluids travel through a nozzle shaped cavity located at the end of the chamber, then into an optically transparent region for laser interrogation (which is made of quartz cuvette in the used CFCs). The speed and width of the sandwiched sample core is controlled by the relative pressure between the sheath fluid and the injected sample. While this pressure setting can be manually configured on flow sorters, it is fixed at pre-defined modes (also referred to as flow rates) by the manufacturer for most of the flow analysers such as the FACSCalibur and FACSVerse.

The samples throughput rate can be controlled by the pre-set flowrate modes for various objectives. For example, low flowrates increase detection sensitivity through the slow interrogation time, which, if compromised, may affect the spread of the measured intensity peak distribution for samples, leading to high values of coefficient of variance (CV). Slow sample speeds also result in narrower sample cores that can reduce chances of coincident events (especially with smaller particles). On the other

hand, an overall short experiment time is desired when working with live cell suspensions to avoid affecting their health negatively, especially when they are labelled with cytotoxic probes such as the studied samples. Therefore, in the current work, higher flow rates (e.g. approximately 60 µL/min) are used for bead or cell samples wherever the detected intensity CVs are found to be insignificantly affected by the sample speed changes. However, for cell cycle analysis where proper DNA quantification is prioritised, we have maintained the recommended flowrate as per protocols (i.e. speeds not exceeding 300 cells/s).
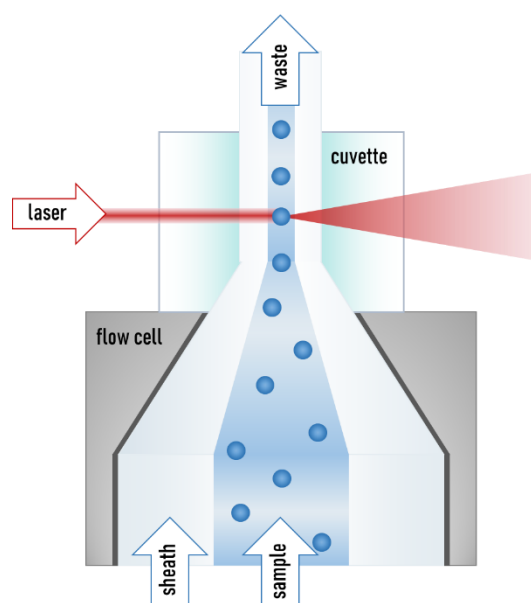
Figure 2-2: A scheme of the hydrodynamic focusing of samples in a typical FC analyser.

In IFC, flowrate speed settings are associated with the desired image sensitivity for an experiment. On ImageStream, fixed pre-set speed modes are once again implemented. These help in managing the introduced complexity in IFC instruments that is associated with focus calibration in detection cameras. For example, manufacturer supplied speed beads are continuously mixed with the sample stream in ImageStream, where they are monitored to automatically adjust the camera focus and imaging synchronisation in real time. For our experiments, image quality is prioritised for subsequent machine learning analysis, so all samples are run on the lowest fluidics flowrate.

Different instruments have different requirements regarding the acceptable volumes or recommended sample concentrations. In some studies, this may not be of much

concern, leading to no considerable changes to their sample preparation protocols. However, in our study, we have found that the sample requirement by ImageStream has introduced some challenges regarding keeping our sample preparation protocol standardised across all the used FC. The issue stems from how the instrument anticipates very condensed samples ($\sim$2x10^7/mL) in low volumes (20-200 μL), which does not align with the standard labelling protocols for DRAQ5, the used DNA label for cell cycle analysis of our studied live cells. As a result, some adjustments have been made to some of the sample preparation steps as would be discussed in detail (along with other challenges) in the methodology (section 5.2.1.3) of Chapter 5.

2.2.2   Optics
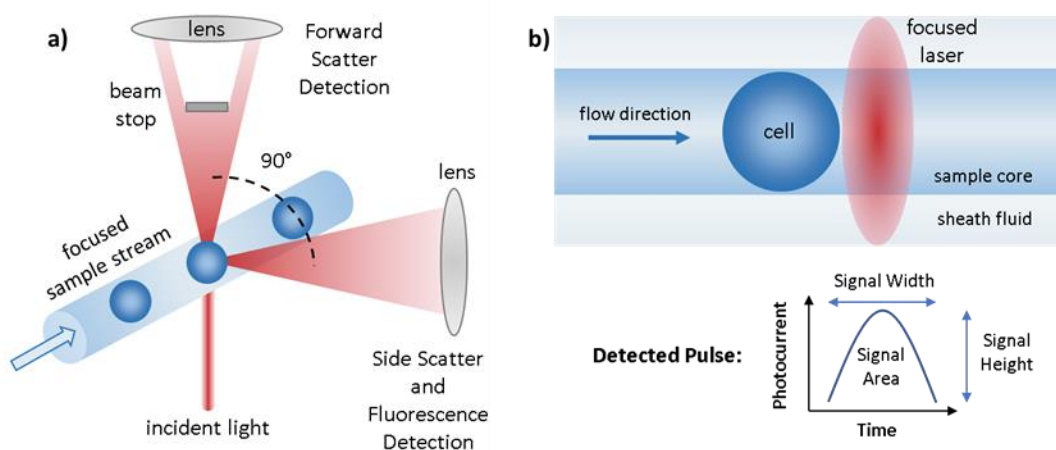
Figure 2-3: The **a)** detection directions and **b)** signal parameters extracted from the pulse generated from a passing cell in a typical CFC analyser.

The optical system in commercial flow cytometers mainly consists of illumination sources, detectors, and optical elements for light collection, directing and filtering. In the following subsections, we highlight similarities and differences in the used instruments regarding the provided light sources and the resulting scatter and fluorescence emission. Light detectors are then discussed where we address some of their advantages and downsides, particularly those associated with our experiments.

2.2.2.1 Light sources, light scatter, and fluorescence emission

Lasers are the main light source used for scatter illumination and fluorescence (FL) excitation in FC instruments. They are preferred for the job because of their coherent nature, which provides narrow spectral bandwidth in high intensities that is ideal for fluorescence excitation and detection. Most manufacturers provide wavelengths that are commonly used with fluoroprobes in biological experiments, such as the 488 nm blue laser used in the current work. Nonetheless, these instruments may have differences regarding how the laser light is delivered or shaped to interact with the samples. For example, the geometry of the laser beam at the focus point is elliptically shaped as 22 x 66 μm and 9 x 63 μm for the FACSCalibur and FACSVerse, respectively (where the shorter axis is parallel to the sample's travelling direction; see b in Figure 2-3. Because power distribution in lasers is nonlinear, variability in the shape of the illumination beams could trigger responses that are relatively different within the experiment samples from one instrument to the other. This is of relevance to scatter results (which is highly dependent on the incident power, see Chapter 3 section 3.3.2.2), especially that of particles with sizes comparable to that of the beam, as would be discussed in Chapter 4 (section 4.3.1.1).

In the used CFC instruments, the forward scattered light (FSC) is detected by a photodiode with bandpass filter, where an obstruction bar is placed around the 0° direction (relative to the optical path) to block most of the transmitted light (see a in Figure 2-3). Around the 90° direction, the side scattered light (SSC) and the FL emission are collected and directed by beam splitters to be detected by photomultiplier tubes (PMT) supplied with the appropriate wavelength filters. Some of the used beam splitters can be polarised, such as the Brewster-angle beam splitter in front of the side scatter detector in FACSCalibur. On ImageStream, a different laser (785 nm laser) is used to create the 90° two-dimensional darkfield (DF) images (in place of the zero/one-dimensional side scatter), but the same blue laser wavelength (488 nm) is available for fluorescence excitation. There, in place of forward scatter, a multi-colours LEDs illumination in a direction perpendicular to that of the other lasers is used to create two-dimensional brightfield (BF) images. All generated light is then collected to travel on the same optical path, passing by band-stop filters before being detected on high-resolution CCD cameras (two cameras collection via a beam splitter is available on

our instrument). Illustrations of the optical paths for both FACSCalibur and ImageStream can be seen in Figure 2-4 b.
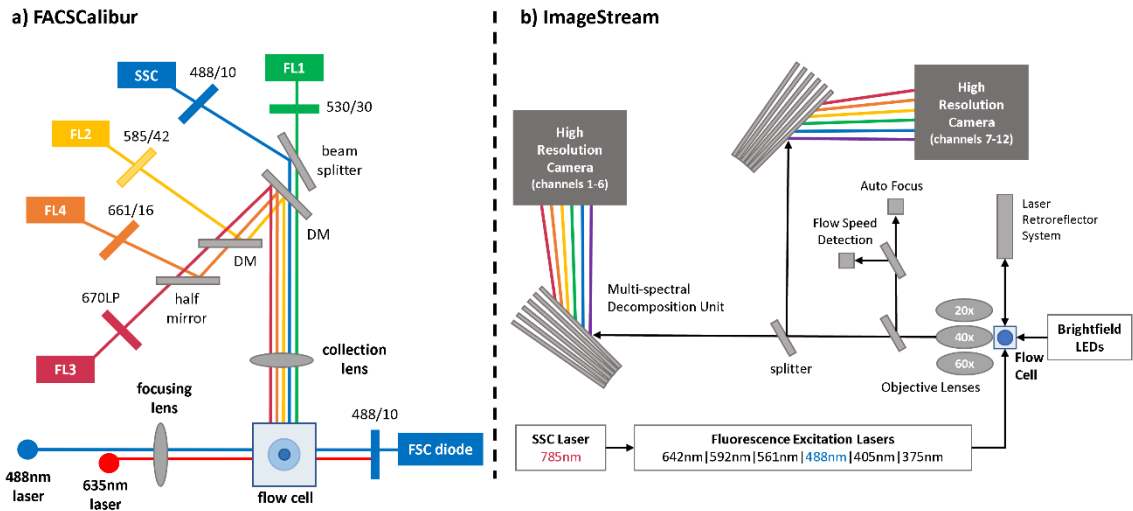
Figure 2-4: Overview of the optical path for a CFC instrument such as the **a)** FACSCalibur (Becton Dickinson Inc., UK), and an IFC instrument such as the **b)** ImageStream (Merck Millipore). Figures are re-created from instruments manual as provided by their respective manufacturer.

Albeit not the same, all instruments are equipped with suitable filters for the detection of the fluorescence of interest, which enables the associated sample preparation steps to be nearly standardised.

2.2.2.2 Light detectors

In commercial CFC, photomultiplier tubes (PMT) are typically used for the detection of side scattered light and fluorescence emission of incoming particles. These detectors capture well the wide range of the detected intensity between the fluorescent and non-fluorescent samples as well as the weak intensities of the side scatter light with their high gain. Avalanche photodiodes (APD) are often used instead for forward scatter detection, which tends to generate higher intensity signals for the particle sizes that are commonly investigated on flow cytometers. For IFC instruments, charged coupled device (CCD) cameras are used instead to produce two-dimensional brightfield, darkfield and fluorescence images of samples.

In the following subsection, we discuss some detection problems we have observed that are mostly inevitable in the current work and thus regarded as methodology limitations.

2.2.2.2.1   Saturation in detectors

There are two main sources of high intensity differences in our investigated samples. The first is related to variability in the area or size of the light intensity source(s) within the particle, and the second is related to variability in the localised spot's intensity. While both types can affect all detectors, we comment here on which type is observed the most with which detector (thus instrument) in our experiments.

The first type of saturation can be found in the PMT detection of the polyploid cell samples that are labelled for DNA quantification. When labelled for their DNA, diploid and polyploid cell samples display variable nuclei sizes and intensities. Normally, researchers utilise logarithmic scaling (logarithmic amplification on analogue CFC) to record and view both dim or small intensity structures along with bright or large intensity structures on the same scale (e.g. as done in the current work with the GFP- cyclinB1 label readout). However, linear scaling is preferred for cell cycle DNA readout because it helps track the expected linear multiplication of DNA content during such an event. As a result, we have found that beyond the regular diploid cell cycle, we could only view the first polyploid cell cycle and half of its second one before the PMT detector saturates (see plot a in Figure 2-5). In this case, the dynamic range of the measured total DNA intensity per cell is limited by the voltage setting of the diploid DNA peaks, whose values increase in multiples when the polyploidy DNA peaks are detected. Lowering the voltage of the diploid DNA peaks can be done to visualise the higher intensity peaks through logarithmic scaling, but this may come at the cost of potentially introducing errors in the measurement for some CFC instruments (see section 2.2.3.1 for details), On the other hand, CCD cameras exhibit a much larger dynamic range when it comes to the measured total intensity of a given parameter per cell. In the example of total measured DNA, detection on CCD cameras can cover the second polyploid cell cycle in our samples properly (see plot b in Figure 2-5), where the dynamic range of the detected intensity is 12-bit per pixel. This is possible because the total intensity of the nuclei is calculated in CCD cameras

afterwards via adding up the pixel values of the imaged intensity areas. Therefore, the intensity integration measurement in CCD cameras can offer a larger dynamic range compared to PMT detectors, as long as the individual pixels are not saturated, which we will discuss next.
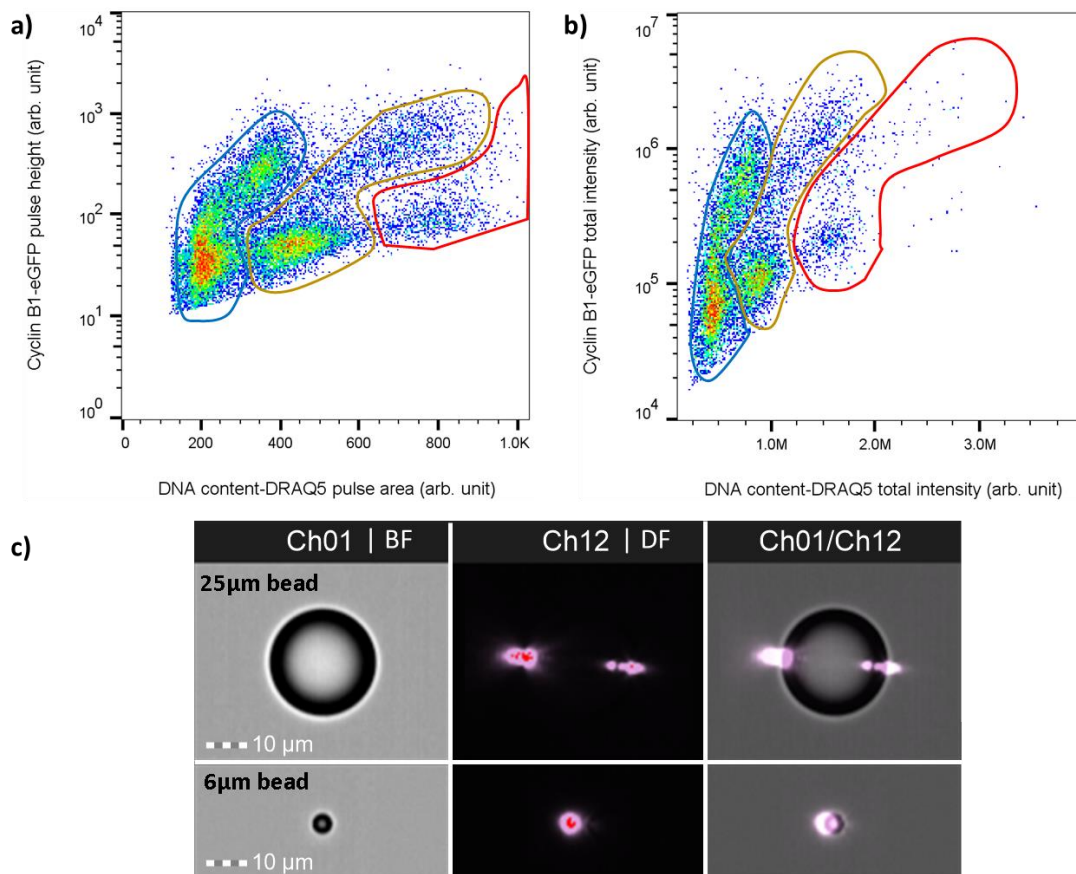
Figure 2-5: Examples from the thesis on the type of observed detection saturation. The top scatter density plots of cell cycle readout show the first type of saturation observed on the **a)** PMT detectors of CFC instruments (FACSCalibur) where only half the second polyploidy (red gate) cell cycle is detected, unlike the **b)** CCD camera of the ImageStream that show them all. In the scatter plots, the shift in cell number density from high to low is represented by the colours red, yellow, green and blue, respectively. The three population round gates in the plots (**a**, **b**) are manually drawn (for illustration purposes and not for analysis), via referencing control samples (see section 3.2.3 in Chapter 3). The second type of saturation is observed on **c)** the localized pixel saturation of ImageStream images, such as those observed with the darkfield images of plain small or large beads (saturation represented by the red pixels in Ch12). The columns in **c)** refer to the brightfield, darkfield (1 mW), and an overlayed image of both, respectively. The top and bottom row images are for a 25 µm and 6 µm bead, respectively.

The second type of saturation is often encountered in CCD cameras as they have lower localised intensity dynamic range compared to the other flow cytometer detectors, and

so they tend to struggle with pixel saturation. This is especially the case when it comes to scatter detection, i.e. darkfield images, of uniform particles such as our control scatter bead samples. In general, focused two-dimensional darkfield images of beads show high intensity spots where the laser beam enters and exists the beads (see images of c in Figure 2-5). This is due to the high sphericity and composition-uniformity of the beads, which then results into high reflectivity at these boundaries. In the CCD cameras of ImageStream, these high intensity spots have shown pixel saturation even when the scatter laser is operated at the lowest power setting (e.g. 1 mW). Therefore, this renders the usage of beads as scatter (or, in this example, darkfield images) controls inadequate on our IFC instrument. Nonetheless, scatter beads are routinely run in our IFC experiments (as done in the CFC ones) to observe the stability of the imaged brightfield intensities between the different experiments. It is worth mentioning that one may apply attenuating filters on the darkfield detection channel to reduce the saturation, but that could risk eliminating the dimly spotted patterns commonly seen in the darkfield images of our cells (whose intensity cover the entire pixel range of the CCD camera moving from diploid to polyploid cells at a laser power of 1 mW).

2.2.3   Electronics

Once the scattered or emitted photons hit the detectors, the electronic circuits handle the job of the pulse conversion and information extraction. Different instruments have their own approaches to how the detected signal is handled in the circuitry of the flow cytometer, but all essentially aim to provide similar readouts: height, width, or area of the detected pulse (refer to b in Figure 2-3). Nowadays, most CFC flow cytometers provide all three measurements, but on analogue CFC instruments, there could be limits on how many measurements are conducted per parameter (scatter or FL channel). For example, on FACSCalibur, users can measure the signal height for each parameter, but the circuitry limits measuring signal's width and area to only one FL channel, which must be set before the data acquisition. Another major difference between digital and analogue CFC is concerned with the analogue signal amplification and digitisation. This is discussed in detail in the following subsection as it is an issue that can considerably affect the subsequent data analysis

2.2.3.1 Pulse digitisation and log transformation (analogue versus digital flow cytometers)

Attention must be given to how or when a flow cytometer extract measurements from the analogue pulses of the events as they pass in front of the detectors. Older FC models perform all pulse measurement extractions (also known as parameters) within the instrument's analogue circuitry before they are digitised for computer display and data storage. Depending on the parameters selected for acquisition, the detected pulse per channel is first passed through the appropriate circuits to have its height, width and/or area measured. Each measurement is then passed to an analogue-to-digital-converter (ADC), where the signal resolution depends on the bitrate. Fluorescence compensation, when desired during acquisition, is also handled within the instrument hardware. Such a compensation process, however, is irreversible because the raw uncompensated data is lost, making it impossible to fix any user-related compensation errors (thus it is preferred to apply any needed spectral overlap corrections afterwards). In FC, logarithmic scales (see section 2.5.3.1.2) are often used to display distributions of weak intensity signals on the same scale as that of high intensity ones. Such readouts are commonly observed with fluorescence labelling of cells as well as the scatter of small versus large particles. However, due to the lower bitrate of the ADCs in the older generation of flow cytometers (e.g. 10-bit on FACSCalibur), linear measurements of the pulse show very poor resolution of the weak signals when displayed logarithmically afterwards. Therefore, if a logarithmic scale display is desired for a given detection channel, linear-to-log transformation is first applied to the analogue signal before it is digitised. This, once again, results in the hardware log-transformation being an irreversible process, where the original linear intensity readout cannot be retrieved. The log decades allowed per detection channel depends on the dynamic range capacity of the instrument and the bitrate of its ADC. Naturally, the process of log-transformation within the circuitry does introduce errors to the transformed measurements, mostly observed at the weakest and highest intensity decades [2]. Therefore, it is recommended, to adjust PMT voltages to centre the detected signals within the dynamic range of the scale wherever possible. It should also be noted that, while the acquired data files store the raw log-transformed data as they are, most flow cytometer data visualising programmes applies a reverse log-to-linear transformation (see section 2.5.3.1.1) when it displays them to users. This also

may introduce some errors because of the inherently non-perfect original log-transformation. Hence, it is best to keep the log-scale data visualisation when handling any analogue log-transformed data on older flow cytometers.

One last thing to discuss regarding older models is their tendency to have limited number of measurements (height, width, or area) per detection channel. For example, on FACSCalibur, the main parameter extracted for all detection channels is the peak of the intensity pulse. The width and area of the pulse measurements can be measured only for one single fluorescence channel (this setting is not available for scatter channels), which the user must select before the data acquisition. This means that data filtering that requires the presence of at least two pulse measurements, such as that of coincidence events, could only be applied for this instrument on an appropriate fluorescence channel (e.g. the labelled nuclei in the current work). On such an instrument, scatter channels cannot be used for this purpose, which, otherwise, would have been a researcher's first choice for the filtering of such doublet events.

Modern flow cytometers perform signal digitisation in the instrument circuitry first before measurements are extracted. These flow cytometers tend to have ADCs with higher bit rate (e.g. 18-bit for FACSVerse), which allows better resolution of weak signals. This omits the need for any analogue log-transformation and all collected parameters are inherently linear regardless of the user's choice of data display during the acquisition. In addition, colour compensation here is performed digitally via the acquisition software either during or after experimental collection.

2.3   Controls in flow cytometry

The validity of the conclusions made in research is dependent on the quality of the considered control samples and how they support the hypothesis. In FC, controls vary from those assessing the performance of the instrument to those related to the experiment design and objectives. The purpose of these controls is to ensure that most known experimental variables are accounted for, so concrete conclusions can be drawn. In this section we give an overview of the controls used in our work such as the quality, experiment, system and compensation controls. Further discussion can be found in each results chapter in the thesis.

Instrument testing controls, or quality controls (QC), refer to the manufacturer-specified quality tests that must be run routinely to ensure the system is calibrated and performing to the standard. They come in the form of plain and fluorescent bead kits (provided by the manufacturer), and are usually run monthly, daily, before and after experiments, or according to recommended in-house protocols.

Experimental controls are concerned with checking the validity of any applied treatment, which are represented in our work by the different sample types discussed in Chapter 3. The polyploidy of cells in our work is drug induced [3], thus it requires non-drug treated samples to observe the drug response against the regular biological status of the cells. This is discussed in depth also in Chapter 3, which describes the assessment of these controls through all the conducted experiments. Therefore, the experimental controls in our work can also be regarded as our biological controls.

System controls refer to those that assist in the standardisation of experiment specific settings. In our study, these involve unlabelled cells as well as polystyrene beads to fine tune the scatter detection channel for any changes between experiments and/or to target cell side scatter as would be explained in later chapters.

Lastly, compensation controls refer to those that help assess any potential spectral overlap between the detection channels and they are discussed in depth in the next section (2.4).

2.4   Fluorescence compensation

Spectral overlap between neighbouring detection channel on flow cytometers is often inevitable, and thus must be accounted for. Uncompensated cell data with huge spectral overlap could falsely consider unlabelled cells as significantly fluorescent cells in its respective fluorescence detection channel. This could lead to inaccurate interpretation of the actual data; hence proper fluorescence compensation should be applied wherever needed. This starts first by choosing the suitable colour panel of biomarkers to identify target populations, along with the associated excitation lasers, to running the needed compensation controls during the experiment (as mentioned in section 2.3) and correcting the data via applying appropriate compensation matrices. Many FC data analysis software help in automatic generation of compensation

matrices, but users are recommended to manually inspect them afterwards and manually optimise them if needed.

When several flow cytometers are considered, it is also recommended to plan a flexible colour panel that can be run with the same sample preparation steps if data reproducibility over different platforms is to be examined. Commercial flow cytometers often carry similar wavelengths for the excitation lasers, such as the 488 nm blue laser or a 633-642 nm red laser, as well as nearly similar bandpass filters for fluorescence detection in the green, yellow, and red wavelengths. This makes it possible to replicate experiments with a panel of limited colours on different instruments, but it could be challenging as the colour panel expands or different laser wavelengths are needed. For example, for the experiments on the current work, we only require two biomarkers (thus, two fluorescence detection channels) excited by the 488 nm blue laser to biologically identify the polyploid cells apart from the diploid ones. One of these biomarkers is cyclin B1, whose levels increase gradually (starting from the end of the first growth [G1] phase) as the cell transition to the second growth (G2) phase of the cell cycle, where they peak at the G2/Mitotic (M) phase (see the y-axis in plots a or b in Figure 2-5). The cell line in the current work is transfected with a green fluorescent protein (GFP) tagged cyclin B1, where GFP is generally regarded as a non-invasive label. Then to tell populations such as diploid cells at G2 apart from polyploid cells at G1, we use DRAQ5 to quantify the DNA in the nuclei of live cells (see the x-axis in plots a or b in Figure 2-5). DRAQ5™ is a cell-permeant DNA binding anthraquinone dye that intercalates between A-T bases of double-stranded DNA [4]–[6]. In this work, it is used for its colour compatibility with GFP (it can be excited by the same blue laser as the GFP's and it fluoresces in the red wavelengths. It is also used for its simple labelling protocol that minimises any extra stress inducing steps on the cells. This makes the biology findings easily reproducible on several CFC because the chosen panel reduces any significant spectral overlap in the detected data, resulting in practically no need for fluorescence compensation.

However, even simple colour panels, such as ours, can be challenging if they require different sample preparation protocols because some flow cytometers do not strictly handle the samples in a similar manner. For example, the IFC ImageStream recommends sample concentrations that are much denser, and prepared in liquid

volumes that are much smaller compared to the ones followed in most CFCs such as FACSCalibur or FACSVerse. This required some changes to the regular labelling protocol of DRAQ5, as will be addressed later in Chapter 5 (see section 5.2.1.3).

Lastly, it should be noted that most CFC data analysis does not consider scatter channels when building compensation matrices. This is based on the simple fact that when a single laser is required for both scatter and fluorescence excitation, it is not feasible for regular users on the CFC instruments to selectively block or filter out the resulting forward and side scatter while keeping their detectors running to catch any fluorescence spillover. For most CFC instruments, the scatter illumination lasers and their respective scatter detection channels are pre-determined in the instrument hardware, where the main excitation laser (commonly the 488 nm blue laser laser) is often used for this job. On the other hand, ImageStream, the used IFC, operates illumination wavelengths for the brightfield and darkfield images that are different from those of the lasers used for fluorescence excitation, so regular users could build up compensation matrices of all detection channels to count for any spillover. Therefore, in the current work, while the spectral overlap for fluorescence channels is not significant on the used CFC instruments, we did consider any increase in the scatter medians of the GFP expressing versus non-GFP expressing cells as a baseline for the statistical significance of the studied scatter (see section 3.3.2.2 in Chapter 3 for related discussion). For the experiments on IFC, the regular fluorescence compensation procedure has been applied.

2.5   Data in flow cytometry

This section discusses how cell data are handled digitally. This starts from what parameters can be controlled during data acquisition, to how it is stored and accessed, and lastly to how analysis is conducted. The section aims to highlight which methods can be standardised across the used flow cytometers as well as discuss any observed limitations and how they are handled in the thesis.

2.5.1   Data acquisition

Flow cytometers provide real time measurements of the detected events. Therefore, it is common practice to first run some sample or control tubes to ensure that the flow cytometer is operating on the appropriate settings for the intended application before the desired data are acquired and saved. Users could also set up various collection gates of interests during the set-up stage to increase the number of meaningful data in their acquisition. In this section, we discuss some configurable options in the instruments used in this thesis that can affect the quality of the collected data.

In analogue CFC instruments such as the used FACSCalibur, PMT voltage settings can be configured separately for each of the width, height, and area parameters for a limited number of FL detectors (none for the scatter detectors on FACSCalibur). In the current work, this is only considered for the DNA-DRAQ5 FL detector channels since DNA linear quantification can be used in doublet discrimination. Because it is a FL readout of the nuclei, there is generally less concern about the variation in physical cell size among the different samples (unlike the case with scatter generally as will be discussed next). However, the cells in our drug-treated samples tend to have irregular multiples of the regular DNA set. Therefore, doublet discrimination in such samples may not result in cleanly separated populations, compared to drug-free samples, which contain cells of regular DNA sets.

In digital CFC instruments, area scaling is a detector-related setting that is commonly present for data acquisition. The setting allows users to configure the scaling factor between the pulse's area and height measurement for the detection channels. For digital CFC instruments, users usually can control the PMT voltages (electronic gain for the FSC's photodiode detector) only for the pulse's height measurement, so the area scaling setting helps ensure that the other measurements (area and width) are within the dynamic range. The linear relationship between the area and the height of the pulse depends on the detector, thus the FSC detector has its own setting independent from those detected on PMTs. The scaling factor could also vary between samples depending on changes in the sheath flow speed or the size of the particles. When set up well (i.e. via achieving a 45° angle linearity), the resulting scatter plot could also be used for doublet discrimination (see Figure 4-4 in Chapter 4). Generally, default values for the area scaling setting are established by the acquisition software

based on QC tests that are routinely run using small beads (approximately 2-6 µm). However, this default scaling becomes inaccurate as the size difference of the experiment samples increases (relative to the QC beads), and users are then recommended to manually adjust (wherever configurable) the area scaling setting for each sample of different size [7]. For the cells studied in the current work, the suspended live cells (kept on ice) can vary in size from approximately 16±1 to 29±2 µm depending on their position on the diploidy or polyploid cell cycle (see results in Chapter 5). Moreover, scatter detection is of importance to the current work, and any changes in the area scaling requires adjustment to the voltage settings of the concerned parameters. Changes in the voltage settings of the scatter channels between the different sample types in our experiments is not desired. As a result, the area scaling in the current work is set up (wherever configurable) based on the scatter and FL of the drug-free control cell samples, and is kept the same for the rest of the experiment (regardless of the cell size differences).

For ImageStream, the IFC in this thesis, the regular operation of the instrument allows users to adjust the detected intensity peaks on the CCD cameras via changing the power of the illumination lasers.

2.5.2   Data file types

Before approaching the data analysis, it is essential to highlight some points regarding how the detected intensity signals are handled or stored digitally. This is especially significant when it concerns experiments conducted across several platforms, be it analogue or digital.

The "Flow Cytometry Standard" or "FCS" is the standardised data file format for CFC (it is currently at FCS version 3.2). Most of the acquisition programmes in digital CFC export data in the FCS 3.0 format (introduced in 1997 [8]), while those in analogue CFC are still using the FCS 2.0 one (introduced in 1990 [9]). In either version, the FCS file consists of three main segments, which are the Header, Text and Data segments. The Header segment states the FCS file version and the byte offsets of the rest of the segments, while the Text segment contains the experiment's metadata (in the form of main and optional keywords), and the Data segment contains the numerical values of

the acquired raw data. More details regarding FCS files and their different versions can be found on the official website of the International Society for Advancement of Cytometry (ISAC) and their related previous publications (https://isac-net.org/).

Many commercial FC analysis programmes are designed to handle both FCS 2.0 and 3.0 well without much user intervention. They also often facilitate FCS file conversion to updated versions, which makes FCS 2.0 files more accessible. However, some of the recently published analysis tools tend to expect FCS 3.0 files that are exported by digital CFCs. This could result in file reading errors and multiple bugs when dealing with FCS 3.0 converted files from older machines. Moreover, when further analysis by other means is desired or the data in the FCS files are exported in other non-FCS formats, the knowledge regarding the original FCS file version and the type of the raw data amplification per parameter becomes crucial. An example of the first issue is presented in the current work (see section 2.5.3.2 here, and section 3.2.3.2 in Chapter 3), where updates to the published script of the tool had to be written to make the analysis on the FCS 2.0 files of our FACSCalibur data possible. Regarding the second issue, any non-FCS numerical data file exports were routinely associated with their respective metadata (e.g. amplification type, range) to ensure accurate representation and handling of the data. An example of that is the numerical data exports of the logarithmically amplified parameters in our FCS 2.0 files from the FACSCalibur (which is addressed in section 2.5.3.1.1).

For IFC, there is currently no standardised file format for the imaged data (though recommendations have been proposed). The IFC in the current work, ImageStream, stores experiment data as raw image files (.rif) or compensated image files (.cif), which are mainly read and processed via the analysis software provided by the manufacturer. Nonetheless, because individual images can be exported in popular image file formats such as .tiff, this allows us to further study them via open-source imaging analysis tools as illustrated in Chapter 5.

2.5.3   Data analysis

In the following subsections we discuss some of the methods and approaches that are referenced throughout the thesis when it comes to data presentation or analysis.

2.5.3.1   Data visualisation

Most of the FC experiments are concerned with examining the existence and/or lack of markers of interest through their fluorescence emission. The collected cell intensities for each parameter are first visualised via traditional two-dimensional scatter dot plots, density plots, or frequency histograms. This visualisation step helps identify well-known distribution patterns, such as that of cell cycle in the current work, and apply population thresholds accordingly. In this section, we discuss aspects of the used data visualisation in the current work such as the common graph types, and the appropriate binning or scaling methods.

Frequency count histograms are the default data presentation option per parameter on most FC data analysis software and they are the simplest to interpret. Representing the data of each parameter through probability distribution functions (PDF) could also be used instead to convey a similar reading. However, the visible bins in frequency count histograms help remind readers of the binning process that is being applied in the data visualisation (even on scatter plots) or even during its acquisition. Therefore, most of our single parameter visualisation is done via histograms. Histograms or PDFs are also plotted sometimes along the axes of scatter plots when multiple populations are graphed (e.g. Figure 3-5 in Chapter 3). In our results, we also use box plots to summarise statistics (i.e. mean, median, and the middle 50%) of individual histograms, but it should be noted that traditional box plots may not reflect multimodal distributions well. For such a job, violin plots could be used instead because they combine distribution density and box plots, but they are less commonly used and can appear overwhelming. Therefore, we will simply report box plots along with frequency histograms wherever we deem needed.

In the current work, when it concerns data binning in frequency histogram for display purposes, the bin width is not fixed and is thus stated in figures' captions wherever applicable. It should be emphasised that it is common for FC analysis software to use the binning or scaling choices of the single parameter histograms as the backbone for the data binning or scaling of the rest of available figure types. These options sometimes can be found in their exported numerical data (in other file formats).

Topics such as data binning and scaling in FC are well explained in previous review articles [10], so the following subsections echo their remarks with reference to our data and the used instruments.

2.5.3.1.1   Data binning

Binning refers to the process of assigning data to discrete categories or bins, which can be of uniform or variable widths. In FC, choices regarding data binning are mostly concerned with data display rather than statistical computations. This is briefly discussed here along with some exceptions that involves data transformation.

Originally, all FC data are binned as the detected analogue signals are converted during the acquisition to digital ones, where the binning resolution is dependent on the instrument's ADC. The binning resolution is often reduced further via analysis software to make the data visualisation sensible. For example, even a reasonably high events count, such as the 30K for some of our samples, is not sufficient to fill up all the 262,144 intensity channels (18-bit ADC) on the FACSVerse. In such a case, any intensity peaks or valleys would less likely be recognised unless the binning resolution is reduced to 1024 channels (10 bits) or 256 channels (8 bits).

As pointed earlier, most FC analysis software compute data statistics on the stored raw data regardless of the user's binning choice for the data display. However, logarithmically amplified data in analogue CFC instruments are often handled differently. For example, it is common for commercial FC acquisition or analysis software to apply a reverse transformation on the raw linearly-binned-but-log-transformed data before computing statistics, where the file's stored metadata are referenced for the transformation parameters. To illustrate further, the ADC in FACSCalibur first perform a 4-decade log amplification on the acquired signals before digitising them linearly into 1024 channels (10 bits). Analysis software would then transform these values into linear readings of 10^4 channels before reporting statistics on them. The transformation is usually done using the formula $y = f2 \times 10^{f1 \cdot x/R}$, where x is the raw channel number, y is the scaled value (also called relative brightness measurement), R is the range of the raw channel data (e.g. 1024 channels in FACSCalibur), f1 and f2 refer to the amplification's decade number and the offset

value of the raw scale, and they can be extracted from the file metadata (e.g. values of f1 and f2 can be extracted from the keyword $PnE, which refers to the amplification type in the FC electronics for a given parameter n) [8], [9]. One should note that such formulas assume ideal logarithmic amplification of the instrument's circuitry and does not count for any non-linear behaviour. Therefore, this data transformations can naturally introduce additional errors when applied on raw data [1], [2] (see section 2.2.3.1).

Binning in FC analysis software often follows the display scale, where linear binning maintains absolute bin widths while logarithmic one maintains relative bin widths. As a result, if two dim and bright intensity distributions of equal CVs are plotted in a histogram, their visible distribution widths may appear unequal on a linear scale but will appear equal on a logarithmic one. Further consequences of such binning methods are discussed in the following section.

2.5.3.1.2   Data scaling

Axis scaling is another point that should be highlighted in FC data presentation. Firstly, all intensity scales are in uncalibrated arbitrary units, and they are titled after their respective instrument's specific detection channel as commonly practiced in FC research. In general, linear scales are often used to visualise cell scatter as well as FL channels where linear intensity peak duplication is anticipated, such as with DNA intensity profile. On the other hand, logarithmic scales are preferred to visualise FL channels most of the times, as it facilitates clear visualisation of both dim and bright populations on the same dynamic range. Caution should still be taken when reading linear data on a logarithmic scale via analysis software. In some occasions, the resolution of the histogram binning at the lowest intensity decade could surpass that of the instrument's ADC, which then results in visual artifacts such as picket fencing [10]. In such examples, a linear and logarithmic hybrid scale could solve the problem such as the biexponential Logicle or HyperLog scales, which display pseudo linear scale on the near zero values then transition to pseudo logarithmic scale for the rest [11]–[13]. Nonetheless, most of our data are plotted in either linear or regular logarithmic scales because we have found that any observed artifacts are mostly

associated with low intensity noise channels where usually no biologically significant populations are present.

2.5.3.2   **Population gating: methods and gating strategy**

Population gating refers to the process of applying computed or user defined thresholds on the collected parameters to identify populations of interest, as well as improve the quality of the data before analysis is conducted. Data are often plotted first before decisions are made on the appropriate thresholds. Therefore, proper data compensation (see section 2.4) and visualisation (see section 2.5.3.1) are essential.

The applied thresholds, also known as gates, can be done manually in various shapes, be it a line or a polygon, depending on the data fit. The gating thresholds can also be automated. This has been shown through various tools in the literature either via employing clustering algorithms, such as the k-means algorithm, or fitting curves on well-studied distributions, such as the Dean-Jett-Fox model [14] or the Watson pragmatic algorithm [15], which were designed to recognise diploid cell DNA peaks for statistical readouts (e.g. see a demonstration of the fits in Figure 3-1 in Chapter 3).

The order of the single or multi parameter figures on which gates are applied is often referred to as the gating strategy or hierarchy. It is ideal to have a united gating strategy that can be applied to all the data collected from various platforms. Such standardisation helps reduce variety in the subsequent analysis, even more so if supplied in an automated manner. For example, bias in data filtering could be introduced through manual gating because opinions on the best boundaries and how they should be drawn could be affected by the individual's expertise and/or in-house protocols. Therefore, to ensure similar application of the proposed gating strategy for our CFC data in the current work, we have utilised supervised automated data methods such as DAFi (Directed Automated Filtering and Identification of cell populations) [16] algorithm.

DAFi [16] is an automated gating method for flow cytometry data written on several languages, including R. It enables users to apply unsupervised clustering algorithms such as k-means on all the data Then it applies user-defined rough gates to assign the resulting clusters (via their centroid position) to populations known by the user (see

Figure 2-6). Therefore, it utilises the potential of unsupervised methods to potentially find unknown data clusters, while making the results interpretable through the supervised filtering (i.e. parent populations are defined based on the user input). However, the original version was written to only expect FCS 3.0 files in the script handling of the parameters. It also works with the assumption that only log scale display (thus transform) is desired for FL parameters (which is not the preferred case for DNA profile readout). Therefore, we have introduced in this thesis some necessary code modifications and updates to the original DAFi gating script for R to increase its application flexibility as well as meet our analysis objectives (update highlights are presented in section 3.2.3.2 of Chapter 3).
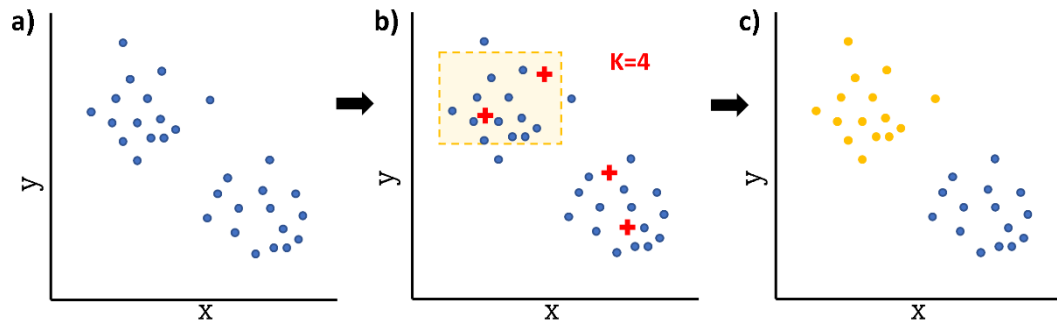
Figure 2-6: A simple scheme summarizing how the DAFi[16] gating programme works on flow cytometry data. The example shows application of a gate (yellow) via the k-means method on some measured parameters x and y. **a)** A scatter plot of some example raw data is presented where it shows two clustered but ungated populations. **b)** The software applies a k-means clustering algorithm to generate centroids (red crosses) of k clusters for the raw data depending on user input (e.g. k=4). The software also takes user-defined rough gates (yellow rectangular) as input, to define the area of a single cluster of interest. **c)** The software combines all k-clusters whose centroid belong to the area of interest and outputs them as a single gated cluster (yellow), regardless of whether all the individual cluster points reside within or outside the user-defined gate boundary.

Both the collected CFC and IFC data in this work share similar gating threshold approaches when it concerns selecting the diploid or polyploid cell populations through their fluorescence. However, because IFC deals with images, this naturally introduces a different approach to the data quality filtering steps in the gating hierarchy. Thus, the population gating in the IFC data of the current work is handled separately and via manual means.

2.5.3.3   Data interpretation

In this section, we highlight some characteristics of FC data that must be considered during analysis. This includes the frequently witnessed data distribution type and how statistical significance is generally handled. Further comments regarding the relevant applied analysis methods for our data set are discussed in either the methodology or discussion sections for each results chapter throughout the thesis.

As commonly seen with biological data, the population distributions of the detected scatter or FL signals in our samples tend to be skewed and not perfectly normal around their mean value. Therefore, robust statistics (e.g. the median reading instead of the mean) and non-parametric statistical tests are used to describe the data. Nonetheless, a normal population distribution can be assumed when describing those of statistical readouts pulled from samples of the same type. For example, the percentages of each cell phase in the DNA profile of control cells are expected to form normal distributions (see Chapter 3). It is also worth mentioning that some of our data distributions could appear lognormal, i.e. the logarithmic transform of the data is normally distributed. Nonetheless, this will not be regarded in the statistical handling of our data because such a method could be considered inconclusive for FC data interpretation [17].

The current work is concerned with observing the scatter intensity changes in our cell samples. One should note, however, that some of the observed intensity shifts may not be biologically significant even if traditional statistical tests regard them as so. For example, when multiple readings are collected for the same sample on the same instrument around the same time, shifts in the intensity distribution between all the collections could still be observed and some statistical tests may flag them as significant. Such intensity shifts are common in FC data and their sources can be attributed to that of the instrument's hardware noise, the sample preparation and handling procedure, or the intrinsic drifts found in all biological samples. While steps are taken to minimise any drifts wherever possible (such as conducting regular instrument calibration, and devising optimal sample preparation conditions), some statistically significant shifts could persist and affect the data interpretation. Therefore, it is important to establish a baseline for biologically non-significant intensity shifts for each parameter on each instrument and/or sample type (e.g. beads or cells). In our research, this is conducted via observing the intensity shifts of repeated measurements

of our control cell samples on the detection channels of interest. These are then used as the baseline intensity difference above which any intensity shifts would be regarded as biologically meaningful (example of similar methods can be found in the literature [18]). Determination of the baseline for biologically relevant intensity shifts also considers noise resulting from spectral overlap that cannot be compensated (see section 2.4). Bead samples are also routinely run along with cell samples on all instruments to monitor any instrument related noise throughout the experiments.

Lastly, because PMT settings are set to keep the same peak positions per channel for control samples on every experiment, we have generally found no significant need in adding the extra step of data normalisation. Moreover, our experimental objectives are concerned mainly with studying the relative peak intensity values in the populations of interest, be it for scatter or FL channels, and not with their absolute intensity values. Though if considered, caution should be taken if scale normalisation is intended to be applied on data from several instruments all at once. This is due to the possible variability in the linear amplification performance of the detectors on the different instruments. This can be checked via running multi-peak control samples across the instruments and observing their relative peak separation. However, we do not recommend it for data that are originally log-transformed, such as those of the FACSCalibur.

2.6   Summary

In the current work, we have mainly used three types of flow cytometers to investigate the label-free characteristics of the target polyploid cells in Osteosarcoma. These are the conventional flow cytometers (CFC) FACSCalibur and FACSVerse, and the imaging flow cytometer (IFC) ImageStream. The chapter describes the instruments and the appropriate data collection and analysis methods through examining the aspects they share, their limitations, and inherent differences, with some examples from the thesis results. These topics are referenced in the results or discussions throughout the thesis. In addition, the presented discussion is expected to assist with any steps taken for the standardisation of experiment design or data replication on other flow cytometry platforms, whether they were benchtop or microdevices.

2.7 References

[1] H. M. Shapiro, *Practical flow cytometry*, 4th ed. New York: Wiley-Liss, 2003.

[2] C. Snow, 'Flow cytometer electronics', *Cytometry*, vol. 57A, no. 2, pp. 63–69, Feb. 2004, doi: 10.1002/cyto.a.10120.

[3] P. J. Smith *et al.*, 'Mitotic Bypass Via An Occult Cell Cycle Phase Following DNA Topoisomerase II Inhibition In p53 Functional Human Tumor Cells', *Cell Cycle*, vol. 6, no. 16, pp. 2071–2081, Aug. 2007, doi: 10.4161/cc.6.16.4585.

[4] P. J. Smith *et al.*, 'Characteristics of a novel deep red/infrared fluorescent cell-permeant DNA probe, DRAQ5, in intact human cells analyzed by flow cytometry, confocal and multiphoton microscopy', *Cytometry*, vol. 40, no. 4, pp. 280–291, Aug. 2000, doi: https://doi.org/10.1002/1097-0320(20000801)40:4<280::AID-CYTO4>3.0.CO;2-7.

[5] P. J. Smith, M. Wiltshire, S. Davies, L. H. Patterson, and T. Hoy, 'A novel cell permeant and far red-fluorescing DNA probe, DRAQ5, for blood cell discrimination by flow cytometry', *J. Immunol. Methods*, vol. 229, no. 1–2, pp. 131–139, Oct. 1999, doi: 10.1016/S0022-1759(99)00116-7.

[6] P. J. Smith, M. Wiltshire, and R. J. Errington, 'DRAQ 5 Labeling of Nuclear DNA in Live and Fixed Cells', *Curr. Protoc. Cytom.*, vol. 28, no. 1, Apr. 2004, doi: 10.1002/0471142956.cy0725s28.

[7] A. L. Hazen, T. Bushnell, and D. L. Haviland, 'The importance of area scaling with FACS DIVA software', *Methods*, vol. 134–135, pp. 130–135, Feb. 2018, doi: 10.1016/j.ymeth.2018.01.006.

[8] L. C. Seamer *et al.*, 'Proposed new data file standard for flow cytometry, version FCS 3.0', *Cytometry*, vol. 28, no. 2, pp. 118–122, Jun. 1997, doi: 10.1002/(SICI)1097-0320(19970601)28:2<118::AID-CYTO3>3.0.CO;2-B.

[9] 'Data file standard for flow cytometry', *Cytometry*, vol. 11, no. 3, pp. 323–332, 1990, doi: 10.1002/cyto.990110303.

[10] D. Novo and J. Wood, 'Flow cytometry histograms: Transformations, resolution, and display', *Cytometry A*, vol. 73A, no. 8, pp. 685–692, Aug. 2008, doi: 10.1002/cyto.a.20592.

[11] C. B. Bagwell, 'Hyperlog — A flexible log-like transform for negative, zero, and positive valued data', *Cytometry A*, vol. 64A, no. 1, pp. 34–42, Mar. 2005, doi: 10.1002/cyto.a.20114.

[12] D. R. Parks, M. Roederer, and W. A. Moore, 'A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data', *Cytometry A*, vol. 69A, no. 6, pp. 541–551, Jun. 2006, doi: 10.1002/cyto.a.20258.

[13] W. A. Moore and D. R. Parks, 'Update for the logicle data scale including operational code implementations', *Cytometry A*, vol. 81A, no. 4, pp. 273–277, Apr. 2012, doi: 10.1002/cyto.a.22030.

[14] M. H. Fox, 'A model for the computer analysis of synchronous DNA distributions obtained by flow cytometry', *Cytometry*, vol. 1, no. 1, pp. 71–77, Jul. 1980, doi: 10.1002/cyto.990010114.

[15] J. V. Watson, S. H. Chambers, and P. J. Smith, 'A pragmatic approach to the analysis of DNA histograms with a definable G1 peak', *Cytometry*, vol. 8, no. 1, pp. 1–8, Jan. 1987, doi: 10.1002/cyto.990080101.

[16] A. J. Lee *et al.*, 'DAFi: A directed recursive data filtering and clustering approach for improving and interpreting data clustering identification of cell populations from polychromatic flow cytometry data: DAFi', *Cytometry A*, vol. 93, no. 6, pp. 597–610, Jun. 2018, doi: 10.1002/cyto.a.23371.

[17] C. FENG *et al.*, 'Log-transformation and its implications for data analysis', *Shanghai Arch Psychiatry*, vol. 26, no. 2, pp. 105–109, Apr. 2014, doi: 10.3969/j.issn.1002-0829.2014.02.009.

[18] F. Lampariello, 'On the use of the Kolmogorov-Smirnov statistical test for immunofluorescence histogram comparison', *Cytometry*, vol. 39, no. 3, pp. 179–188, Mar. 2000, doi: 10.1002/(SICI)1097-0320(20000301)39:3<179::AID-CYTO2>3.0.CO;2-I.

Chapter 3 :

The control, specifications, and handling of the biological samples

3.1 Introduction

The thesis experiments seek to characterise the label-free parameters of cell-cycling polyploid cells in osteosarcoma cells (U-2 OS) through flow cytometry. The parameters of interest are the forward and side scatter intensities as well as the brightfield and darkfield images. However, it is essential to ensure first the stability and reproducibility of the fluorescent biology readout for the examined cell samples on any platform before discussing their label-free readout. This is particularly of importance, considering that the polyploidy in the current work is drug induced, and that the studied live cells can be fragile or sensitive to stress. The method of inducing polyploidy in the U-2 OS cell line with ICRF-193 has been explained in literature along with its underlying biology (where cells bypass mitosis after drug application) [1]. Nonetheless, the inherent non-rigid nature of cells means that perturbations in the biology readouts are possible. Such variations may then extend to their physical characteristics, thus impacting our investigation.

Therefore, this chapter looks to outline the observed biological drifts in the studied control and drug-treated samples. These drifts can either be inherent, caused by the sample handling conditions or instrument noise. Thus, the chapter aims first to define the cell biology specifications for the different types of samples in this work. Next, it will define the optimal sample handling protocol that ensures that said specifications can be delivered each time regardless of the instrument or operator. These aims are directly concerned with the thesis's 1st secondary objective, where such protocol is expected to help produce well-described and reproducible scatter data of the target cells throughout the thesis experiments, and for any work that wishes to replicate the study. The chapter also addresses the thesis's 2nd secondary objective, where it explains the updates we have introduced to an existing gating method to help centralise the gating application of the data collected from different instruments.

3.2 Methodology

In this section, we outline the methodology of the experiments in this chapter, which is also referenced in the next chapter (Chapter 4).

3.2.1 Flow cytometry samples

The flow cytometer samples consist of trypsinised suspensions of live cells in media, as well as beads used for calibration and a cross-platform reference. The subsections below introduce the cells first, then explain the sample preparation procedure.

3.2.1.1 The U-2 OS cyclin B1-eGFP cells

The osteosarcoma U-2 OS cell line (ATCCHTB-96) consist of epithelial cells obtained from the sarcoma of the tibia of a 15-year-old female back in 1964 [2]. The U-2 OS cells in this work are transfected to stably express cyclin B1-eGFP, where the model has been described previously [3]. Cyclins are a family of proteins that form complexes to regulate the cell cycle, where cyclin B1 is associated with the regulation of the G2/M phase. As mentioned previously (section 2.4 in Chapter 2), the levels of cyclin B1 exhibit a slow increase at the end of G1 phase, reaching a peak at the exit of G2 phase/entry of M phase, then dropping sharply when the chromatids separate (anaphase) near the end of the M phase [4]. The used cell model (U-2 OS cyclin B1-eGFP) has been shown to cycle into polyploidy under drug ICRF-193 perturbation [1], where a similar procedure for sample preparation is described next.

3.2.1.2 Samples preparation

The cells are cultured in McCoy's 5a medium with 10% Foetal Bovine Serum, 1% Glutamine, and 1% G418, and the culture is incubated under 5% CO^2 concentration in air at 37°C. For experiments, plates of the cells are cultured in media over the span of four days (unless otherwise stated). To induce a considerable number of cell cycling polyploid cells, the cells are first cultured for 24 hours in 6-well plates or 75-T flasks, then treated with 2µg/mL ICRF-193 (bis[2,6-dioxopiperazine]). In the current work, the duration of the ICRF-193 treatment is either 24 or 48 hours. In the first case, the drug-treated media is washed after 24 hours have passed, then replaced with fresh media where cells are left to recover for 48 hours before being processed by a flow cytometer (sample type is referred to as 24T48R). In the 48-hour case, the same steps are followed, except it is a wash after 48 hours then recovery for 24 hours (referred to as 48T24R). Control cells are cells of the same culture that are not treated with any

drug, but they receive the same wash or media change steps as the drug-treated samples. In other words, two types of control cell samples are prepared every time, but in this thesis, we will only report on the control cell sample for the 24-hour drug-treatment. To monitor the drug response in an experiment and to identify the cell cycle phases, the samples are labelled with DRAQ5 (Biostatus Ltd, Shepshed, UK) at 20 μM to determine their DNA content. Cell viability is also monitored with DRAQ7 (Biostatus Ltd, Shepshed, UK), but it is not reported in the current work.

For instrument monitoring and setting calibration, unlabelled and labelled beads are run along with the cell samples. These are the unlabelled plain polystyrene beads kit of four sizes 4, 6, 10,15 μm (F-13838, Molecular Probes, Inc., USA), the 25 μm (Sigma-Aldrich Inc., UK), as well as the BD Calibrite™ 3-color kit (Becton Dickinson Inc., UK), which includes polymethylmethacrylate beads of approximately 6 μm labelled with either FITC or PerCP. For experiments on the CFC instruments, the unlabelled beads are diluted in DI water at $2 - 4 \times 10^5$ beads/mL, while the FITC and PerCP beads are diluted in BD FACSFlow sheath fluid (Becton Dickinson Inc., UK) at $1.5 - 2.5 \times 10^6$ beads/mL.

3.2.2   Flow cytometers

The samples in this chapter have been run through two different CFC instruments: the analogue (log-amplification) BD FACSCalibur and the digital BD FACSVerse (Becton Dickinson Inc., UK). Both machines illuminate the cells in the flow with a 488 nm argon-ion laser, at 15 mW (a specification, and 13-15 mW actual reported power; gaussian beam spot of 22x66 μm) for FACSCalibur, and approximately 20 mW for FACSVerse (9x63 μm gaussian beam spot). Details on their working principle can be found in Chapter 2. For both instruments, the forward scatter is detected on a silicon photodiode with a 488/10 BP filter, while the side scatter and fluorescence are detected by PMTs. A scheme of the optical path for FACSCalibur can be found in Figure 2-4 in Chapter 2, where the side scatter is detected with a 488/10 BP filter (passing through a brewster-angle beam splitter), CyclinB1-eGFP is detected on channel one (530/30 BP), and DNA-DRAQ5 on channel three (670 LP), and the area and width measurements are turned on for channel three. The optical path is different for the FACSVerse as it uses an optical fibre to direct the collected scatter and emission to a

detector array with bandpass filters and longpass dichroic mirrors for each detection channel, arranged in a heptagon (except for forward scatter, which is detected at a different direction as per usual). For FACSVerse, the side scatter is detected with a 488/15 BP (no mirrors), CyclinB1-eGFP on the FITC channel (527/32 BP after passing a 507 LP mirror), and DNA-DRAQ5 on the PerCP channel (700/54 BP after passing 665 LP).

3.2.3   Cell data gating

The cell biological response to the drug treatment is assessed through monitoring the diploid and polyploid populations, which are identified via their fluorescence readout. The characterisation of cell scatter (Chapter 4) is also mainly based on the gated cell phases in the labelled samples. Therefore, in this section, we discuss the different gating methods and the gating strategy used for the cell data presented in this chapter (and any CFC cell data in the thesis), along with our approach for centralised gating application.

3.2.3.1   Diploid DNA fitting models

Automated fitting models that require no normalisation steps are useful when monitoring the diploid cell cycle readout in control samples across different platforms. Thus, for this purpose, we use here either of the Watson pragmatic model [5] or Dean-Jett-Fox model [6] (mentioned previously in Chapter 2) to apply a fitting on the DNA profile of control samples. Both of these algorithms assume normal distribution fits for the G0/G1 and G2/M peaks. For S phase, the Watson's applies a probability density fit, while the Dean-Jett-Fox's applies a smooth curve fit via combining a normal distribution with a second-order polynomial curve. Figure 3-1 shows an example of both models fitting on a control sample done via FlowJo software (FlowJo, LLC, USA).

Note that these fittings do not count the cyclin B1-eGFP readout, thus the gated diploid cell cycle phases here may differ in percentages from those performed on the two-dimensional scatter plot of DNA-DRAQ5 versus cyclin B1-eGFP (see section 3.2.3.3).

Thus, these are not used in gating cell cycle phases for the analysis of label-free parameters in the next chapter.
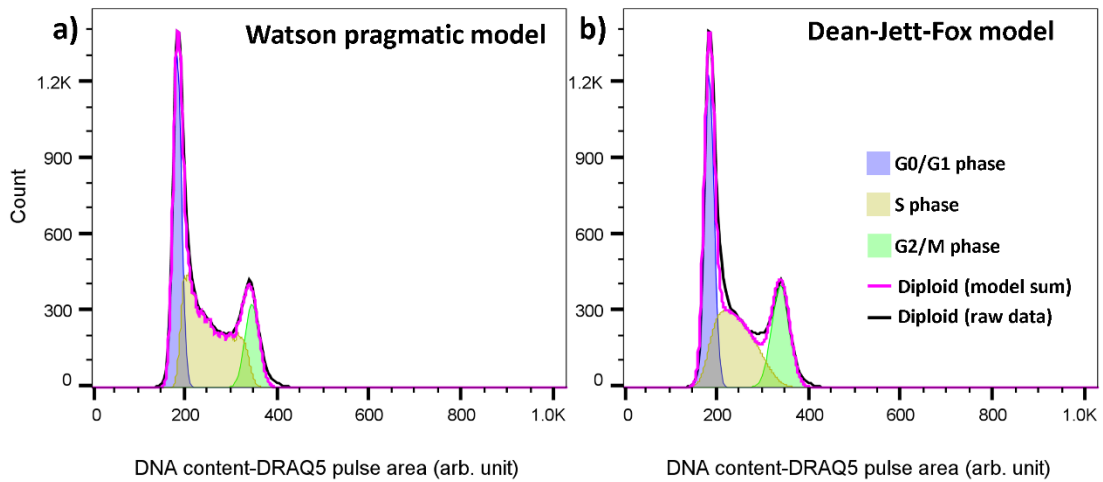
Figure 3-1: Example of a Watson (**a**) and a Dean-Jett-Fox (**b**) DNA profile fitting performed via FlowJo on the same control sample. The sample is collected via FACSCalibur on linear mode, then gated out of debris and doublets using our customised DAFi gating script on R.

3.2.3.2 Automated gating with a customised version of DAFi gating on R

For the CFC data in this thesis, unless otherwise stated, all the analysed cell cycle populations are gated through the application of a customised version, updated by this work, of the DAFi gating script on R [7] (briefly described in Chapter 2 section 2.5.3.2). The updates were essential to make the script suitable for cell cycle analysis (linear scale reading of the DNA), as well as to enable it to handle data from different CFC instruments (e.g. analogue FCS2.0). Thus, the updates help standardise our gating application on the different used instruments, as per the thesis objectives.

In order for DAFi to run, it requires two inputs: an FCS data file and a configuration text file. The configuration file contains information on the user-defined rough manual gates, which are constructed and optimised via data inspection on FlowJo. The coordinates of the gates are then normalised by the user to 200 (a requirement unchanged from original DAFi script) before they are input in the configuration file. This allowed us to gate several FCS files by referencing a single configuration file (or

more than one if an instrumental setting shift was too high to be covered by the rough gates).

The highlights of the customised version of the DAFi gating on R are presented in Table 3-1. In this updated version, scale transformation on output data is applied based on the desired display scale for each parameter in an FCS 3.0 file. For the investigated parameters in this work, that will be applying logarithmic transformations before gating on the Area and Height measurements of cyclin B1-eGFP parameters, while no transformation is applied (i.e. linear) for the DNA-DRAQ5 measurements (width measurements are kept linear for all parameters, scatter is also kept linear for analysis convenience). For FCS 2.0 files, raw data are read as they are with no transformations. This is because raw data in FCS 2.0 are already hardware transformed based on the collection mode (logarithmic or linear) by the hardware (see Chapter 2 section 2.2.3 for more). Within the script, the data binning is kept at 10-bits for FACSCalibur's data but binned at 12-bits for FACSVerse's.

Update highlights	Original DAFi on R	Thesis customised DAFi on R
Accepted FCS files	FCS 3.0	FCS 2.0 and FCS 3.0
User-defined gates	Rectangle gates.	Polygons (up to 6 vertices in current version, and it can be adjusted for more).
K-means cluster size	Single value (for all re-clustering events down the gating hierarchy)	Different values allowed (for any re-clustering event down the gating hierarchy)
Scaling	• Applies log-transformation on all parameters except scatter and time • Axis scaling is fixed to 12-bit (i.e. maximum channel number is 4096)	• FCS 2.0: no scale transformation is applied (raw data are already hardware transformed, see Chapter 2 for more). • FCS 3.0: excludes PerCP channel for cell cycle analysis with DRAQ5 from log-transformation. • Axis scaling is user-defined. (It avoids resolution errors with low bit data)
R version 4 test	N/A.	Compatible.

Table 3-1: Highlights of the thesis customized update on the original DAFi [7] gating script for R.

The output of the original DAFi script is a text file that contains the transformed data each along with information of its gating hierarchy. In our application, the output

consists of FCS 3.0 files of each gated population ready for analysis (available as a python script; see Appendix A). In this thesis, we have kept the transformed output as it is (i.e. exported FCS 3.0 files of filtered populations are not based on the raw data except for FCS 2.0). This explains the apparent linear scaling of most parameters displayed in the thesis, including those that are commonly read on a logarithmic scale for better visualisation (e.g. all axes in Figure 3-2 except for the DNA's). For simplicity, the constructed FCS 3.0 files only consist of essential keywords that allow it to be read by most flow cytometry analysis programmes. Nonetheless, keywords can be customised by the user, for example, to put back information of relative brightness calculation for the original FCS 2.0 files hardware (see Chapter 2 section 2.5.3.1.1 for more).

3.2.3.3   Gating strategy

Unless otherwise stated and other than debris exclusion (where scatter is commonly used), all gating on the cell populations is based on the intensity measurements of the GFP and DRAQ5 fluorescence. This includes filtering for single cells, then for labelled cells or GFP reporting cells, and then finally the different cell cycle phases in each sample type are gated to identify the diploid and polyploid cells in the samples.

Figure 3-2 illustrates these steps via examples of a control (drug-untreated) sample (a-d in Figure 3-2) and drug- treated samples (d in Figure 3-2) on FACSCalibur. The figure shows the gating hierarchy for the studied populations, where automated gating is applied through the customised version of DAFi (plotted via FlowJo for display) (section 3.2.3.2). As plotted, the resulting gates show smooth boundaries based on the predictions of the k-means clustering algorithm. More about single cell gating can be found in Chapter 4 (Figure 4-2). The control drug-free sample (diploid majority) is referenced for establishing the overall diploid user-defined gate, while the 48T24R drug-treated sample (polyploid majority) is referenced for the overall polyploid ones. The overall diploid rough gate is then adjusted to consider the broader distributions of the DNA signal of the drug-stressed diploid cells in the 24T48R drug-treated sample. The rough gates for cell cycle phases are determined by referencing the relative increase of cyclin B1-eGFP signal, along with that of DRAQ5-DNA, for either the diploid or polyploid population.
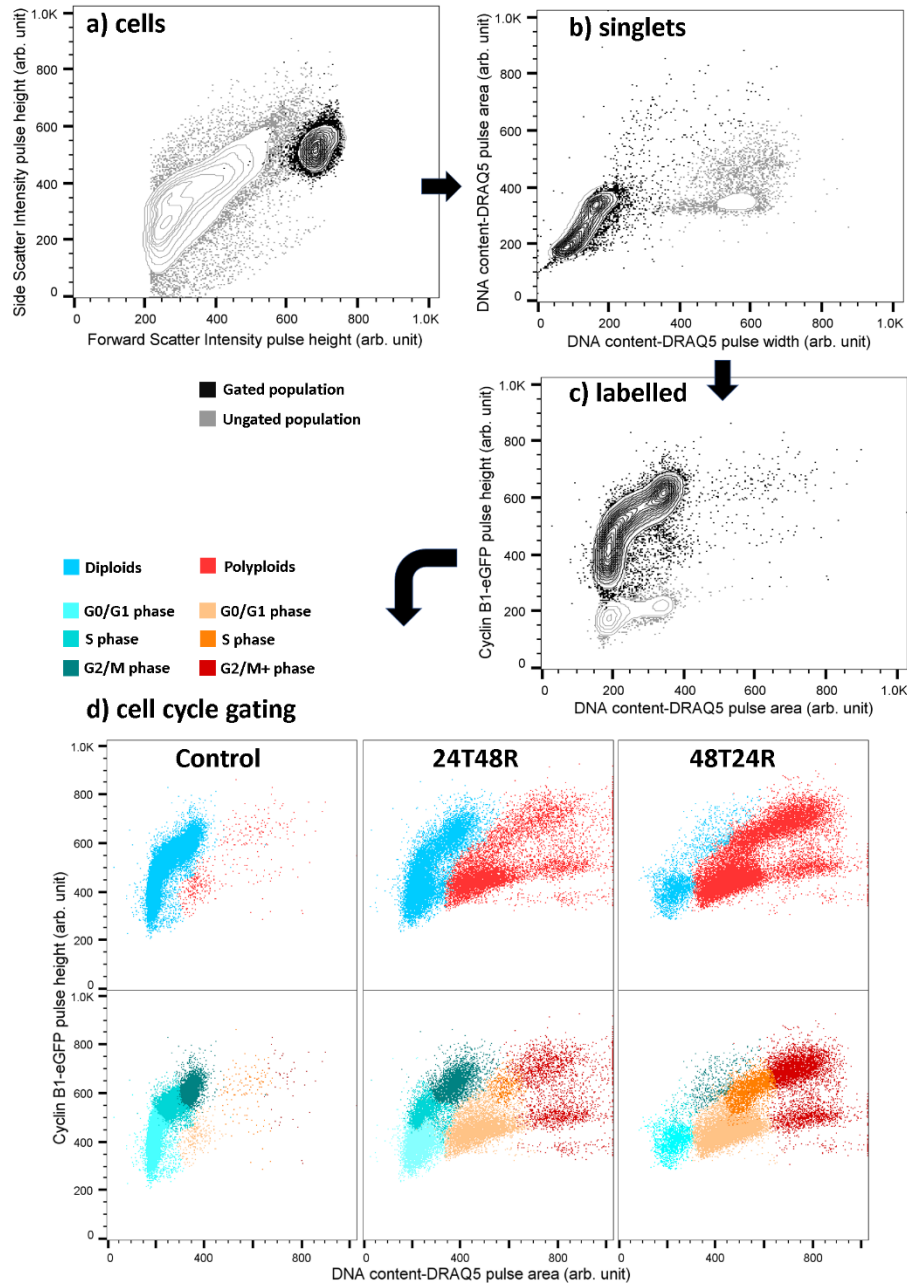
Figure 3-2: General gating strategy followed in flow cytometric data analysis of all cell sample types labelled with DRAQ5-DNA and cyclin B1-eGFP. The first three scatter plots show data in density contour lines with dots as outliers (20 lines at 5% cells spacing). First, **a)** the whole cells cluster (black) is separated from debris (grey) based on forward scatter versus side scatter. Next, **b)** single cells (black) are identified via plotting the DRAQ5 pulse width against its area. The cell cycle readout is then plotted with DRAQ5-DNA against cyclin B1-eGFP where we first filter for **c)** labelled cells (black) and exclude GFP non-reporting (grey) cells. The same parameters are re-used to finally filter for **d)** cell phases. In **d)** the columns refer to sample type (control, 24-hour, or 48-hour drug-treated). The first row shows a gating of the parent diploid (blue) and polyploid (red) populations, while the second row shows gating of their subpopulations. All gating is performed using our customised DAFi gating script on R, where the figure examples are collected on FACSCalibur on logarithmic mode for scatter and cyclin B1-eGFP.

3.2.4   Data analysis

Whenever data pooling in a single (or more) platform is needed, care should be taken on whether some samples can be grouped together. For example, for FACSCalibur, we do not group in our analysis any logarithmically amplified parameters with those that are linearly amplified. This is because of the potential instrument errors that could be present in the log amplification by the analogue circuits (see section 2.2.3 in Chapter 2). However, this does not affect the parameters for the biology readouts, which are consistent in the thesis in terms of data collection modes (linear circuits for DNA content, and logarithmic ones for GFP). Therefore, in all sections of this chapter, we do group samples that are different in terms of scatter-mode collection while favouring consistency in its biological-sensitive conditions, e.g. similar wait time in DRAQ5 (mostly <45 minutes), or any other condition of interest. However, if an observation is made based on scatter intensities, only similar scatter-mode collections are grouped together, such as those reported in section 3.3.2.

Lastly, unless normality can be assumed, the median, the robust standard deviation (RSD), and robust coefficient of variance (RCV) values are used to describe the data (which are commonly reported by flow cytometry software such as FlowJo, see equations 3-1 and 3-2). As previously stated in Chapter 2, these measurements take into account the skewness commonly observed in the data distributions, and are less affected by outliers, unlike normal distribution statistics, which are based on the mean (i.e. an average measurement that can be sensitive to outliers). The RSD is calculated based on the percentile values around the median such that 1 RSD $\approx$ 34.13%, the equivalent of one standard deviation in a normal distribution. In the current work, when comparing results from repeated measurements to verify the reported biology, the analysis is performed on some relevant descriptive statistics of the measured intensity distributions, such as the mean value or standard deviation of the different cell cycle phases. In other words, unless otherwise stated, the entire intensity channel distributions are not fed into the statistical tests (reasons explained in section 2.5.3.3 in Chapter 2). All presented statistical test results have been produced with either FlowJo or IBM SPSS Statistics (IBM Corp., UK).

$$RCD = \frac{(84.13^{th}percentile - median) + (median + 15.87^{th}percentile)}{2},$$

$$where\ 84.13^{th}percentile - 15.87^{th}percentile \approx 68.26\%\ of\ the\ data.$$

3-1

$$RCV = \frac{(84.13^{th}percentile - 15.87^{th}percentile)/2}{median} \times 100\%\ .$$

3-2

3.3 Results

As mentioned previously, there are different factors that contribute to variations observed on repeated experiments of live cell samples. Some of them are based on the experiment environment such as the temperature in which samples are kept before processing, their wait time in tubes, the instrument performance of that day, or the operator preparing the cell samples. Other factors are related to the cell sample's intrinsic nature that may affect the production of regular portions of target cells, such as the possible impact of cell age or cell culture confluency. The amount and health of the target cells can also be affected by drug-related factors during the experiment culture such as the prepared drug dose or the drug application duration.

Control experiments have been conducted to stabilise and minimise effects of said factors on the studied cells' biology. These observations will help ensure the reproducibility of the results on different instruments.

The experiments in this section can be split in three categories: control and specifications of the cell sample, control and specifications of the sample delivery, and general sample data reproducibility. The first covers effects of cell age and drug treatment on the production of target cells, while the one for the sample delivery covers the experimental-environment-based factors mentioned above. Lastly, sample reproducibility discusses the reproducibility of the biology on a single platform or more.

3.3.1 Control and specifications of the cell samples

This subsection first discusses some of the perturbations observed in the fluorescence cell cycle readout of control cells, which can be attributed to intrinsic or drug-related stimuli, or general cell growth conditions. Next, it defines the cell cycle specifications for the drug-treated samples that will be referenced throughout the thesis. The biology

perturbations for the control samples are assessed statistically in the reproducibility section 3.3.3.1.

For the first part, we check how cell age may affect the cell cycle readout per experiment for the cultured cells. The experimental data in this research have been collected throughout this project, where new cells are cultured routinely for consecutive 2-3 months. Thus, checking this factor is critical to the stability of the study's results. In the results, cell age is expressed in terms of the passage number of the original cell cultured, where a passage number refers to the count of the subculturing events (continued from that of the thawed cells).
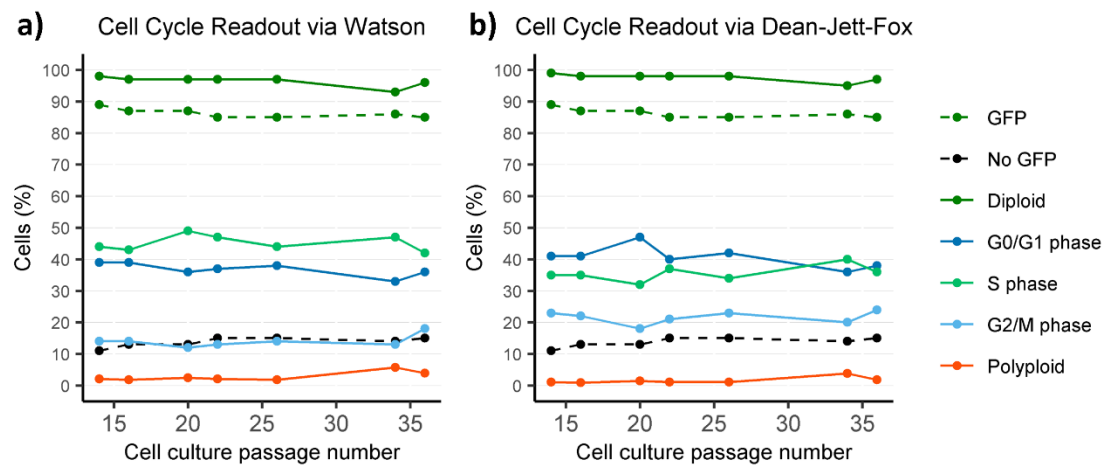
Figure 3-3: Time trend of the cell populations sizes in drug-free control samples collected on FACSCalibur from 7 cell passages (one experiment each). GFP and non-GFP refer to cells that are reporting or non-reporting of cyclin B1-eGFP, respectively, which were gated via the customised DAFi script. G0/G1, G2/M, S, diploid (their sum), and polyploid are the cell cycle phases extracted either via Watson (**a**) or Dean-Jett-Fox (**b**) fitting models for cell cycle.

Cells in different cell cycle phases have been monitored over time for control cell samples (i.e. drug-free samples), to establish the expected cell cycle behaviour, without adding the drug treatment as another variable. This is plotted in Figure 3-3, where the cell cycle readout is estimated via either the Watson or Dean-Jett-Fox models. In total, the presented data set consists of measurements of labelled samples from 7 cell passages (one experiment each). The figure shows how the fitting of each model can estimate the readout of the DNA cell cycle (expressed in solid lines) differently, e.g. low G2/M and high S phase estimates in Watson's versus higher G0/G1 phase estimates in Dean-Jett-Fox's. The dashed lines refer to the GFP reporting

cells (green) versus non-reporting ones (black), which are gated via our customised DAFi script and plotted for each model as a reference.

Overall, neither model estimation showed any considerable changes (e.g. over 10%) in cell cycle over the span of 22 passage numbers. We took notice of two points at which some changes in the line trend have been observed, which are at passage numbers 20 and those above 30. For the first point, the experiment samples were cultured from a stock culture confluency of 50% (while the rest from approximately 70-80%), which is a variable that is generally well-known to potentially affect cell growth rate [8]. For the two measurements at the later passage numbers (34 and 36), these belong to a 48-hour in well plate cell cultures (i.e. instead of the usual 4-day in culture in the rest of the experiments). This is also another variable that could impact cell growth rate and makes it incomparable to the rest. No changes have been observed in the cells expressing the GFP label throughout the reported 22 passage numbers. In general, keeping a steady cell cycle measurement ensures that we have no changes in the input samples for the drug-treatment, thus we can expect similar drug-response output each time. Therefore, it can be inferred that cell age is not a concern to the biology reproducibility of the samples as long as cells are cultured in the same conditions (i.e. regarding stock confluency or culture duration), or at least for the reported time period. In our work, no data were collected for cell passages beyond the reported duration.

In the second part, we check the drug response through the cell cycle readouts in drug-treated samples. This is used to define the target cell sample specifications that will be observed on different platforms. These numbers could be affected by the drug dose (fixed for the current work at 2 µg/mL) or dose application period (either 24 or 48 hours). The top pie chart in Figure 3-4 (a) shows the mean percentages of relevant populations in cell samples that have been treated with the drug for 24 hours. This has been constructed from samples labelled with both DRAQ5 and GFP from three experiments (where approximately 20K cells were analysed each after excluding doublets and GFP-nonreporting cells). The 24-hour drug-treatment has reduced the percentage of diploid cells to be 57±0.4% of the entire population, unlike control cells that have a majority of diploid cells at 97±1% (based on DAFi gating on the same samples in Figure 3-3). The polyploid cells at the G2/M phase of cell cycle or at the

second ploidy cell cycle (i.e. the target cells that have shown resistance of the drug through a continuous cell cycle) make up approximately 10% of all the analysed single cells (highlighted in a red rectangle in Figure 3-4).
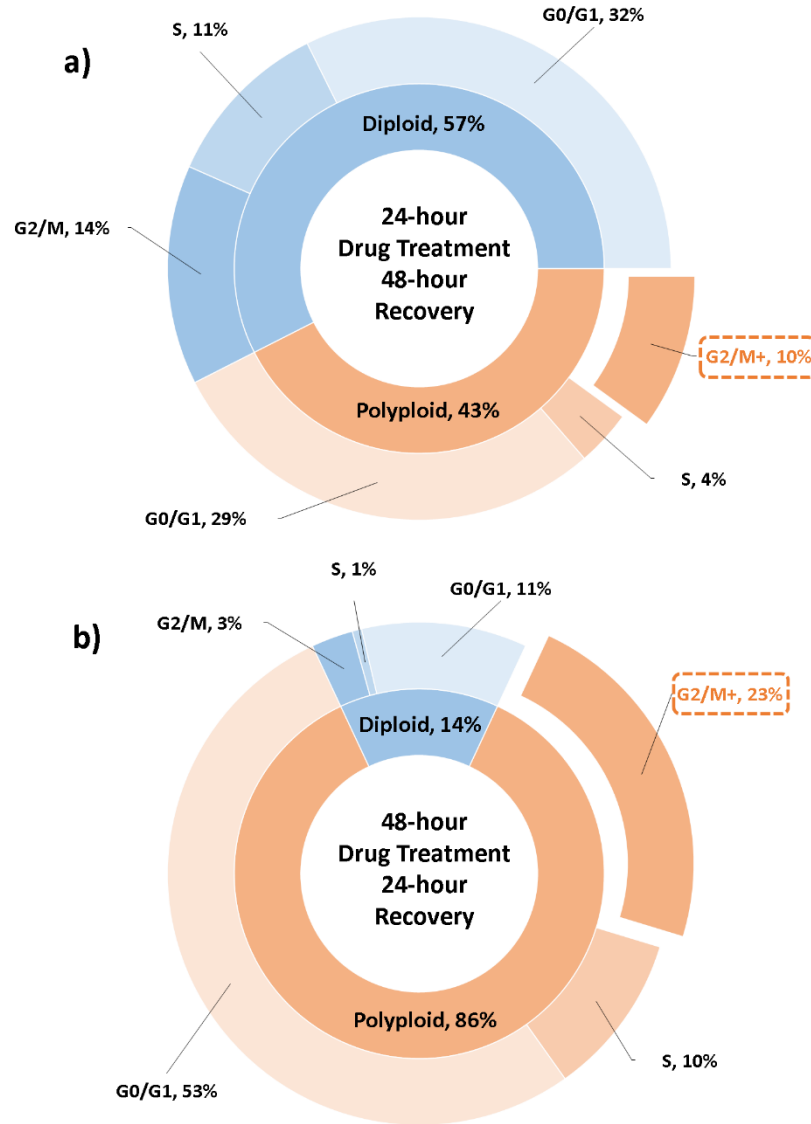
Figure 3-4: Specifications of the target cells (highlighted in red rectangles) and the overall cell cycle presented as average percentages in the prepared cell samples that underwent ICRF-193 drug treatment for 24 (**a**) or 48 hours (**b**), and allowed 48 or 24 hours recovery, respectively. The figure presents a hierarchy of cell populations, starting from the inner circle with cells gated according to the nature of cell cycle (diploid or polyploid), and the outer circle representing their cell cycle phases. For each chart, all percentages are relative to the total number of gated singlets GFP-reporting cells (approximately 20K cells per experiment), which were taken from three repeats (passage numbers were 14, 16 and 22). Error margins aren't reported in the graph.

The percentage of the target cells can be improved by leaving the cells for one more day in the drug (a total of 48 hours). It results in approximately 23% of cells at the G2/M+ phase of the polyploidy cycle, as the overall polyploid cell population increases from less than half to be approximately 86% of the cell sample as shown in b in Figure 3-4.

In general, we have observed low numbers of cells clustering at the S phase within parent populations of low overall density, e.g. less cell density at S phase in the diploid population of the 48T24R drug-treated samples (see d in Figure 3-2). Because such cells are more dispersed and less clustered around each other, a k-means clustering method may struggle to assign a centroid for these cells within their user-defined cell gate (instead, they could be assigned to centroids within neighbouring cell phase gates). As a result, the (k-means based) automated gating with DAFi may still miss some of these S phase polyploid cells and count them as G0/G1 phase polyploid cells. We have kept it as it is in the reported Figure 3-4, because the bias in the automated method appears nearly similar. Moreover, we are mainly concerned with monitoring the overall diploid versus polyploid populations when checking drug-response, as well as our target cells. Therefore, we only adjust the input of the configuration files for the automated gating when these populations are not finely gated, or if an overall gating error exists.

All in all, while Figure 3-4 may not be taken as absolute measures of cell phase proportions, these can still be referenced in monitoring general cell sample specifications provided that the same gating method is used for any analysed new data.

Lastly, we briefly mention that consideration has been taken whenever new materials are acquired during the study. For example, newly prepared drug aliquots are tested against previous ones first before further experiments are conducted, to ensure stability of the prepared dose.

3.3.2 Control and specifications of the sample delivery

This section discusses some external sources that may affect the results of fluorescence (and subsequent scatter) measurements during the sample data collection. In other words, we discuss variations observed in the results that could be attributed to a change

in the studied sample's biology due to stress imposed by external sources directly or indirectly. In addition, we discuss variations that are attributed to purely instrumental-based effect such as the instability in the power of the illumination laser.

3.3.2.1   Sample wait time in room temperature versus on ice (5ºC)

Here, we look at the proper storage or handling conditions for the live biological samples in the current work during experiment sessions. Proper conditions are those that allow minimum irregularity in the measured cell scatter or biology. The variables studied are those of time: 0.5-2 hours (to accommodate sample transport time whenever needed), and of temperature: room temperature (RT) (21-23ºC) versus slowing the sample's biology via keeping them on ice (5ºC). This is done in a single experiment setting to exclude effects of other external variables. The parameter measurements that will be looked at are the forward and side scatter intensity collected in linear signal amplification mode on FACSCalibur for approximately 30K cells gated out of debris. Specifically, changes in the shape and size of the scatter distributions will be observed, as they can be taken as indications of unstable biology or overall sample quality (i.e. increase of cell aggregations). The control sample (for changes in either temperatures or time duration) will be a freshly prepared (wait time < 20 minutes) sample at RT (common procedure in all cell sample preparations).

In Figure 3-5, the population profiles of the cells that were kept in RT for more than one hour (orange and red) appear different from that of the fresh control sample in green. On the adjacent histograms, a second intensity peak can be seen emerging at higher forward scatter intensity channels for these samples. This is unlike the control sample's, which shows single peak scatter distribution. Higher forward scatter intensity can be attributed to increase in cell sizes, thus a potential increase in cell aggregations for samples of the same phenotype. Because the live cell samples in the current work can be fragile to external stress, especially when drug-treated, we generally avoid the application of any extra steps that could minimise aggregations such as size filtering. Other than reducing the analysis yield, aggregations can generally impact the quality of the rest of the cells. Thus, any handling conditions that increase aggregations are less desirable. For side scatter, the values of the RCV in all samples show larger distribution spreads across the higher intensities as cells wait

longer in RT. This is an indication of increased cell granularity that can be attributed to cell death, which is naturally a non-desired outcome.

On the other hand, the sample kept on ice (blue), for the same time duration as the orange RT sample (i.e. 1.7 hour), shows a similar intensity cluster shape (i.e. no additional peaks) as that of the fresh sample, albeit shifted. The RCV values of the side scatter intensities in both the fresh and on-ice samples are the same (30.6%), compared to the more spread RT sample (at 45%). This suggests that during a relatively long wait time there has been no concerning changes in the cell scatter, hence its biology.

	Subset Name	Median : FSC-H	Robust CV : FSC-H	Median : SSC-H	Robust CV : SSC-H
	<0.3hr wait in RT	376	15.0	162	30.6
	1.7hr wait on Ice	319	16.3	134	30.6
	1.7hr wait in RT	417	18.1	171	45.0
	2.7hr wait in RT	408	19.5	167	50.3
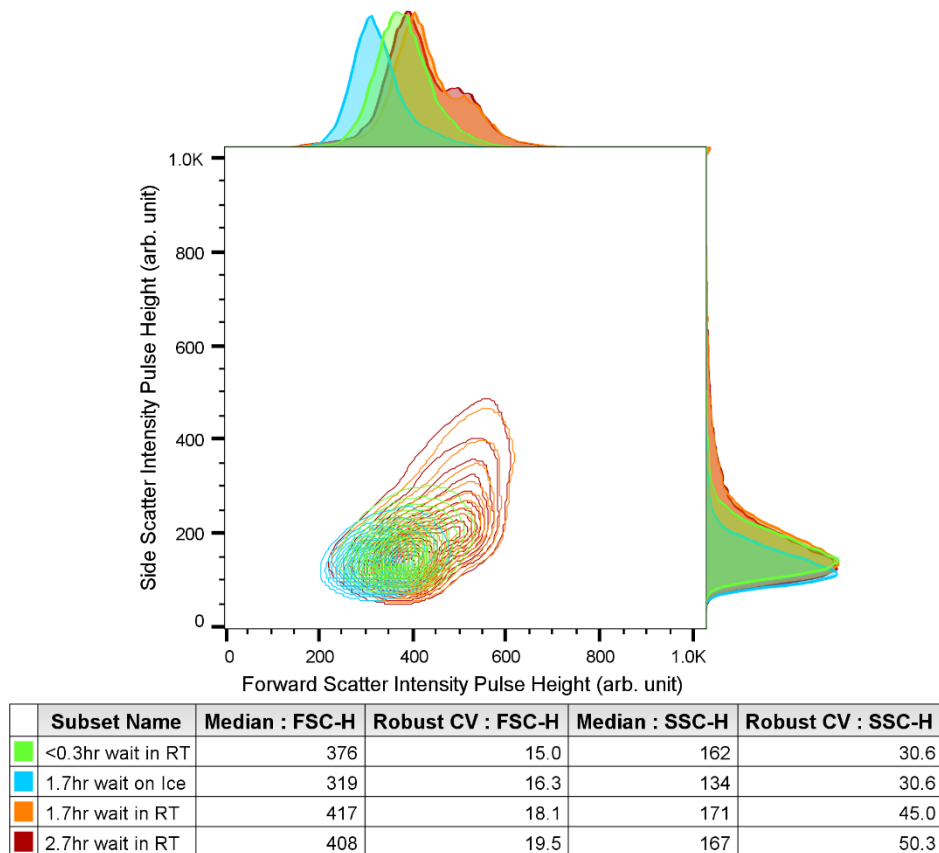
Figure 3-5: Effect of wait time and temperature on delivery of live cell samples displayed as contour plots (20 lines at 5% cells spacing). Adjacent histograms are also added for the forward scatter (horizontal) and side scatter (vertical) intensities. Data were collected on FACSCalibur on linear mode.

In either case, whenever possible, it is generally recommended that live cells be analysed as soon as they are prepared. Otherwise, keeping them on ice for up to 2 hours is an acceptable threshold regarding preserving their scatter and health for the experiments in the current work. This threshold is mainly for DNA-unlabelled samples, because cell health and viability for DNA-DRAQ5 labelled samples tend to

decrease with time (due to DRAQ5's irreversible binding to the cellular DNA). It should be noted that the current work does not investigate any potential long-term effects on cells preserved on ice for such long periods, which may impact cell health if they are brought back for culture.

On a related note, we have found that any extra centrifuging steps during sample preparations, mainly for DNA-DRAQ5 labelled samples, can impact the cell cycle readout significantly. This was noted via the broadening of the cell cycle profiles of control samples, and the appearance of cell fragmentation (<2N DNA populations) in the drug-treated samples (data not presented). Thus, reducing any mechanical stress during the sample handling is overall recommended.

3.3.2.2   Instrument performance

Here we discuss observations we have made regarding the instrument performance stability that can affect the cell analysis. Specifically, we will discuss the uncontrollable illumination power fluctuations that can be encountered in some flow cytometers.

The recommended instrument-specific Quality Control (QC) tests have been run routinely in the duration of the experiments discussed in this thesis. Nonetheless, passing QC tests is not always an indication on the stability of the instrument performance during the experiment session itself. Therefore, labelled and unlabelled calibration beads have been run as well before and after each experiment session to monitor the stability of the measurements throughout the experiment.

The real-time power readings of the illumination laser are only available on some flow cytometers like the FACSCalibur, and they are not stored in metadata of the exported data files. It is also not a common practice in the FC field to keep track of them. In FC studies, where researchers are mostly interested in the FL readout of biological samples, this type of observations tend to be less of a concern. This is because of the tendency for some of the biological FL emissions to be less sensitive to small illumination power shifts (e.g. shifts of sub milliwatts). On the other hand, cell scatter can be thought of as a mirrors system, where the intensity of the output light is directly

affected by the intensity of the input. Thus, minimal power shifts can be detected on sensitive detectors like the PMT on flow cytometers.

To illustrate this point, we discuss here the measured scatter intensity shifts we have observed for the labelled and unlabelled calibration beads that were run before and after experiments. During which, the real-time illumination power readings for each sample have been recorded and averaged, and readings have been collected over 4 consecutive days. Figure 3-6 shows how small fluctuations in the illumination power on FACSCalibur correlate to observable peak shifts in the scatter of the calibration beads. For the scatter data set in the figure, these shifts are considered statistically significant ($p < 0.01$) when applying the chi-square test on the entire scatter distribution (via FlowJo). For the same illumination power shifts, shifts in intensity have also been observed in fluorescent calibration beads, but not much so with the fluorescent readout of the biological samples (data not shown).
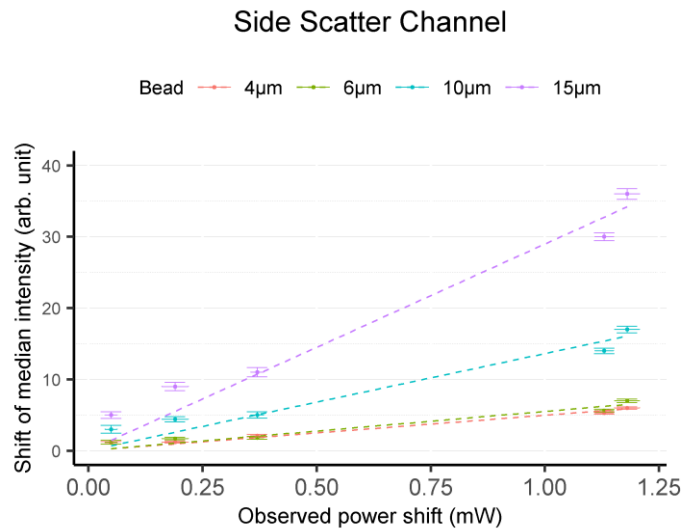
Figure 3-6: Shift in the median intensity channel in the side scatter of four polystyrene beads against the shifts in observed illumination laser power. Scatter data are collected over different days on FACSCalibur using same detector settings (logarithmic amplification mode, transformed to linear before calculating shifts). Error bars are based on the standard error.

Fluctuations in illumination power as low as observed in Figure 3-6 can theoretically cause detectable reduction on the PMT of the scatter signal, and we have observed that experimentally as well. Stability in the detection of light scatter is a topic of concern in this thesis. Nonetheless, it is not practically possible to separate the effect of power shifts from those of other noise sources in the instrument. For example, shifts in the

detected bead intensities have also been observed in the side scatter measurements collected on two different days with the same observable power and detector settings (data not shown). This suggests the existence of another noise source, a detection-circuitry-based one, that is statistically significant.

Therefore, regardless of the suspected source of noise, the data in Figure 3-6 can help in counting for such noise factors by providing us with the baseline for significant intensity channel shift. In other words, any observed intensity shifts in the control beads' measurement will also be used to define the threshold for a meaningful statistically significant intensity shift in experiment analysis (i.e. an intensity channel difference is considered meaningful if shift > threshold, otherwise it is dismissed as a potential instrument noise (shift < threshold) even if they were statistically significant). Naturally, shifts observed in the control cell samples (be it for FL or scatter channels) are also considered with those of beads when establishing the baseline for significance in intensity shifts.

3.3.3 Reproducibility

Reproducibility refers to the ability to repeatedly acquire our reported observations and results following the specified methodology. For this work, it can be looked at in two ways. The first one is reproducibility of the biology itself within a single platform, which also helps observe the tolerable drifts in the biological sample specifications. The second is reproducibility of said sample across multiple platforms. This refers to successfully meeting the biological sample specifications presented by the previous steps on other platforms. For both, we validate this through applying statistical tests on repeated measurements of the drug-free FL labelled cell samples, as they are the biological control for the FL measurements of all cell samples in our experiment design (similar to how beads are the control samples for the scatter measurements).

3.3.3.1 Biological sample reproducibility

In this section, we examine the reproducibility of the sample biology as described through cell cycle readout. A reproducible and stable cell biology acts as a foundation

for reproducible and stable drug-response as well as label-free readouts throughout the thesis experiments.

This has been touched upon earlier in section 3.3.1 where the specifications for the different cell samples in this work were defined through the observed relative population sizes of the different cell cycle phases. Here, we expand on that by testing the significance of any perturbations observed in repeated measurements of control cells on a single platform. However, as mentioned in the Methodology section, we will present here descriptive statistics (instead of the entire data distribution) of only some selective features of the measured parameters.

A total of 9 measurements (variables) are drawn from five out of the seven experiments previously presented in section 3.3.1 (where the experiments of the two late passage numbers are excluded). We also add three more experiment repeats, so a total of nine (n=9) different readings on the FACSCalibur for each variable. We have chosen to use those extracted from the Dean-Jett-Fox model because they are closer to the cell cycle estimations of our DAFi gating (e.g. both estimated S phase populations to be less than G2 phase ones, unlike the Watson model [see Figure 3-3 and Figure 3-4]). The variables are summarised in Table 3-2 for the nine experiments. Other than the cell phase proportions, these include the diploid DNA peaks linearity (i.e. ratio of G2/G1). This is calculated from the mean channel values for the G0/G1 and G2/M peaks, predicted by either of the Watson or Dean-Jett-Fox fittings, for each data point.

feature	mean (%)	SD (%)	exact p-value (2-tailed)
GFP	86.6	2.9	0.294
No GFP	13.4	2.9	0.294
G1	42.2	3.0	0.700
S	33.7	2.7	0.977
G2	22.1	2.2	0.588
G2/G1	1.83	0.02	0.201

Table 3-2: Summary of a selection of features representing the biological DNA readout of control samples collected from 9 repeated FACSCalibur experiments. Reported significance is the exact p-values corrected for ties.

Our null hypothesis for this test is that the distribution for each of the nine variables, as acquired through several sample collections, follows a normal distribution. To test the hypothesis, we run the non-parametric one-sample Kolmogorov-Smirnov test via the SPSS software. The test results (also in Table 3-2) showed that none of the

variables deviated from the normal distribution (i.e. p-value > 0.05), thus we can retain the null hypothesis. From which, we can conclude that the reported biology drifts in the cell biology are not statistically significant to the reproducibility of the biology on this instrument. Results are also checked with other normality tests such as Shapiro-Wilk's (i.e. all variables reported p-values > 0.05).

3.3.3.2   Sample reproducibility across platforms

Reproducibility across platforms is tested through repeated experiment days with control cell samples, which have been cultured with the same conditions and run on the same day on two different CFC instruments, FACSCalibur and FACSVerse.

The purpose of the analysis performed here is to answer the question of whether there is any statistically significant difference between the regular cell biology reported by each platform. This is done through testing the hypothesis that the mean values of each measured variable are the same for both instruments. Here we use the same nine variables introduced in the previous section (3.3.3.1), but through three experiment repeats on either instrument. Because the test sample size is only n=3 for each variable (corresponding to one result from each day), the normality assumption cannot be held. Therefore, we use the non-parametric version of the independent samples t-test, which is the Mann-Whitney U test.

A summary of the results is presented in Table 3-3 and it shows that all the samples' features have scored p-value > 0.05, i.e. retaining the null hypothesis, with only the G2/G1 linearity reporting a statistically significant difference between the two machines.

feature	total mean (%)	total SD (%)	asymptotic p-value (2-tailed)
GFP	85.5	0.5	0.754
No GFP	14.5	0.5	0.754
G1	34.2	4.7	0.251
S	38.6	2.5	0.251
G2	24.3	3.0	0.917
G2/G1	1.84	0.01	0.005

Table 3-3: Summary of a selection of features representing the biological DNA readout of control samples collected from 3 repeated experiments on FACSCalibur and FACSVerse. Reported significance is the asymptotic (approximated) p-values.

Generally, cells in the G2 phase of the cell cycle are expected to have double the DNA contents of that of cells in the G1phase, i.e. the ratio of G2 on G1 would ideally equal 2 for human cells. For the reported three parallel experiments, the diploid peak linearity had mean values of 1.83 and 1.85 for the FACSVerse and FACSCalibur, respectively. Because both are set at the same flowrate during collection, we believe that this difference is instrument based, and not a consequence of a biological change or potential under-labelling. Naturally, a larger sample set would confirm these assumptions. This is supported by how the reported values earlier (Table 3-2) show that the G2/G1 linearity for FACSCalibur is more disperse when more data were collected, with a mean (1.83±0.02) that approaches the one reported for FACSVerse here.

In conclusion, we can state that the reproducibility of the cell sample biology across the two platforms has been achieved as relevant sample specifications have been met without any significantly meaningful differences.

Lastly, it is worth mentioning that using DNA QC kits (e.g. using Chicken erythrocyte nuclei [CEN] or calf thymocyte nuclei [CTN]) can help monitoring DNA peaks profile linearity and resolutions on different instruments. However, these are not considered in the control sample panel for the current work because of logistical constraints. In addition, the bimodal DNA profile for our control diploid cell samples is sufficient to observe any instrument changes, just as has been reported here.

3.4   Biological sample handling protocol

Based on the chapter's results, and as an answer to the 1st secondary objective of the thesis (see Chapter 1), we outline here our biological sample handling protocol for the rest of the experiments in this thesis. In general, the proportion of cell phases in the cell-cycle readouts can be affected by the cell culture confluency, be it in stock culture or in an experiment's plates. This could impact the percentage of the diploid and polyploid populations in the drug-treated samples. Thus, we recommend monitoring this factor if a steady cell population numbers is desired, and to have a better assessment of the drug-response (where underdosing can affect the outcomes as well). During experiments, non-ideal circumstances for cell survival and data stability may

be encountered such as cells spending long period of time at room temperature and risking changes in their biology. Therefore, all cell samples will be put on ice (~5ºC) to slow their biological activity, after they are prepared or incubated for any labelling. For long flow cytometer experiments (such as those on the IFC), it is recommended to process DRAQ5-labelled samples first before the unlabelled samples due to the cytotoxicity of DRAQ5, which can impact cell health negatively with time [9]. In addition, during the cell preparation itself, it is desirable to minimise any additional external stress such as multiple centrifuging steps. To monitor potential instrument noise, we recommend running suitable bead samples on all detection channels before and after the session (in addition the initial QC tests). These should help establish the baseline for meaningful statistical significance in the detections (along with the drifts observed in control cell samples). Whenever samples are run on new instruments, a parallel one should be run on a different instrument to ensure that the cell biology readout is well-known and within the desired specifications (defined in the previous sections). This protocol should assist any future works aiming to replicate the thesis results with the studied live cell samples, be it for label-free detection or sorting, on benchtop FC instruments or even microfluidic devices.

3.5   Summary

This work is concerned with studying label-free parameters of the cell-cycling polyploid cells. Nonetheless, it is crucial to ensure first that any such observation is an actual characteristic of the target cells, and not a natural drift of the sample biology or a noise in the instrument. Therefore, this chapter addressed these concerns by showcasing our control measures where experiments were conducted to observe different relevant variables in the cell experiments, and how they may affect the samples. Next, the desired biology specifications of the cell samples were defined through repeated measurements on a single platform. Then, reproducibility of these results has been observed on two CFC instruments. The automated DAFi gating method has also been optimised for the current work to assess in the centralisation of cell cycle analysis on the diploid and polyploid cell populations. The chapter concluded by summarising the findings through proposing a cell sample handling

protocol that aims for stable and reproducible cell sample specifications on any platform.

3.6   References

[1] P. J. Smith *et al.*, 'Mitotic Bypass Via An Occult Cell Cycle Phase Following DNA Topoisomerase II Inhibition In p53 Functional Human Tumor Cells', *Cell Cycle*, vol. 6, no. 16, pp. 2071–2081, Aug. 2007, doi: 10.4161/cc.6.16.4585.

[2] 'U-2 OS - HTB-96 | ATCC'. https://www.atcc.org/products/htb-96 (accessed Jun. 28, 2022).

[3] N. Thomas *et al.*, 'Characterisation and Gene Expression Profiling of a Stable Cell Line Expressing a Cell Cycle GFP Sensor', *Cell Cycle*, vol. 4, no. 1, pp. 191–195, Jan. 2005, doi: 10.4161/cc.4.1.1405.

[4] N. H. Lents and J. J. Baldassare, 'Cyclins and Cyclin-Dependent Kinases', in *Encyclopedia of Cell Biology,* Elsevier, 2016, pp. 423–431. doi: 10.1016/B978-0-12-394447-4.30057-8.

[5] J. V. Watson, S. H. Chambers, and P. J. Smith, 'A pragmatic approach to the analysis of DNA histograms with a definable G1 peak', *Cytometry*, vol. 8, no. 1, pp. 1–8, Jan. 1987, doi: 10.1002/cyto.990080101.

[6] M. H. Fox, 'A model for the computer analysis of synchronous DNA distributions obtained by flow cytometry', *Cytometry*, vol. 1, no. 1, pp. 71–77, Jul. 1980, doi: 10.1002/cyto.990010114.

[7] A. J. Lee *et al.*, 'DAFi: A directed recursive data filtering and clustering approach for improving and interpreting data clustering identification of cell populations from polychromatic flow cytometry data: DAFi', *Cytometry A*, vol. 93, no. 6, pp. 597–610, Jun. 2018, doi: 10.1002/cyto.a.23371.

[8] V. Noel and M. D. Berry, 'Culture of Adherent Cancer Cell Lines', in *Cancer Cell Biology*, vol. 2508, S. L. Christian, Ed. New York, NY: Springer US, 2022, pp. 19–29. doi: 10.1007/978-1-0716-2376-3_3.

[9] P. J. Smith, M. Wiltshire, and R. J. Errington, 'DRAQ 5 Labeling of Nuclear DNA in Live and Fixed Cells', *Curr. Protoc. Cytom.,* vol. 28, no. 1, Apr. 2004, doi: 10.1002/0471142956.cy0725s28.

Chapter 4 :

Scatter-based label-free detection of cell-cycling polyploid cells in osteosarcoma

4.1 Introduction

The first hypothesis of the thesis states that cell-cycling polyploid cells in osteosarcoma can be detected through their high forward and side scatter intensities in heterogenous samples of diploid cells. This is postulated based on the understanding that polyploid cells tend to be larger in size and organelle numbers, which may contribute, on top of other potential physiological changes [1], [2], to increases in cell scatter, as discussed in Chapter 1. In the previous chapter, we have mentioned that when osteosarcoma cells are treated with a DNA topoisomerase II drug, such as ICRF-193, some cells become polyploid (acquiring twice or more the regular diploid genetic content) as they undergo endoreduplication, i.e. DNA multiplication without cell division [3]. We have also illustrated the common detection method of the polyploid cells, which relays on fluorescence labelling of the DNA content; a method that is not compatible with long-term studies on live cells.

In this chapter, we aim to experimentally test the first hypothesis. The chapter also aims to provide scatter-based detection specifications for the target cells, which can help in target cell sorting on commercial systems. Firstly, the forward and side light scatter is characterised for the entire cell sample. This is done by studying the scatter of the cells in different cell cycle phases for both drug-free and drug-treated samples. Next, scatter detection thresholds of the target cells are recommended by referring to the discussed relative scatter intensity changes between the different phases, as well as that of a standard reference. Lastly, the efficiency of said method will be estimated via applying an automated data gating method to find the cell-cycling polyploid population out of the collected flow cytometer results.

All results are based on experiments collected on the FACSVerse (rescaled to 12-bit channels) using the same materials and methods as those presented in Chapter 3 (section 3.2). Unless otherwise stated, the displayed results are concluded from three repeats of the cells scatter experiment, where stability of the sample biology and sample delivery are observed according to the conditions established in Chapter 3.

4.2 Light Scatter characterisation of the osteosarcoma cells

This section looks at the characteristics of the forward and side scatter of the studied U-2 OS osteosarcoma samples, including the target cell-cycling polyploid cells. The characterisation is performed through comparison with a control sample, which is drug-free, i.e. a sample that consists mostly of diploid. In addition, through fluorescence, cells are divided into their respective diploidy or polyploidy G0/G1, S, and G2/M cell cycle phases. This helps us identify our target cells, the cell-cycling polyploid cells, (denoted as cells at the G2/M+ of the ploidy cell cycle), to examine their scatter properties.

In the first subsection, we look first at the overall (forward and side) scatter of control samples against that of the drug-treated ones, where polyploid cells make up half or more than half the entire sample. In next subsection, we look closely at the overall scatter of diploid cells in both control and drug-treated samples. Similarly, we discuss in the following subsection the forward and side scatter of polyploid cells, including the target cells. We start it by looking at the overall scatter of diploid versus polyploid cells within the drug-treated samples. Then we check how the scatter changes throughout the cell cycle phases. We also look at how the height and width measurements of the detected light scatter pulse contribute to the pulse's area for both the forward and side scattered light, and how this relates to utilising both scatter directions in the detection of polyploid cells.

4.2.1 Overall scatter of whole cell samples at different stages of drug-treatment

The first result plot at which researchers commonly look during a flow cytometry data collection session is the dot plot of the forward and side scattered laser light. It is mostly used for debris exclusion, but also can be used to observe any irregularity in the cell sample (as illustrated in Figure 3-5 of Chapter 3). In the following subsections, we look at this data plot to observe how the presence of various doses of polyploid cells (different stages or durations of the drug-treatment) make notable changes to the overall scatter of samples. In addition, we also address the potential effect of DRAQ5, the DNA label that is used here to classify the cells in their respective position of the

cell cycle. Lastly, we look at doublet exclusion via scatter, as commonly practiced in flow cytometry.

4.2.1.1   Light scatter of DRAQ5-unlabelled samples

In this section, we present the three different types of the cell samples whose biology was discussed in Chapter 3: a drug-free control, a 24-hour drug-treated cells that were allowed to recover for 48 hours after drug wash, and a 48-hour drug-treated cells with 24-hour recovery. We have also added a fourth sample: a 24-hour drug-treated cells that were given no recovery time, to represent the intermediate stage in treatment between the first and second samples (listed previously). Unlike the two other drug-treated samples, this fourth sample is still under the drug arrest at G2/M phase. In other words, its cell cycle readout shows that the cells are either at the diploidy G2/M phase or polyploidy G0/G1 phase (both have same DNA count) [3]. All the presented samples in this section are DRAQ5-free, to observe their scatter without the potential stress of the label.

The scatter profiles of these samples are presented as density plots in Figure 4-1, where the colours red, yellow, green and blue, represent the shift from regions of high to low cell concentration, respectively. To examine further the changes in the forward and side scatter for each sample, the scatter profiles are replotted in a single figure (see Figure 4-2), where the intensity histograms are co-presented for both forward and side scatter.

From the figures, we first have the control (diploid) osteosarcoma cells setting mostly in a dense population at the relatively lower end of the intensity scale (see a in Figure 4-1, and green population in Figure 4-2). This changes in the next sample (24-hour drug with no recovery), where the overall light scatter shifted up in intensities for both the forward and side directions (see b in Figure 4-1, and blue in Figure 4-2). Such observed scatter shifts seem to agree with the equivalent transition in biology introduced above. Further notes regarding this scatter result are discussed in section 4.2.3.3.
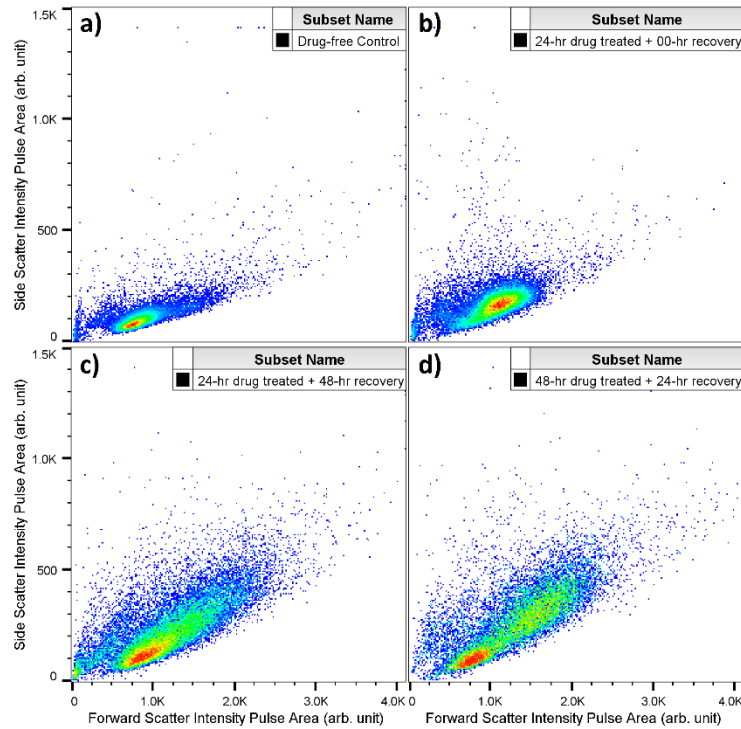
Figure 4-1: Forward and side scatter intensity dot plot of four samples: **a)** drug-free control, **b)** 24-hour drug treated with 0 recovery time, **c)** 24-hour drug treated with 48-hour recovery, and **d)** 48-hour drug treated with 24-hour recovery time. All samples are taken from a one-day experiment, and plotted after cleaning out a portion of the debris.

Next, cells are released from the one-day drug treatment and allowed to recover for two days (c in Figure 4-1, and orange in Figure 4-2). Even though the scatter intensity peak shifts back (from the second sample's) towards that of the diploid cells, the intensity medians remain close to that of the one-day drug-treated with no-recovery sample (blue). In addition, the intensity distributions for both forward and side scatter show larger robust coefficient of variance (RCV) values with a tail that extends over higher intensity channels.

In the last sample (d in Figure 4-1, and red in Figure 4-2), cells are kept in the drug for two days then left to recover for one day. For such a sample, a bimodal intensity distribution forms for both forward and side scatter, where the first intensity peak mostly overlaps with the scatter intensity peak for the control diploid sample, and the second peak is nearly twice as high.

These observations show that as the number of polyploid cells increases in a sample, the cells' scatter changes and starts to spread over higher intensities. In later sections

of this chapter, this gradual shift in scatter intensities is noted as cells move from diploidy to polyploidy through different phases of cell cycle.
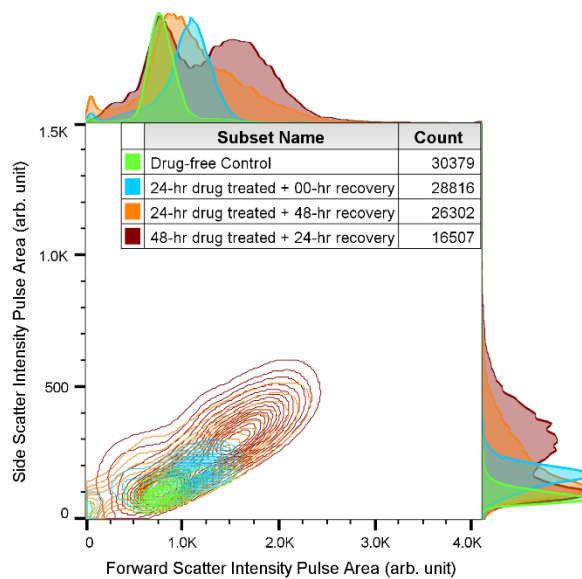
Figure 4-2: Forward and side scatter intensity dot plot of four samples: drug-free control, 24-hour drug treated with 0 recovery time, 24-hour drug treated with 48-hour recovery, and 48-hour drug treated with 24-hour recovery time. Figure also displays adjacent forward and side scatter histograms, where frequency is scaled by the mode value for each sample.

Lastly, the results also suggest that the area measurement of the detected light scatter pulse in either direction, forward and side, could serve as a good candidate in the detection of our target cells.

4.2.1.2   Effect of DRAQ5 labelling on cells scatter

Because the current work handles live cells, DRAQ5 is used for the DNA labelling to identify the live cell-cycling polyploid cells and study their corresponding scatter. Beside following the measures established in the previous chapter to ensure stability of the sample biology, the DRAQ5 labelling protocols have been observed in terms of conducting experiments as quickly as possible (see section 3.4 in Chapter 3). Nonetheless, it is worth noting how the scatter profile of DRAQ5-labelled samples seem to shift towards higher intensities compared to a DRAQ5-free (data not illustrated).

Statistical tests may be used to comment on the significance of observed intensity shifts. However, as previously established, with intensity data in flow cytometry, these tests often give false significance to drifts in data that are sampled from the same population. Therefore, significance in intensity channel shifts for both forward and side scatter could be set up based on the maximum intensity drifts in the light scatter of diploid cells in a control sample (along with those observed in control beads. See section 3.3.2.2 in Chapter 3). Peak separation resolutions can be calculated where the spread of the intensity peaks is also considered.

Based on such measures, the shift in the DRAQ5-labelled samples could be considered significant, which then suggests the presence of potentially physiological changes in the labelled cells that affect their scatter. Nonetheless, these shifts do not seem to affect the shape of the overall scatter distributions that we can already observe in DRAQ5-free samples (section 4.2.1.1). Therefore, it can be assumed that the scatter results investigated in the rest of the chapter using DRAQ5-labelled samples can also be observed in the DRAQ5-free ones.

4.2.1.3   Light scatter and single cells
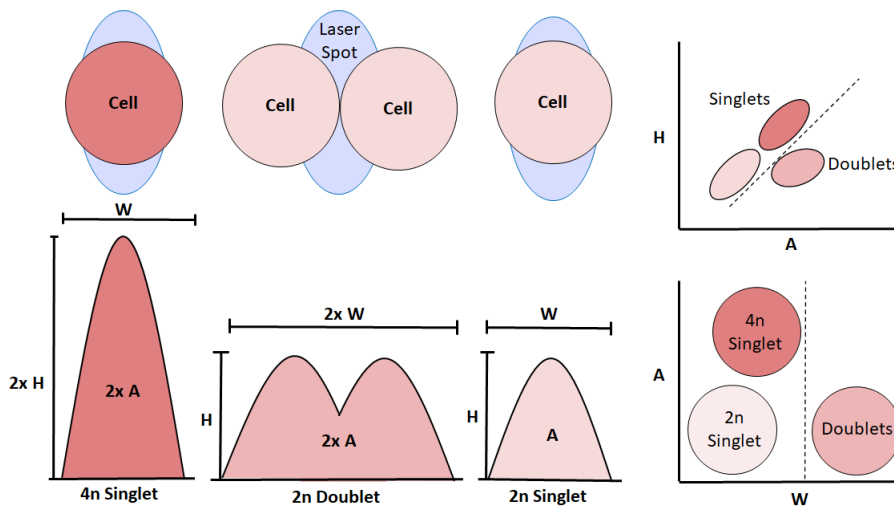
Figure 4-3: An illustration that shows three types of detected DNA-content fluorescence pulses (left) for cells travelling horizontally. The pulse parameters (width, height, and area) are then unlisted in doublet events exclusion via two types of scatter plots (right).

In the previous chapter, doublet events are recognised and excluded from the analysis through plotting the area-height or width-area of the DNA-content fluorescence (DRAQ5) pulse as illustrated in Figure 4-3.

For label-free samples, a similar gating strategy using the forward scatter intensity channel is often followed in flow cytometry. In the case of drug-free control samples, cells are expected to have tighter size spreads (excluding mitotic cells). Therefore, linearity in their area (A) versus height (H) scatter signal readings can be observed (See Figure 4-4, a), allowing for single cells to be gated.
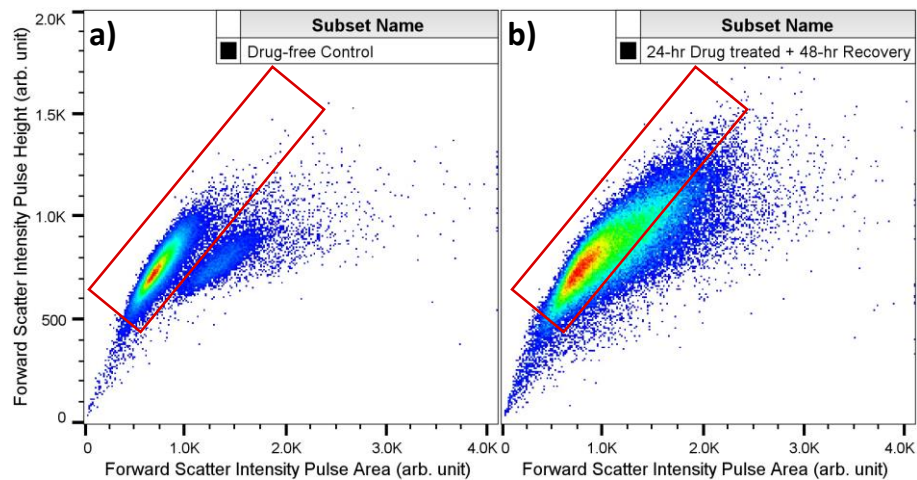
Figure 4-4: Dot plots of the pulse area versus height of the forward scatter intensity for a drug-free control (**a**), and a 24-hour drug treated sample with 48-hour recovery (**b**). The red rectangle is a rough single cells gate where the gate angle is set based on the control sample's singlet population. All samples are taken from a one-day experiment, and plotted after debris exclusion (see Figure 3-2 in Chapter 3).

However, in the case of drug-treated samples, the scatter A-H linearity settings held true for the diploid cells provide no clear distinction between the singlets and doublets (See Figure 4-4, b). Because scatter is the main interest of this work, the laser power and photomultiplier tube settings for the scatter channel are not changed between samples to attempt any corrections for the A-H linearity. Such observation can be of value on its own to understanding the scatter of polyploid cells. Further notes regarding the width and height parameters of the measured scatter of polyploid cells are addressed in section 4.2.3.4. A wide and rough scatter-based singlets gate could be applied to the drug-treated samples guided by the linearity angle of a diploid control, as illustrated in Figure 4-4. Nonetheless, for our scatter characterisation in the current

work, we have mainly relied on the DNA fluorescence for doublets exclusion to minimise such noise for the analysis.

4.2.2   Light scatter of diploid cells

While the diploid cells are not the main interest of this research, they serve as the scatter control or cell standard, to which any abnormalities in polyploid cells are noted. In this section, we first look at the overall scatter of diploid cells in drug-free samples (green population in Figure 4-5), against the diploid cells that are drug-stressed, such as those in the 24-hour drug-treated samples with 48-hour recovery time (blue population in Figure 4-5).

	Subset Name	Count
■	Diploids (Drug-free Control)	66425
■	Diploids (24-hr Drug treated + 48-hr Recovery)	29128
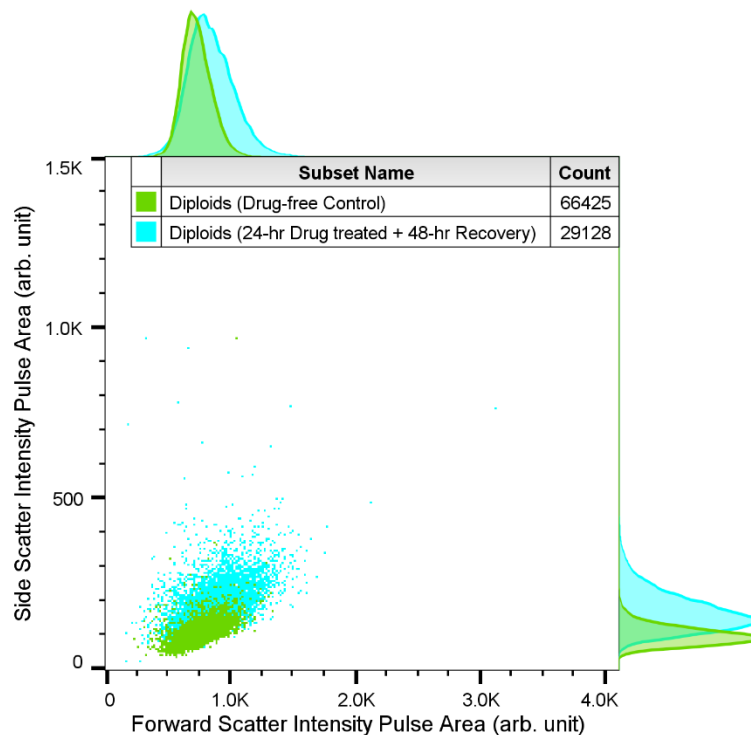
Figure 4-5: Forward and side scatter intensity dot plot of diploid cells for a drug-free versus a 24-hour drug-treated with 48-hour recovery samples (both are DRAQ5-labelled). Samples consist of combined data from three experiment repeats over different days. Figure also displays adjacent forward and side scatter histograms, where frequency is scaled by the mode value for each sample.

As illustrated in the adjacent histograms in Figure 4-5, significant shifts can be observed in the mean intensities for both the forward and side scatter of the drug-stressed diploid cells towards higher values. This can be expressed by the observation that approximately 6% and 45% of the drug-treated diploid cells had higher forward

and side scatter intensities, respectively. In addition, the RCV of the drug-stressed diploid distribution increased by 5.7±0.6% and 6.5±0.6% for forward and side scatter, respectively, compared to the drug-free control sample. From these results, it can be seen that the scatter of diploid cells, especially in the side direction, changes when they undergo the drug-induced stress. We can relate this observation to that of the previous chapter's, where the diploidy cell cycle of drug-treated cells show broader DNA content peaks than those of normal diploid cells. Combined together, it can be suggested that diploid cells with drug-induced genome irregularity may have different physiological properties than those of normal cells, which results in the observed increase in their light scatter intensities, especially that of large angle scatter.

It should be noted that the generated adjacent histograms on the dot plot figures used in this chapter are scaled based on the mode of the respective population for visualisation purposes. This is done to help the reader see all the distributions clearly and to follow any shifts in the mean intensity between the overlapping populations. However, this display choice may falsely exaggerate the percentage of the non-overlapped cells, such is the case with the forward scatter histogram in Figure 4-5, where the non-overlap region may look comparable to the side scatter's, when in reality the former is much smaller.

Lastly, we comment on the relative scatter changes of the cell cycle phases within diploid cells of drug-free and drug-treated samples. Figure 4-6 (green) shows gradual scatter mean shifts towards higher intensities for the drug-free control, as its cells cycle starting from the G0/G1 phase, then duplicate their DNA in the S phase, and divide in the G2/M phase. A high degree of scatter intensity overlap can be observed as well for these tightly packed distributions. For diploid cells in drug-treated samples (Figure 4-6 blue), the first notable observation is the increase in the RCV values of the scatter distributions, similar to the ones observed in the parent population (Figure 4-5). Due to that, the shift in the intensity means between G0/G1 phase and S phase is less significant. The figure also helps to illustrate the observed increase in the overall scatter intensity of the drug-treated samples.
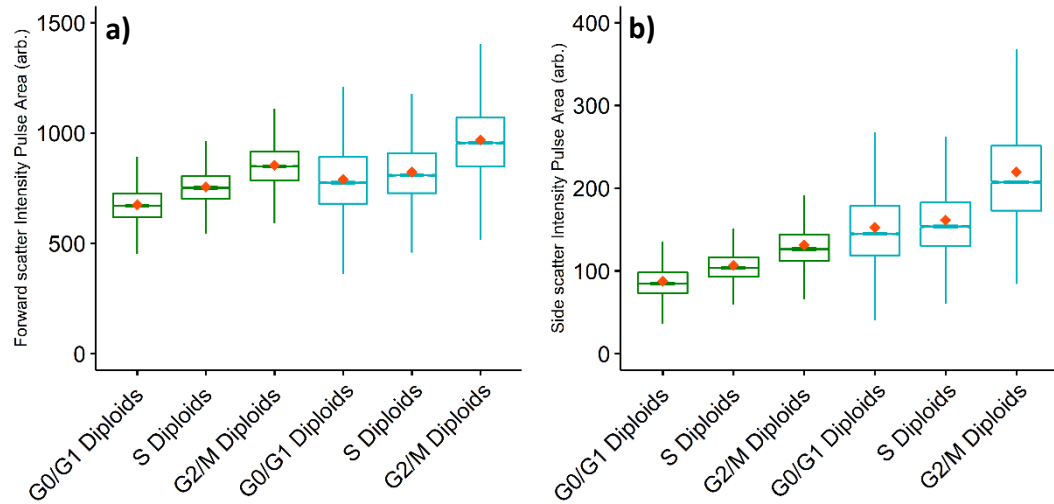
Figure 4-6: Box plots of the forward (**a**) and side (**b**) scatter intensity pulse area channels for diploid cells shown in their G0/G1, S, and G2/M cell cycle phases for a drug-free control (green), and a 24-hour drug-treated with 48-hour recovery samples (blue). Both samples are DRAQ5-labelled and consist of combined data from three experiment repeats over different days. The line inside the box refers to the median, the red diamond indicates the mean, and the box encloses 50% of the events. The upper (lower) whisker is at the smaller (largest) of the maximum (minimum) scatter value and Q3 + 1.5IQR (Q1– 1.5IQR).

4.2.3   Light scatter of polyploid cells

In this section, we look at the light scatter intensity properties of the cells that have transitioned to the polyploidy cell cycle. Firstly, the overall scatter of the polyploid cells is compared to that of the diploid cells from the same 24-hour drug-treated samples with 48-hour recovery. Next, for the same samples, the diploidy and polyploidy cell cycles are broken down to their respective cell phases to observe the trend of light intensity shifts across them all. The following section brings attention to how the previous observation compares to that of the light scatter of newly formed polyploid cells, i.e. polyploid cells in samples that are 24-hour drug-treated with no recovery. Lastly, the contributions of the event pulse widths and heights to that of the areas is discussed for both forward and side scatter. This helps understand the significance of the area measurement of the two label-free parameters in the detection of polyploidy cells.

4.2.3.1   Light scatter of overall diploid versus polyploid cells

The main interest of the chapter is to test whether the potentially high light scatter of polyploid cells can be used to differentiate them from diploid cells. Section 4.2.1.1 gives a first impression on the validity of the hypothesis through the high shifts in the scatter of the overall sample control versus drug-treated samples. In this section, we look closely at three collections (pooled) of scatter results for fluorescence-classified diploid and polyploid cells.
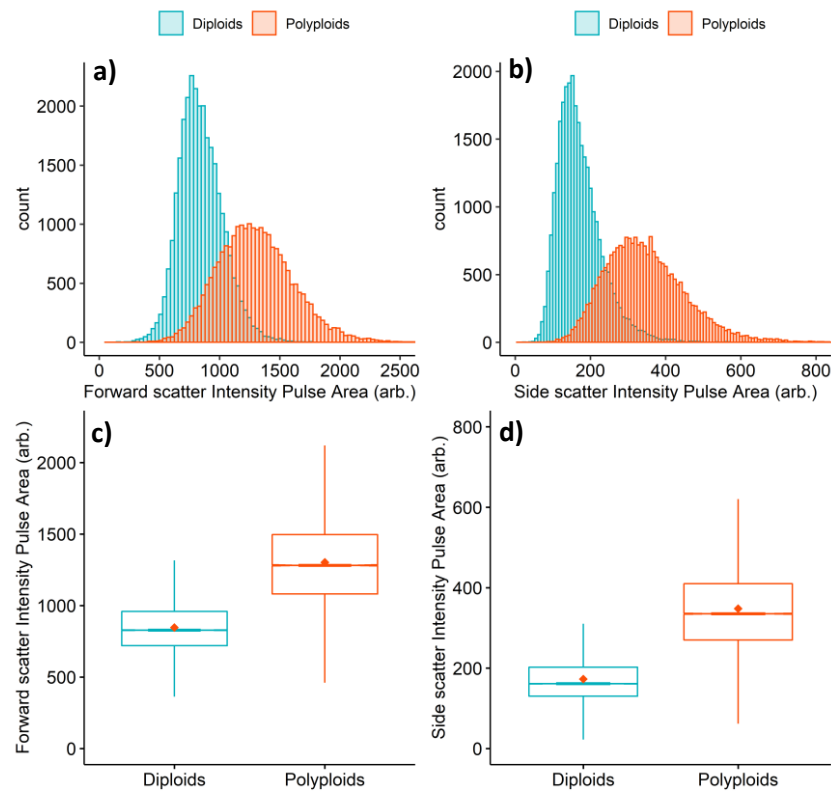
Figure 4-7: Histograms (top) and box plots (bottom) of the forward (**a, c**) and side (**b, d**) scatter intensity pulse area channels for both diploid and polyploid cells. The data are acquired from 24-hour drug-treated samples with 48-hour recovery collected over three days on the FACSVerse. The line inside the box refers to the median, the red diamond indicates the mean, and the box encloses 50% of the events. The histogram bins are 32 and 8 channels wide for forward and side scatter, respectively. The sample sizes for the 3-day pooled data are approximately 29K cells for the diploid cells and approximately 24K cells for the polyploid cells. The scatter scales are cropped for better data display.

In Figure 4-7 box plots, the polyploid cells show a significant shift in the median of the signal area parameter for both the forward and side scatter intensities. A considerable population separation can be observed in top histogram plots of Figure

4-7. For the forward scatter, the overlapped region covers approximately 17% of the data total, where approximately 78% of the polyploid cells had higher scatter intensities than 87% of the diploid cells. For the side scatter, the overlapping encompasses 14% of the data total, where approximately 85% of the polyploid cells had higher scatter intensities than 88% of the diploid cells. For these figures, it can be seen why the scatter distribution becomes bimodal in samples that have higher polyploidy concentration, e.g. 48-hour drug-treated samples with 24-hour recovery (section 4.2.1.1).

Adapted from chromatography, the separation of the two peaks can be expressed by the resolution R, which takes into account the intensity peak widths W when estimating the separation degree of the peak intensity channels I, where W for a normal distribution is four times the standard deviation $\boldsymbol{\sigma}$.

$$R = \frac{I_2 - I_1}{0.5(W_1 + W_2)} \, , \qquad where\ W_i = \boldsymbol{4}\ \boldsymbol{\sigma_i}, \qquad \boldsymbol{i} = \boldsymbol{1, 2}\ \ and\ I_2 > I_1 \qquad\qquad 4\text{-}1$$

In equation 4-1, an R value of 1.0 means a peak separation of one W wide for two normally distributed peaks of equal heights and widths. To put it in another perspective, an R value of 0.5 is nearly equivalent to the minimum resolvable objects according to Rayleigh criterion for resolution in imaging, where the first minimum of the one object's diffraction pattern is at the maxima of the other objects beside it.

For our application of equation 4-1, robust metrics are used for the scatter intensity peaks I of the diploid and polyploid cells (i.e. median peak instead of mean, and the robust standard deviation (RSD) instead of $\boldsymbol{\sigma}$). This resulted in $R = 0.47$ and $R = 0.55$ for forward and side scatter peaks, respectively. The baseline on what is the minimum accepted resolution depends highly on the impurity tolerance for the applications down the line. In later sections, we will illustrate the impurities expected when refer to these scatter results for the polyploidy detection in drug-treated samples.

4.2.3.2   Light scatter of cell-cycling polyploid cells in drug-treated samples

At different cell phases within the polyploidy cell cycle, a gradual increase in the median of light scatter intensities can be observed (Figure 4-8) in the drug-treated sample.
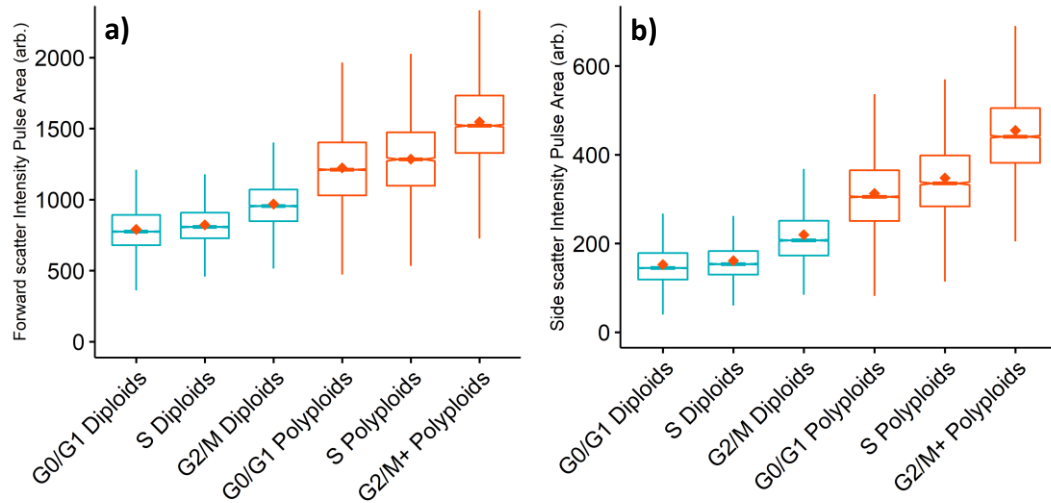
Figure 4-8: Box plots of the forward (**a**) and side (**b**) scatter intensity pulse area channels for the cells of different diploidy (blue) and polyploidy (red) cell cycle phases. The data are acquired from 24-hour drug-treated samples with 48-hour recovery collected over three days on the FACSVerse. The line inside the box refers to the median, the red diamond indicates the mean, and the box encloses 50% of the events. The box notches around the median, wherever visible, extend to ±1.58 IQR divided by the square root of the sample size, where IQR = Q3 – Q1. The upper (lower) whisker is at the smaller (largest) of the maximum (minimum) scatter value and Q3 + 1.5IQR (Q1– 1.5IQR).

This observation is similar to the one discussed in section 4.2.2, except that it shows even higher RCV values in the scatter intensity when cells move to the polyploid cell phases. Both previous and current observations indicate the presence of a positive correlation between the light scatter of the osteosarcoma cells and their DNA content ($r_s = 0.82$ and $r_s = 0.88$, for the overall forward and side scatter, respectively), which can be explored experimentally as a detection parameter for the ploidy cells.

Setting at the highest light scatter intensities (relative to the cell sample) are our target cell population, the cell-cycling polyploid cells at the G2/M phase along with the polyploid cells that transitioned to a second ploidy cycle. From the box plots, more than approximately 75% of the cell-cycling polyploid cells have higher forward and

side scatter intensities than the diploid cells of the same sample. Considering the relative rarity of these cells within one-day drug-treated samples, such high yield value gives us a positive impression about the potential of using scatter light for the detection and sorting of the target cells with high purity against diploid cell contaminants.

4.2.3.3   Diploid versus polyploid cells in 24-hour drug-unwashed sample

In the previous sections, the results showed the changes of light scatter as cells move through the different phases in the diploidy and ploidy cell cycle. They also show increase in scatter between the most scattering phase in the diploidy cell cycle (G2/M phase) and the least scattering one the polyploidy cycle (G0/G1 phase). This last observation, however, depends on the time at which the scatter is measured in a drug-treated sample.
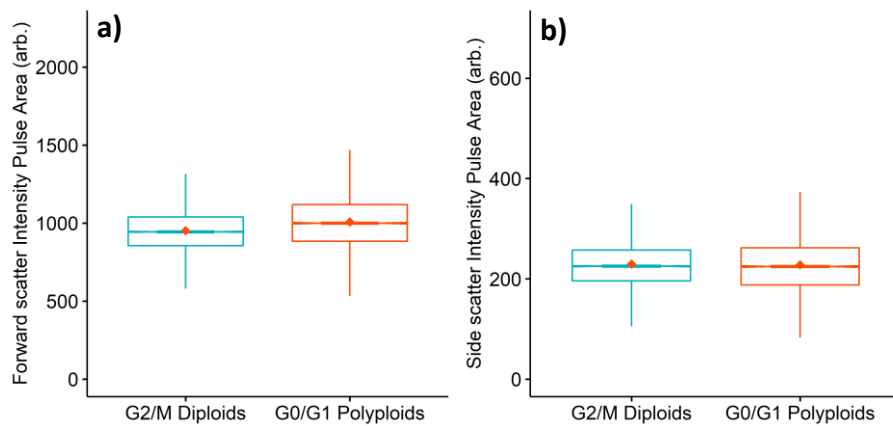
Figure 4-9: Box plots of the forward (**a**) and side (**b**) scatter intensity pulse area channels for the G2/M diploid and G0/G1 polyploid cells in a 24-hour drug-treated sample with zero recovery time, collected on the FACSVerse. The line inside the box refers to the median, the red diamond indicates the mean, and the box encloses 50% of the events.

To illustrate this, the cell scatter is measured after exactly one day of drug treatment, i.e. without giving the cells enough time to recover and cell-cycle. At this stage, the sample has a majority of cells that are either arrested at the diploidy G2 phase or have just transitioned to the G0/G1 polyploidy cells, but not yet committed to the ploidy cell cycle. In Figure 4-9, the forward and side scatter of these two populations is presented. Unlike the results from the previous section (Figure 4-8), the polyploid cells

in the G0/G1 phase have indistinguishable side scatter intensity profile from that of the diploid cells at G2, and a nonsignificant forward scatter shift.

This result suggests that during the drug-arrest, the cell shapes, sizes, and contents in both of the studied cell phases are most likely of similar nature due to the recent transition of the polyploid cells from the diploidy cell cycle. In other words, it suggests that polyploid cells develop high forward and side scatter sources only after they are released from the drug and allowed to grow and proliferate.

4.2.3.4   Light scatter pulse height for forward and side scatter in the polyploid cell detection

All the light scatter comparisons presented in the previous sections are based on the detected pulses' area measurement. From those results, it is inferred that both forward and side scatter serve as good candidates for label-free detection of the cell-cycling ploidy cells. Generally, the pulse area measurement is often recommended because it gives a better assessment of the overall brightness of the scattered intensity light for a particle in the flow stream. Depending on the laser beam spot shape, size and gaussian power distribution, a large particle that is twice as big and bright as a small particle could record a lower intensity peak height. In this example, the area measurement that incorporates information of the pulse duration, i.e. width, could help reveal the actual brightness difference between such two particles [4]. Nonetheless, for some cases, the area measurement may still not be sufficient to represent the detection between platforms with different illumination exposures (see discussion in section 4.3.1.1 for more). Overall, it is best to examine all the available measurements of a detected pulse to have a better understanding on the reported parameter. Figure 4-10 shows the same diploid versus polyploid cell populations from the FACSVerse data in section 4.2.3.1, but in terms of the height and width of the scatter intensity pulses.

For forward scatter, the strength of the area pulse measurement in polyploid cell detection seems to relay on the pulse's width. In Figure 4-10 (a, c), the forward scatter signal widths show a peak resolution of $R = 0.50$ between the diploid and polyploid distributions, which corresponds to an overlap of approximately 16%. In comparison, the forward scatter signal heights (b, d in Figure 4-10) have a lower diploid-polyploid

peak resolution ($R = 0.26$), which corresponds to a larger population overlap of approximately 29%. The area measurement of the forward scatter in section 4.2.3.1 informs us that the polyploid cells have considerably brighter scatter than the diploid cells. However, if only the height measurement is used, the forward scatter of polyploid cells may not be enough to be considered for their label-free detection against diploid cells.
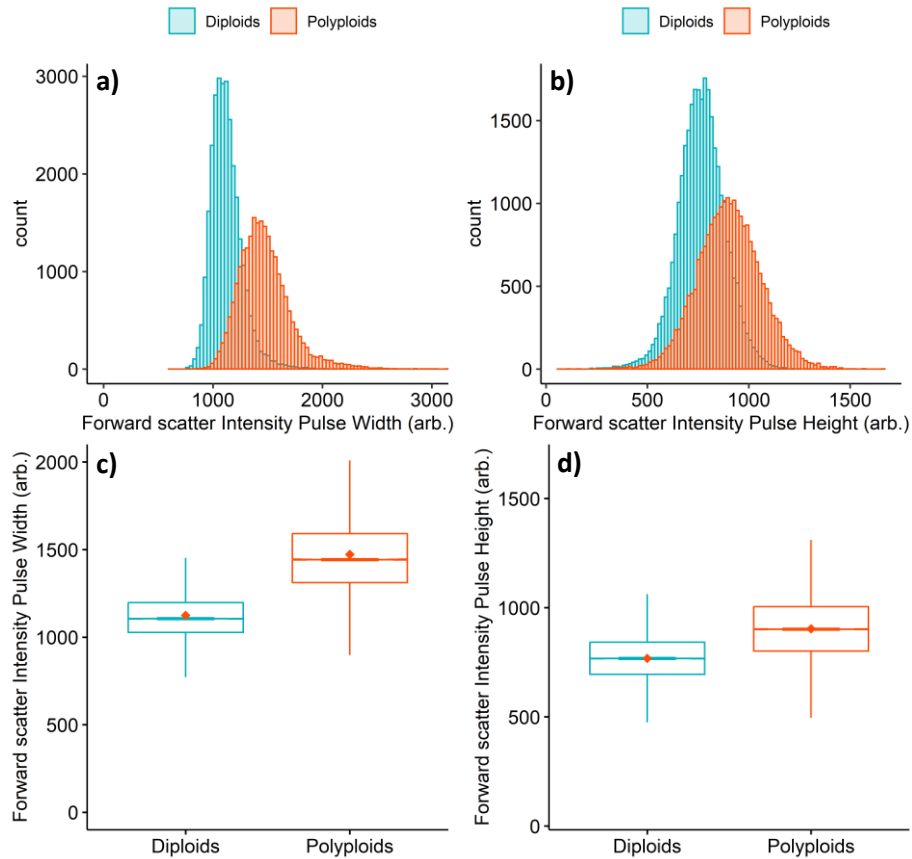
Figure 4-10: Histograms (top) and box plots (bottom) of the forward scatter intensity pulse widths (**a, c**) and heights (**b, d**) for both diploid and polyploid cells. The data are acquired from 24-hour drug-treated samples with 48-hour recovery collected over three days on the FACSVerse. The line inside the box refers to the median, the red diamond indicates the mean, and the box encloses 50% of the events.

On the other hand, Figure 4-11 (b, d) shows that the height measurements of the side scatter's intensity pulses provide a higher resolution between the peaks of the diploid and polyploid cells $R = 0.42$, i.e. an overlap of approximately 21%. The width measurement (a, c in Figure 4-11) behaves similarly to that of the forward scatter because the event detection on the instrument is triggered by the forward scatter's signals.
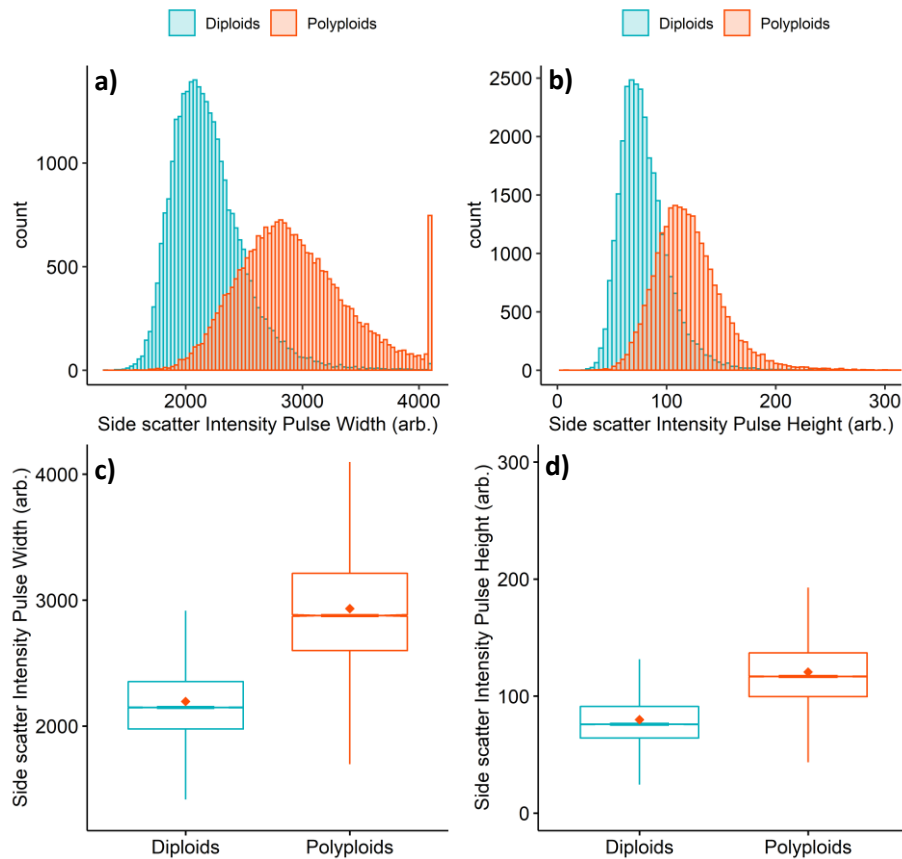
Figure 4-11: Histograms (top) and box plots (bottom) of the side scatter intensity pulse widths (**a, c**) and heights (**b, d**) for both diploid and polyploid cells. The data are acquired from 24-hour drug-treated samples with 48-hour recovery collected over three days on the FACSVerse. The line inside the box refers to the median, the red diamond indicates the mean, and the box encloses 50% of the events.

These observations may explain why the side scatter seems to be the better detection parameter for polyploid cells, than the forward scatter on zero-dimensional CFC instruments, i.e. instruments that only measure the height for the pulse of the scattered light intensity such as the FACSCalibur (data not presented). As previously discussed in the literature review section (1.4.2) of Chapter 1, the forward light scatter for cells correlates with larger cellular structures such as the cytoplasm or the nucleus, while the side scatter correlates with the smaller structures inside the cells. Depending on the laser beam size relative to the cell, the individual small side scatter sources within cells may have better chances at being fully illuminated with high power at the same time when the cell passes through the laser spot (see Figure 4-13), compared to a potentially partial illumination of the cell body for forward scatter. This then could explain how the height measurement of the side scatter intensity can give better diploid-polyploid

peak separation than that of the forward scatter, even without relying on the width measurement for a better overall intensity readout. However, the finding does not exclude the possibility that, when uniform illumination of the entire cell area is guaranteed, the peak signal of the forward scatter sources within the cells may still appear less significant than the side scatter's in detecting polyploid cells.

Lastly, it should be mentioned that the saturation in the high intensity channel of the width measurement of side scatter in Figure 4-11 is a limitation of the instrument setting for the experiment. As previously mentioned in Chapter 2 (section 2.5.1), the area-height scaling setting on FACSVerse is different for both forward and side scatter channels (where the latter's is associated with that of other fluorescence channels). While linearity has been observed as much as possible for this setting on both channels, channel saturation such as the one observed in the width channel of side scatter in Figure 4-11 is hard to avoid when the area scaling of the cells in the DNA fluorescence channel is favoured for doublets exclusion.

4.3   Light scatter for the detection of cell-cycling polyploid cells

This section utilises the previous results to find the cell-cycling polyploid cluster through their forward and side scatter. First, scatter controls are discussed for the detection of the target cells, particularly polystyrene beads for side scatter. The detection thresholds defined through such controls are then used in the automated cell data clustering of target cells. This is performed using the customised DAFi programme of this thesis, aiming to mimic a sorting experiment for either high yield or purity of target cells. The findings are expected to analytically verify whether the target cells can be recognised apart from the diploid cells within a heterogeneous sample.

4.3.1   Scatter controls for the detection of cell-cycling polyploid cells

As illustrated in Figure 4-2, the overall scatter profiles of the cell samples show broad distributions that extend over to high intensities when polyploid cells are present. Ideally, one would use a uniform reference sample that is an exact replicate of the target cells in terms of size, shape, refractive index distribution, and granularity. This

would increase the reproducibility of the scatter detection thresholds regardless of any instrument-specific differences, such as those related to the optical system (e.g. laser's spot size, wavelengths, or the numerical aperture of the collection lens), or to data processing (e.g. whether the detected events are processed as pulse height-only versus all three pulse measurements). However, such requirements for a reference sample are very difficult to realise, especially when dealing with cells of abnormal physiology such as the human osteosarcoma cell-cycling polyploid cells.

Therefore, one could simply refer to the scatter of drug-free cell control samples to define the thresholds above which the polyploid cells could be detected (Figure 4-2). This is similar to the concept of sorting with FL positive or negative peaks, albeit with scatter, the thresholds will be applied on a continuous distribution of single population, and not on discrete populations. It should be noted, however, that setting a sorting gate at the scatter boundary of such control cells does not count for the expected increase of diploid cell scatter in drug-treated samples (see Figure 4-5). Thus, contamination of drug-stressed diploid cells in the sorted drug-treated sample could be higher than potentially desired.

For finer gating on the target cells, another approach for defining scatter controls is to utilise beads with high refractive indices, such as polystyrene beads, as standard samples for the target cell detection, particularly for side scatter. Through analysis on the collected data, the used label-free polystyrene beads kit has shown potential as side scatter threshold references for target cell detection. Figure 4-12 shows an example of side scatter data collected on FACSVerse for the diploid (blue) and polyploid (light red) cells against that of unlabelled polystyrene beads (shades of purple and pink). From the figure, we can see that the side scatter area measurement of drug-treated osteosarcoma cells can roughly be covered on FACSVerse with the resulting scatter of the 4-15 µm polystyrene bead kit, which is the "Flow Cytometry Size Calibration Kit" from Molecular Probes, Inc. (F-13838, refractive index of 1.591 at 590 nm). In this example data, the 10 µm beads could be used to set a finer detection threshold to roughly find the median of the entire polyploid intensity distribution, which also approximately sets at the starting tail of the target cells' distribution (dark red).
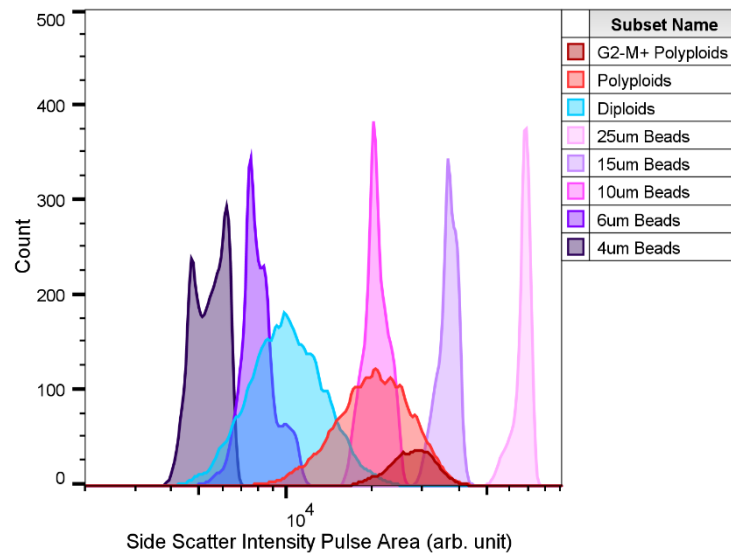
Figure 4-12: Histogram of the side scatter intensity pulse area for diploid and polyploid cells against polystyrene beads of five sizes: 4, 6, 10, 15, and 25 μm. The data are acquired from 24-hour drug-treated samples with 48-hour recovery collected over three days on the FACSVerse. Populations are manually gated.

Depending on the application, researchers may favour the target cell yield over purity in sorting and vice versa. Therefore, another example of a finer gate, in the example data, is to reference the valley between the side scatter peaks of 10 µm and 15 µm beads as a threshold for minimum diploid contamination with target cells. Therefore, large polystyrene beads can serve as reference candidates for the side scatter detection of the cell-cycling polyploid cells. However, it should be noted that their forward scatter may not be as suited for a scatter reference of the investigated osteosarcoma cells even with the largest bead. For example, polystyrene beads as large as 25 μm (Sigma-Aldrich Inc., UK) had forward scatter intensities that were less bright than the smallest viable cell in the heterogenous cell samples on either CFC instrument (data not presented), even though their side scatter intensity can be higher than that of the largest cells (see Figure 4-12).

4.3.1.1   Instrument-based limitations for scatter controls

One point to emphasise regarding scatter, whether of cells or beads, is that it is highly reliant on the instrument, such as its illumination size at the focus, or its detection system. The internal composition of beads cannot be equated to a cell's, even if they produce the same overall scatter sum as that of cells. As a result, instrumentation

differences could cause the studied cells to have a scatter at different intensities relative to the scatter of a beads kit. In other words, polystyrene beads as side scatter references for the target cells should be optimised for each instrument if needed, as the detected relative side scatter of each instrument may not be comparable.

To illustrate further, Figure 4-13 shows example results (histograms at the bottom) of the relative side scatter height measurements between the same polystyrene beads kit and osteosarcoma cells, as collected from either FACSCalibur (left) or FACSVerse (right). From the figure, we can see that the side scatter peaks of the cells were lower, relative to beads, for FACSVerse compared to FACSCalibur. As a result, bead-based side scatter thresholds for target cell detection on FACSCalibur will be different. This is believed to be a consequence of the different laser spot shapes between the two machines, as illustrated in Figure 4-13 (top). The scheme shows an example of how the largest cross-section of the same cell can be read differently by the FACSCalibur (a laser beam width of 22 μm along the flow direction) and the FACSCalibur (a laser beam width of 9 μm along the flow direction). In general, narrow beam widths along the direction of the flow are preferred in FC platforms to increase the resolution against doublet event detections. However, this comes at the cost of decreasing the simultaneous exposure area of objects that are larger than that beam width. This especially affects cells, whose side scatter sources can be the small organelles within it, thus resulting in a compromised pulse peak measurement due to the uneven illumination (see Figure 4-13, top).

For the FACSVerse, the side scatter reported by peak measurement of cells appears shifted for to the lower intensities relative to beads, compared to the readings of area (Figure 4-12). This shows the importance of using area measurement on FACSVerse, as it adjusts the peak measurement with that of width's, where FACSCalibur does none for scatter channels. Nonetheless, the height measurement on FACSCalibur is a result of a broader cell illumination, thus it may be closer to representing the actual relative difference between the cells and the beads (if seen under a full-coverage and uniform illumination). Even if the area measurement on the FC instruments were to sample the entire detected intensity pulse (i.e. using an integrator, instead of correcting the height with width), we believe that a beam shape difference could still affect the detected side scatter of cells relative to that of uniform spheres between instruments.
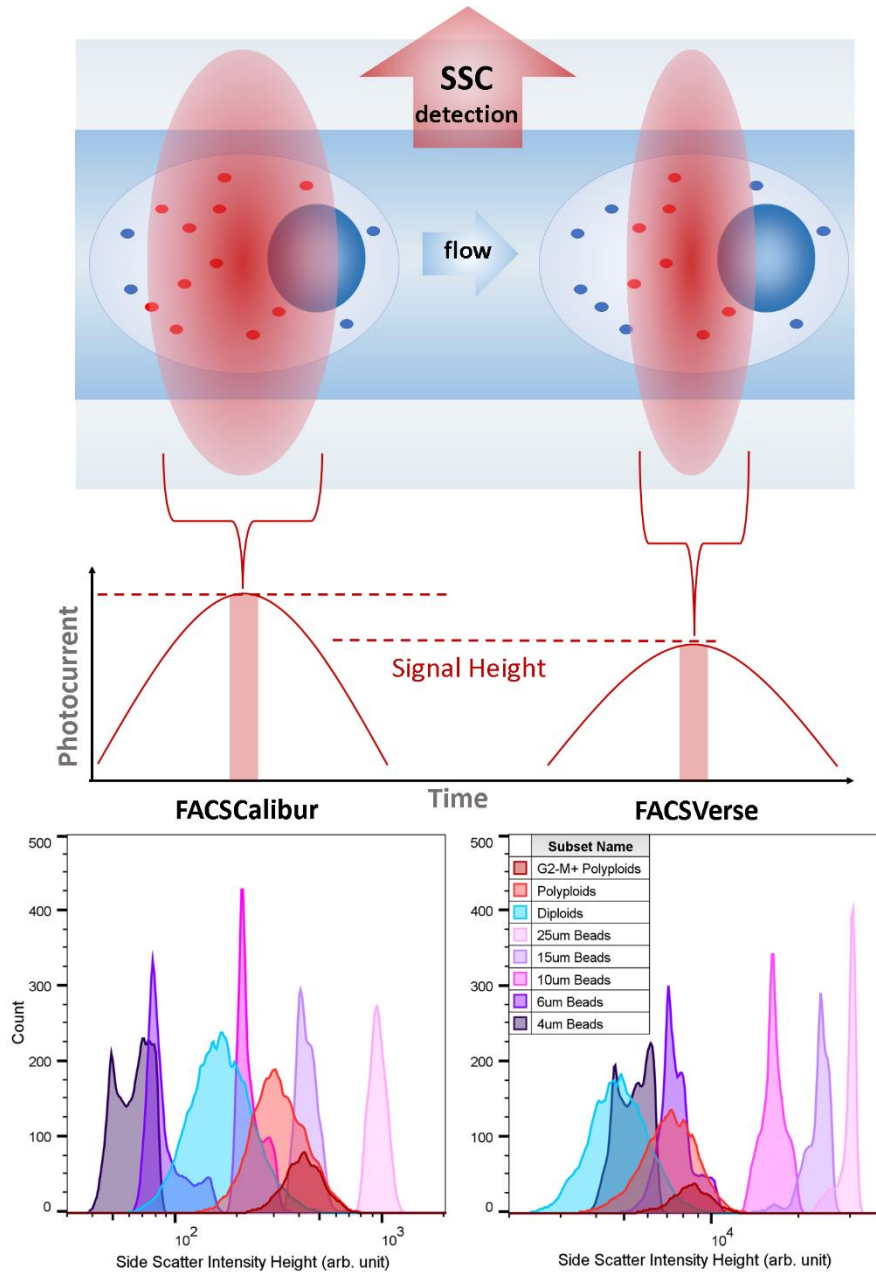
Figure 4-13: A scheme (top) illustrating the difference in laser beam widths across the flow direction between the FACSCalibur (left) and FACSVerse (right) as the potential cause for the observed shifts in a histograms (bottom) of the detected height measurements of the side scatter for the cells relative to polystyrene beads. Populations are manually gated. For the top scheme, the direction of the side scatter (SSC) detection is indicated, where the illustrated beam is incident into the plane of the page.

Lastly, we would like to note that, while the figure illustrates an example of a cell, the same can be said for beads of diameters higher than the beam width. In other words, this reasoning explains the less resolved distributions of the beads with diameters above 6 μm on FACSVerse, compared to their distributions on FACSCalibur. On the other hand, the distributions of the 4 and 6 μm beads can be seen overlapping in a

similar degree for either instrument. This is assumed to be a result of their varying travelling position within the fluidic stream core, relative to the collection lens's focal plane. In fact, the same reason is assumed here to be behind the observed high RCV values for the side scatter intensities of all beads compared to their forward scatter's. In conclusion, instrument variations, be it arising from the optics, fluidics or electronics, need to be counted for when scatter data reproducibility is desired on different instruments. This is especially important when referencing high refractive index beads as a side scatter standard, where preliminary experiments may be needed per platform to optimise the beads set for the target detection.

4.3.2   Automated target cell cluster detection using a customised DAFi gating for R

In this section, we mimic a sorting experiment setting through automated clustering methods, where the cluster application is guided by the scatter thresholds as an input for the customised DAFi gating programme for R of this thesis. Threshold optimisation depends on whether either yield or purity of the target cells is prioritised. Here, we present examples of both, where we apply such thresholds on the collected data of a DRAQ5-labelled 24-hour drug-treated with 48-hour recovery sample from FACSVerse. For both examples, only two scatter gates are applied, which are the whole cells gate (a in Figure 3-2 in Chapter 3), then a rectangular gate whose lower boundaries are defined by the proposed forward and side scatter thresholds. The steps are then repeated but with omitting either of the two scatter directions, to assess their potential effect on the overall filtering process.

For high yield filtering of target cells, we utilise control samples as a reference for suitable thresholds. Such thresholds are more suitable for actual sorting experiments, because control samples can verify and adapt to any potential instrument setting changes (during the sorting session). Based on the discussion in the previous section, two samples are referenced here to set up the scatter detection thresholds for the cell-cycling polyploid cells. For the forward scatter, the drug-free sample is referenced by setting a detection threshold at the median + 3RSD of its overall scatter (i.e. a gate

above approximately 99.85% of the drug-free diploid cells). For side scatter, the 10 μm beads are referenced by setting a detection threshold at its scatter median value. This resulted in a filtered population that is approximately 37% of the original sample with cells re-classified as presented in Figure 4-14 (a). The classification presented in the charts is based on the true label of the filtered cells, so debris (approximately 20% of the filtered sample) and GFP non-reporting cells were excluded from the results in Figure 4-14. In this example, the scatter-based filtering resulted in a nearly 100% filtering of all the assigned target cells in the original sample with a purity of approximately 45%, where near 54% of the filtered sample were other types of polyploid cells. All in all, only approximately 1% of the cells were diploid contaminants (which represents approximately 0.4% of the original diploid population). Thus, this showcases that forward and side scatter intensities have a great potential for label-free polyploidy detection and sorting. When omitting either of them (see b, c in Figure 4-14), we can see that the diploid contamination is highly affected by the presence of a side scatter threshold, as the contamination rises to approximately 5% when it is omitted. Furthermore, using the forward scatter only to filter target cells reduced their purity, as it allowed more of other polyploid cells (near 42% of their cell population in the original sample) to contaminate the target cell gate. In conclusion, filtering with only side scatter (area measurement) reference beads can be sufficient in a label-free gating of target cells. On the other hand, referencing only the forward scatter (area measurement) of drug-free control cells can be utilised for gating on overall polyploid cells (with a yield of approximately 85% of the overall polyploid cells in the original sample).

For high purity filtering of target cells, the scatter analysis of the target cells is referenced where thresholds were set on the 25th percentile on both forward and side scatter. Note that this does not necessarily mean the yield of the filtered outcome would be 75%, because this is done with DAFi gating, which does not apply strict cut-offs like the case with manual gates. Instead, the resulting yield was 85% (naturally for both or either parameter's results). The results when both parameters are used (Figure 4-14, d) showed a purity of approximately 68%, and a diploid contaminant of approximately 0.1% (representing approximately 0.02% of the original diploid population). They also showed that only near 14% of the non-target polyploid cells in the original sample were filtered as target (making approximately 32% of the filtered

cell, as per chart). We also see that, for the same percentile on either scatter parameter, the k-mean clustering of DAFi gated more of the non-target polyploid cells when only the forward scatter is used (Figure 4-14, e), thus reducing its target cell purity. On the other hand, using only side scatter to filter for the target cells nearly had no change from the filtered cells using both directions (Figure 4-14, f). As a matter of fact, thresholds could be pushed further for higher purity if desired. Such process would need to be optimised, considering that it could result in higher yield of doublets as contaminants, which is discussed next.

The scatter-based thresholds discussed in this section did not consider a third scatter gate to filter out doublet events. This is due to how the scatter of singlet events is less resolved in drug-treated samples (see Figure 4-4). In both examples discussed here, there were approximately 20% and 26% doublet events (estimated via DNA-DRAQ5 gating) in the filtered target population for high yield and purity, respectively. Because these cells are not imaged data, we cannot validate how much of these events were diploid cell aggregations.

Another point to highlight is the fact that sorting recovery in actual experiments can be affected by the size of the target cells in the original sample. Enriching the target cells in the original sample (e.g. through drug dose), could help increase the recovery rate of healthy cells post sort. The target cell recovery and the effects of doublet contamination could only be assessed through an actual sorting experiment; thus they cannot be considered in our analytical estimates.

Lastly, we would like to emphasise that the label-free detection of target cell is generally dependant on the design of the optical system in an instrument and the sensitivity of its detectors for either scatter direction. To illustrate further from our results, the standard photomultiplier tubes on the CFC instruments could observe the intensity peak shifts in the side scatter between the populations of interest. Their forward scatter detectors (avalanche photodiodes) did the same for the forward scatter, except with a weaker peak resolution. Because the sensitivity for either detector type is generally different, our results do not necessarily mean that the intensity measurements of forward scatter are absolutely less significant.
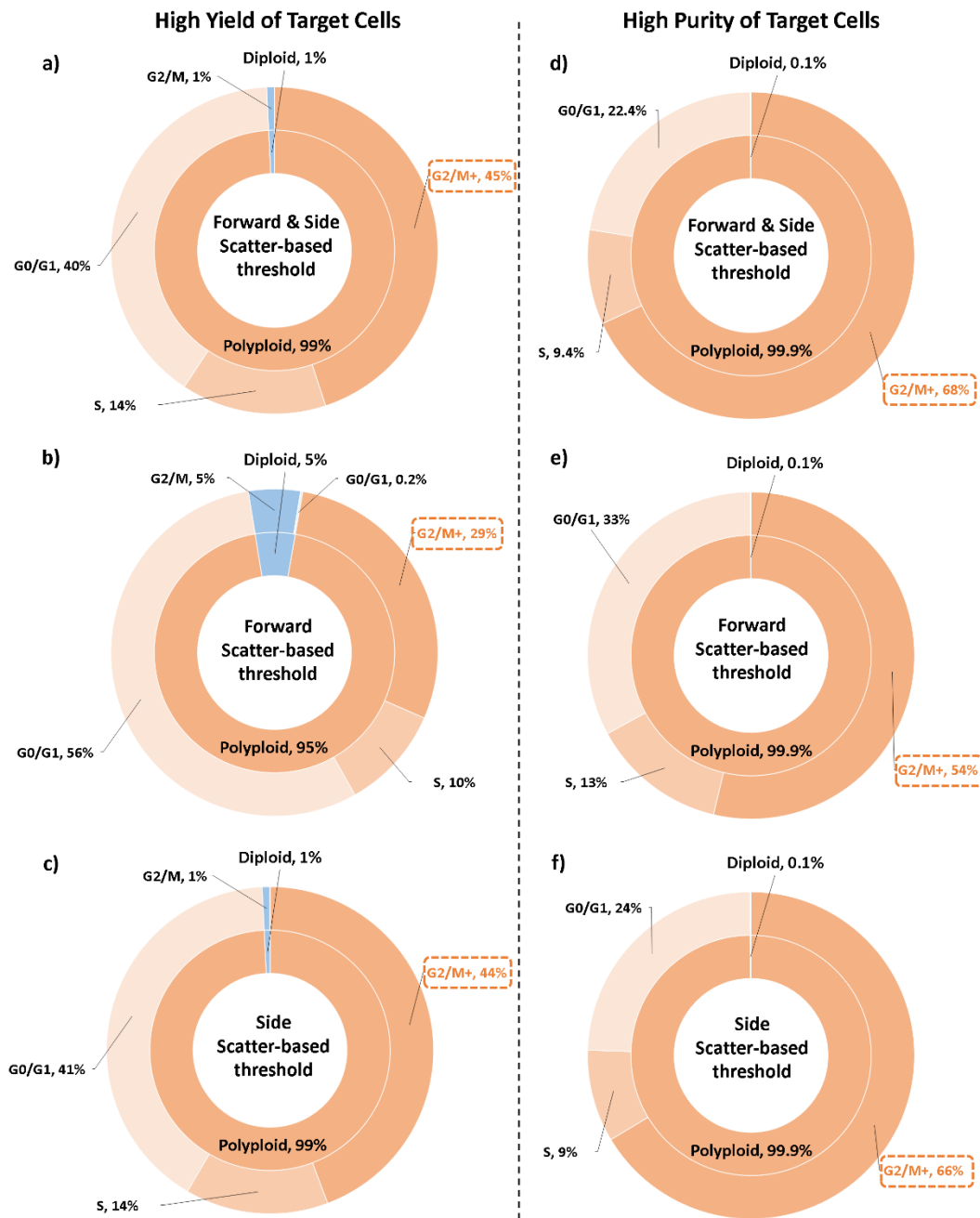
Figure 4-14: Results of applying scatter-based thresholds to target the cell-cycling polyploid cells with high yield (**a-c**) or purity (**d-f**), via customised DAFi for automated gating. The top charts (**a, d**) show when thresholds from both forward and side scatter are applied, then with only forward (**b, e**) or side (**c, f**) scatter. The scatter-based filtered cells are first gated afterwards for singlets and then their cell cycle readout is expressed as percentages of the total GFP-reporting population. The results are presented in a hierarchy of cell populations, starting from the nature of cell cycle (diploid or polyploid), and then their respective cell cycle phases, where target polyploids are highlighted with a red rectangle.

4.4   Summary

In this chapter, we have experimentally characterised the cell scatter of heterogeneous samples of diploid and polyploid cells. We have also analytically validated the hypothesis that the cell-cycling polyploid cells in osteosarcoma can be detected via their high forward and side scatter intensities in mixed samples of diploid cells. This has been performed by first characterising the forward and side scatter of the osteosarcoma cells at different drug-treatment stages via flow cytometry. A diploid-polyploid peak separation of $R = 0.47$ and $R = 0.55$ has been observed for the forward and side scatter intensity (pulse area) on FACSVerse, respectively. In addition, a positive corelation has been observed between either forward or side scatter and the cell DNA content through the different cell cycle phases. In this chapter, we have also identified some scatter controls for the detection of target cells, such as polystyrene beads. Based on these results, thresholds on the area measurements of the forward and side scatter for either high yield or purity of the target cells have been recommended and discussed. For the high yield objective, the scatter thresholds utilised both of the drug-free control sample and polystyrene beads as forward and side scatter references, respectively. The high purity thresholds were based on the scatter characterisation of the cells, which showed that nearly 75% of the target cells had higher scatter than the diploid cells. Next, automated clustering of the target cells have been performed on the data from a one-day drug-treated sample on FACSVerse using the customised DAFi programme of this work. This resulted in approximately 45% and 68% purity in the classified target cells via the thresholds set for high yield and purity, respectively. In both cases, the diploid cell contamination was equal to or less than 1% of the classified cells. When aiming for high yield, thresholds based on side scatter alone (collected via PMTs) can be sufficient in detecting the target cells with the least diploid contamination or high target purity, while thresholds based on forward scatter can be used for general polyploid identification. The chapter also highlighted how the width, height, and area parameters of the detected scatter pulses may impact the diploid-polyploid peak separations.

4.5 References

[1] T. Davoli and T. de Lange, 'The Causes and Consequences of Polyploidy in Normal Development and Cancer', *Annu. Rev. Cell Dev. Biol.*, vol. 27, no. 1, pp. 585–610, Nov. 2011, doi: 10.1146/annurev-cellbio-092910-154234.

[2] J. Coward and A. Harding, 'Size Does Matter: Why Polyploid Tumor Cells are Critical Drug Targets in the War on Cancer', *Front. Oncol.*, vol. 4, May 2014, doi: 10.3389/fonc.2014.00123.

[3] P. J. Smith *et al.*, 'Mitotic Bypass Via An Occult Cell Cycle Phase Following DNA Topoisomerase II Inhibition In p53 Functional Human Tumor Cells', *Cell Cycle*, vol. 6, no. 16, pp. 2071–2081, Aug. 2007, doi: 10.4161/cc.6.16.4585.

[4] H. M. Shapiro, *Practical flow cytometry*, 4th ed. New York: Wiley-Liss, 2003.

95

Chapter 5 :

Image-based label-free detection of cell-cycling polyploid cells in osteosarcoma

5.1 Introduction

The second hypothesis of the thesis states that the label-free brightfield (BF) and darkfield (DF) cell images can be used to classify cell-cycling polyploid cells in osteosarcoma apart from their diploid cells. This has been hypothesised for the same reasons mentioned in the previous chapter, i.e. samples containing polyploid cells tend to be observably different in shape or size compared to diploid majority samples. To assess this hypothesis, we utilise imaging flow cytometry (IFC) in the detection of our target cells. As described previously, IFC instruments, such as ImageStream, collects high-throughput cell image data, which provides increased dimensionality compared to data collected on conventional flow cytometers (CFC). A high cell image count with uniform illumination also serves as suitable data input for machine learning-based analysis. Therefore, supervised machine learning (ML) classifiers are used here to test our hypothesis. In particular, we reference a ML data analysis pipeline that has been recently introduced for ImageStream data [1], [2], as it employs open source methods for managing, extracting, and analysing cell image features.

In this chapter, we also discuss our labelling protocol for DRAQ5 that we have drafted to work around the limitations introduced by the sample preparation requirements of ImageStream. These limitations prevent us from following the standard DRAQ5 labelling protocol for CFC instruments [3], [4], thus affecting our ability to replicate the cell cycle readout via DRAQ5 as observed on CFC instruments.

The chapter first presents the methodology in section 5.2, which consists of four main steps. The DRAQ5 labelling protocol for cell cycle analysis on ImageStream is addressed in subsection 5.2.1.3. The results are then presented and discussed in section 5.3, where we first address the chapter's second hypothesis through cell size measurements, an example of a statistically relevant readout pulled directly from the brightfield images (section 5.3.1). Then we examine the results of the label-free ML classification (section 5.3.2) and present our conclusions.

5.2 Methodology

For this chapter, we can describe our method in four main steps (Figure 5-1): IFC data acquisition, image quality check and cell population gating to determine the ground

truth, pre-classifier image processing, and finally performing the cell classification via different ML models. These steps are described in detail below. In addition, we also discuss in section 5.2.1.3 our DRAQ5 labelling protocol for cell cycle analysis on ImageStream, which is needed to achieve well-resolved DNA fluorescence peaks for the samples.
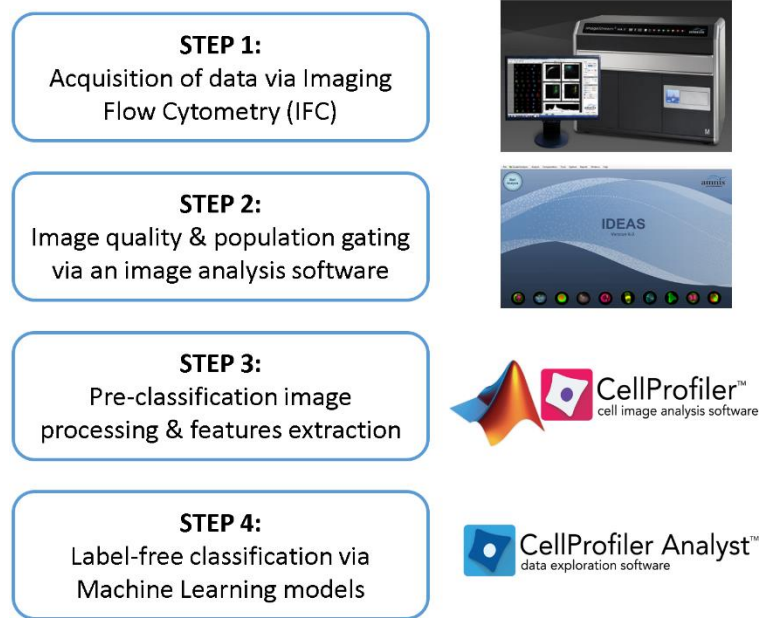
Figure 5-1: A summary of the main four steps, along with the used hardware or software, which have been followed to perform an image-based label-free detection of cell-cycling polyploids in Osteosarcoma.

5.2.1 Acquisition of data via imaging flow cytometry

The first step in testing the hypothesis is to acquire the label-free (brightfield and darkfield) images of our target cells. These images are intended to be used in the training or the testing of the supervised ML classifiers. The following subsections explain the instrument set-up and the cell sample preparations. In addition, we address here our optimisation experiments to the standard DRAQ5 labelling protocol for cell cycle, which are essential to biologically define our target cell while still accommodating the sample preparation requirements of the imaging flow cytometer.

5.2.1.1   Imaging flow cytometer

Fluorescence and label-free images of the cells have been collected using Amnis[®] ImageStream[®X] Mark II (Merck Millipore) imaging flow cytometer. The experiment settings are managed via Amnis[®] INSPIRE software, where instrument calibration is performed before each experiment, and each collected sample data is exported as a raw image file (.rif). For the illumination system, the 488 nm laser (for fluorescence excitation) was set at 100 mW, while the reported brightfield (multi-colours LEDs) power was within 30-33 mW. The darkfield 785 nm laser was set at the lowest power (1 mW) to help minimise pixel saturation in our sample (which carry a variety of cell sizes). The high-speed CCD camera magnification was set at 40x (with a numerical aperture of 0.75, a resolution of 0.5 µm/pixel, a view field of 60 x 128 µm, and 4 µm field depth). The fluidics were set at the lowest speed option for higher sensitivity (with a reported focused core size of 10 µm wide, and a speed of 66 mm/s). Brightfield images were collected on channel one (457/45 BP), CyclinB1-eGFP images on channel two (528/65 BP), DNA-DRAQ5 images on channel five (702/85 BP), and darkfield images on channel twelve (765/40 BP) (refer to Chapter 2 for the instrument's optical path image).

5.2.1.2   Cell samples and sample preparations

For this chapter's experiments, we worked with the same three cell sample types as previously described (see Chapter 3): drug-free control samples, 24-hour drug-treated with 48-hour recovery samples, and 48-hour drug-treated with 24-hour recovery samples.

Our cell samples are first prepared in the same steps and conditions as the ones previously described for the experiments conducted on the CFC instruments (see Chapter 3, section 3.2.1). ImageStream requires samples to be suspended in 20-200 µL at a recommended concertation of $2.0x10^7$/mL. To meet ImageStream's requirements for high sample concentration, these cells are cultured in high numbers using multiple 75-T flasks. The same cell experiment culture is also used for a parallel experiment on an CFC instrument to ensure the stability of the cell cycle readout (see

section 3.4 in Chapter 3). Next, samples are collected and suspended in 150 µL media each, where the final cell concentration depends on their type (see section 5.2.1.3).

With our cells, we have found that the acquisition rate drops often during a long session run despite how much volume is left in the samples. Retrieving said samples in the middle of sessions to manually mix or vortex it seems to also cause a loss in cell numbers. Therefore, each sample volume has been split into three further tubes (50 µL each) to increase the acquisition's efficiency, where the resulting three data files are then merged into one during analysis.

5.2.1.3   DNA labelling with DRAQ5 for cell cycle analysis of live cells on ImageStream

In literature, DRAQ5 labelling on ImageStream is usually used to locate the nucleus, so diluted DRAQ5 labelling is sufficient. However, in our case, identifying diploid cells (regular DNA count) and polyploid cells (high DNA count) to build up the ground truth of the ML classifiers requires a sufficient DRAQ5 probe binding to the DNA of each cell. When applied to our samples on ImageStream, the standard DRAQ5 labelling concentration for cell cycle (20 µM at ~4x10^5 cell/mL) has been found inadequate for several reasons. For one, an extra centrifuging step following the labelling would naturally be required to readjust the final cell concentration for the ImageStream session. However, the studied live cells have been found to be sensitive to such external stress, resulting in noisy cell cycle profiles with high <2N aneuploidy populations. On the other hand, applying the 20 µM DRAQ5 concentration to the final (high) cell concentration directly will lead to insufficient labelling of the sample that could barely resolve it from unlabelled samples. Therefore, we present here our optimisation steps to the standard DRAQ5 labelling protocol for cell cycle analysis. This will help produce well-resolved DNA intensity peaks on ImageStream for the live diploid and polyploid cells in our samples.

The first step is to determine the maximum cell concentrations that can be processed without affecting the health of the cells within the condensed sample volume. Such concentration is first determined for the sample type with the largest cell sizes (i.e. more susceptible to cell aggregations) in the experiment. The concentrations for the

rest of the sample types are then adjusted accordingly to keep the DNA content weight equal between all three sample types. In our experiments, the optimised cell concentrations were 4.2×10^6/mL, 2.8×10^6/mL, and 2.0×10^6/mL, for drug-free control, 24-hour drug treated, and 48-hour drug-treated samples, respectively.

The second step is to optimise the maximum recommended DRAQ5 labelling concentration for cell cycle by the standard protocol (20 μM), without compromising the quality of the sample or the instrument. Through some preliminary tests, we have found that an 80 μM DRAQ5 concentration with the proposed cell concentrations produces well-resolved DNA peaks on ImageStream, similar to those observed in CFC (see Figure 2-5 in Chapter 2). This concentration, however, appears to be at the lower limit of sufficient labelling. In addition, while there are no reported upper limits to applicable DRAQ5 concentrations, the standard protocols do draw attention to the probe's cytotoxicity with time. They also report its tendency to stain the inner tubes of the instruments, which may lead to cross contamination, thus the standard protocols do recommend observing reasonable DRAQ5 concentrations (along with proper cleaning procedures) [3]. Therefore, we recommend further future tests to determine the tolerance degree of DRAQ5 labelling for cell cycle analysis of live cells on ImageStream.

Lastly, we would like to highlight the fact that our ability to reproduce well-resolved cell cycle readouts on ImageStream does not necessarily mean that we have reproduced the same biology seen on CFC. This stems from the observed tendency of uneven cell sampling on ImageStream during the long experiment sessions. Therefore, a parallel experiment on a CFC instrument was run each time to observe the stability of the drug treatment response in the cell samples.

5.2.2   Image quality check and population gating via image analysis software

The second step in the methodology is to perform top level image analysis on the collected .rif files. This is done via IDEAS® (Amnis, USA, ver 6.2.187), the image analysis software for ImageStream, where numerical measurements of cell morphology and fluorescence features are extracted. In the following subsections, we

discuss first our gating hierarchy based on the relevant features for refining the quality of the collected images. Regarding identifying cell population of interests, we also explain the difference in our approach for cell cycle gating from the previous chapters, which helps prepare the data for the next steps.

5.2.2.1 Gating for single cells and image quality

Figure 5-2 shows our gating hierarchy for all cell samples, before cells are assigned to their respective cell cycle phases. By default, ImageStream refers to the brightfield channel for particle detection and lens focus calibration. Therefore, brightfield images are often referenced first to look for single cell populations (i.e. as usually done with forward and side scatter on CFC), and for focused images. Single cells are determined first via plotting the aspect ratio measurement of the detected objects against their area (Figure 5-2, a). Due to their spheric shape in suspension, singular osteosarcoma cells will display higher aspect ratio than their doublets and will have higher area than ImageStream's calibration beads.

Next, focused cell images are determined via the gradient root mean square (gradient RMS), the intensity, the contrast, and modulation measurements of the brightfield images (Figure 5-2, b and c). The gradient RMS gives an indication of cell image sharpness via looking at the average slope changes in the intensity profile (i.e. higher values of gradient RMS correlate with focused images). The same applies to the contrast and modulation (modulation = [max pixel - min pixel] / [max pixel + min pixel]) features, which help filter out streaked images (due to error in camera's speed tracking) as well as some apoptotic looking cells. Next, the circularity feature measurement is used to remove highly cropped images, as well as any apoptotic cells with blebbing membranes (Figure 5-2, d). Apoptotic cells also tend to show nuclear fragmentation. Therefore, we filter next for cells expressing the full fluorescent labels (Figure 5-2, e), then quantify any segmentation in the DNA images (Figure 5-2, f). Naturally, the gate is optimised to include viable mitotic cells.
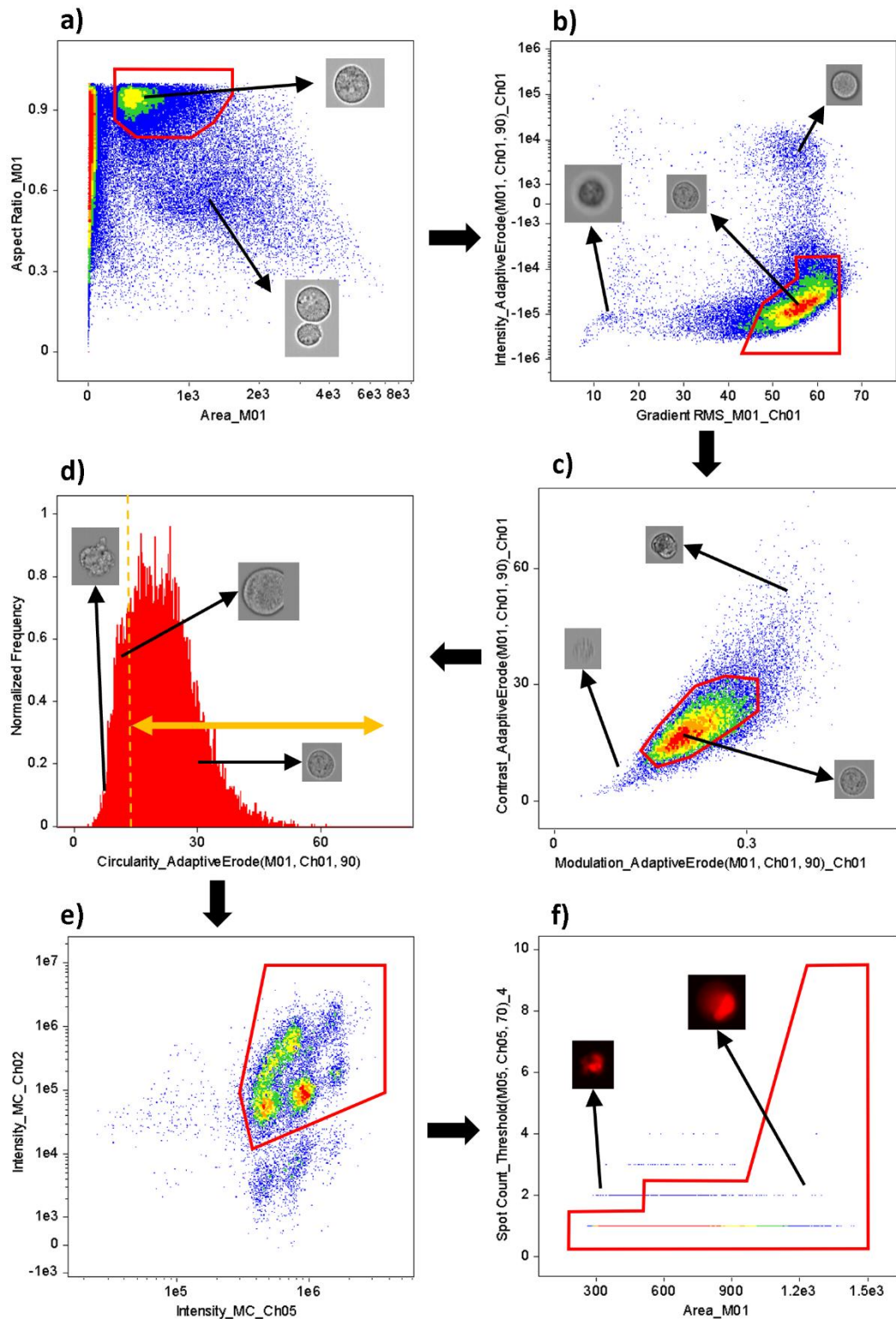
Figure 5-2: Gating hierarchy for image quality and cell viability check on ImageStream using brightfield (a-d) and fluorescence (e-f) images. These gates inspect the data for single (a), focused (b and c), viable (c, d, e, and f), non-cropped, (d), and fluorescently labelled cells (e), for the subsequent gating of the diploid and polyploid populations. The gating is performed via the image analysis software IDEAS, where the graph axes show the default title of each used image feature.

By this stage, we have referenced two user-defined object masks created via IDEAS's mask fitting functions masks (see Figure 5-3). The first mask is defined based on the brightfield images by setting the adaptive erosion coefficient in IDEAS's adaptive erode mask function to 90. Erosion is the process of deleting pixels from the edges of a mask, and the adaptive erosion coefficient in IDEAS refers to the percentage of the remaining pixels of the original mask (which is here the brightfield image default mask [M01]). This was found to conform to the cell shape better than IDEAS's default masks, which tend to exaggerate cell size. This can be seen in the mask displayed on the cell image in Ch02 in Figure 5-3 (figure shows the default combined mask [MC], but the same can be said about the brightfield default mask [not illustrated]). The second mask is created to locate the DRAQ5-labelled cell nuclei by setting the threshold mask function's intensity to 70%.
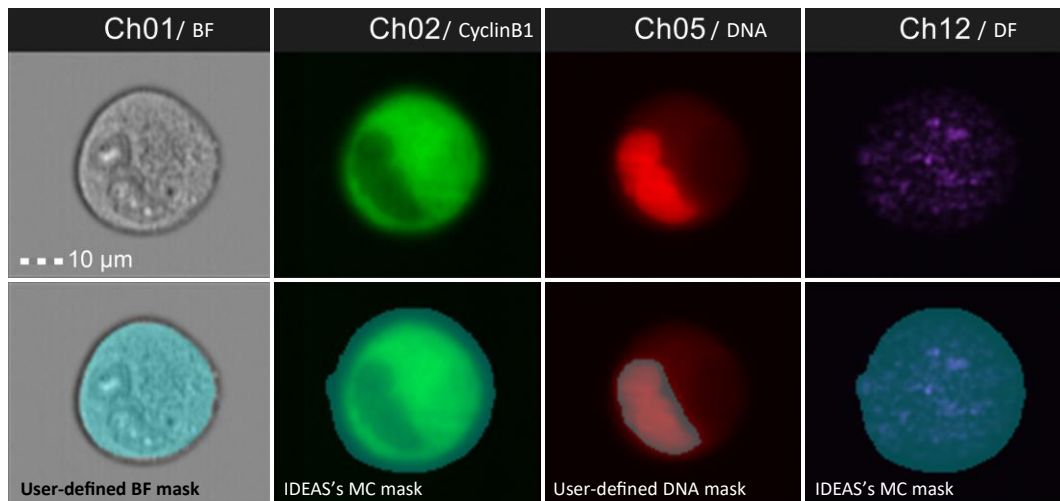
Figure 5-3: An example of a cell's brightfield (BF), CyclinB1-eGFP, DNA-DRAQ5, and darkfield (DF) images collected via ImageStream. The second row shows the same images but with the different mask fittings (highlighted in blue) that are used to calculate image features on IDEAS. These include IDEAS's default combined mask (MC), as well as our two user-defined object masks that we use for stricter cell boundaries (brightfield image) and locating the nucleus (DNA image). The picked example is a cell at the S phase of the polyploidy cell cycle in a 24-hour drug-treated with 48-hour recovery sample.

Lastly, histograms of the raw intensity pixel for all the channel images of are checked against any saturation before passing the filtered cells to the cell cycle analysis stage, discussed in the next section.

5.2.2.2   Gating for cell cycle and further cell analysis

The cell cycle gates are applied in the same direction followed in the previous chapters, except that the gating is performed manually, and it also takes note of the cells located at the boundary between the diploid and polyploid cell populations. Because of natural biology drifts, it is difficult to determine with high certainty to which population these cells belong. Therefore, these cells will not be used in the training of the ML classifiers.
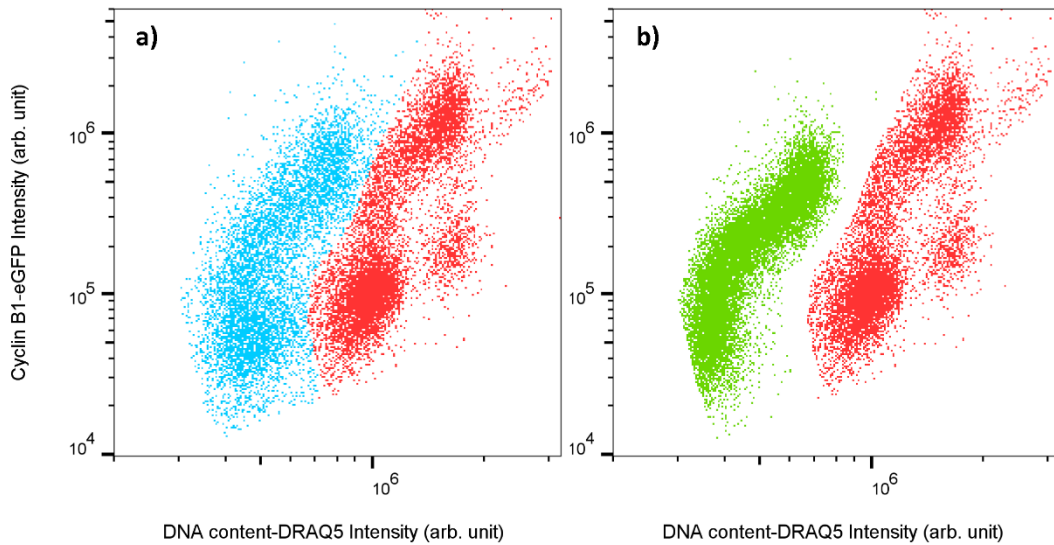
Figure 5-4: Cell cycle scatter plots showcasing the difference in DNA intensity spread between the drug-free (green) and drug-stressed (blue) diploid cells, relative to the polyploid (red) cell population. **a)** Shows the diploid (blue) and polyploid (red) populations in a drug treated sample. The same drug-treated polyploid population is plotted again in **b)**, where gated diploid cells from a drug-free (green) sample are referenced. Gating is performed via the image analysis software IDEAS, where the y-axis is the CyclinB1-eGFP intensity (a.u.) detected on channel 2, and the x-axis represents the DNA-DRAQ5 intensity (a.u.) detected on channel 5.

To illustrate, Figure 5-4 shows the diploid (blue) and polyploid (red) populations in drug-treated samples, in reference to the diploid cells in drug-free (Figure 5-4, green) samples. In the current work, we prioritise the certainty in defining the target cell-cycling polyploidy population against the diploid cells. Therefore, for cell cycle gating, we first place a boundary between the two populations via referencing the 48-hour drug-treated samples (polyploid cells majority). Next, we reference the drug-free control samples (diploid cells majority) to draw the diploid cells gate that will define the diploid cell candidates for the classification's training set. While excluded from

the training set, any diploid cells that belong to the overlapping region (mostly from drug-treated samples) will be considered later in the final classification step.

After all cell classes are biologically identified, TIFF images are generated and exported via IDEAS to train and test the ML classifiers in the next step.

5.2.3   Pre-classification image processing and features extraction

In this step, two actions are performed: the first is to tile the exported raw images from the previous step into cell montages, and the second is to import the montage images into the open-source image analysis software CellProfiler (Broad Institute, Inc., ver 4.0.7) for cell segmentation and extraction of morphological features.

CellProfiler is mostly suited to process images with multiple cells commonly seen in microscopy, as opposed to single cell images from IFC. Therefore, the first step aims to speed up CellProfiler's processing time for the thousands of cell images in our data (during the second step). For our application, 15x15 cell montages have been produced via MATLAB [1] with a cell unit measuring at 120x120 pixels, to accommodate the various cell sizes in our sample without cropping them (an example of a tiled image is provided in Supplementary Figure B-1 in Appendix B).

In the second step, we use CellProfiler to segment the cells, and extract morphological features, which are categorised in the software as: area and shape, intensity, granularity, radial distribution, and texture (texture step size is 3 pixels). This step also covers any needed background normalisation within the feature calculation procedure of CellProfiler, where the feature lists and description are provided by the software documentation [5]. The original method's CellProfiler pipeline [1] has been mostly followed here, except that we only use one primary object mask (based on brightfield images) for feature extractions of both brightfield and darkfield images. Other than reducing processing time, this helps minimise bugs arising in the next analysis tool (CellProfiler Analyst), which does not handle well classification of features extracted via multiple primary object masks.

It should be noted that all the features extracted from the images in this step are solely used for the subsequent classification step, and not to infer exact measurements on the

cells (especially area-based ones). This is due to the tendency of the used CellProfiler object mask to exaggerate the cell boundary fitting (see cell mask highlighted in red boundary in c, Figure 5-5). However, because the same mask is referenced in the analysis of both brightfield and darkfield images, the overfitting helps to ensure proper coverage of all relevant parts of the cell in either image. We assume this would have no meaningful effect on the following classification, as this aspect of the mask fitting does not appear to be biased towards a particular phenotype.
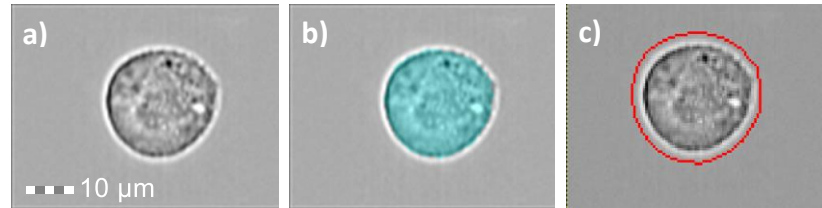
Figure 5-5: An example of a brightfield image of a G2/M diploid osteosarcoma cell from a 24-hour drug-treated with 48-hour recovery sample, showing it with **a**) no mask , with **b**) our user-defined cell mask on IDEAS (highlighted in blue), and **c**) the used CellProfiler pipeline's mask (outlined in red).

The reported cell sizes in the results of this chapter (section 5.3.1) are based on the area feature measurement extracted via IDEAS, where our user-defined cell mask tends to be more accurate in terms of locating cell boundary upon visual inspection (mask highlighted in blue in b, Figure 5-5).

5.2.4   Label-free classification via machine learning

The performance of the ML classifiers relies on choosing a non-biased and well-defined ground truth for each cell class, as well as suitable ML algorithms for the classification, and the tunning of its hyperparameters. For this step, we import our data into the open-source image analysis software CellProfiler Analyst (Broad Institute, Inc., ver 3.0.3), where we build the training set and perform the cell classification. The following subsections describe our criteria for the training set, along with other details regarding the used ML algorithms and the models' evaluation.

5.2.4.1   Training set criteria and the chosen label-free features

To test the hypothesis of this chapter, we have defined a 3-class training set as follows: 1. Diploid cells, 2. G0/G1 and S Polyploid cells, and 3. G2/M+ Polyploid cells (i.e.

our target cells). Once again, the plus sign in the G2/M polyploid phase refers to polyploid cells that transitioned to a second ploidy cell cycle (i.e. cells whose G0/G1 phase has >8N DNA content) (see Figure 5-6 for the three cell classes positions in the cell cycle).

The training set was built using 600 randomly selected cell images (200 for each class), which were equally pulled from two different experiment results. To avoid oversampling the original samples, the training set for each class was built from the sample type in which the said cell class is abundant. For example, the 24-hour drug-treated samples were used to select the training set for the diploid cell class. The other two polyploid classes were trained from the 48-hour drug-treated samples. In our approach, we did not refer to the drug-free samples to train the diploid cells because these samples have not been exposed to the drug stress. It is also worth mentioning that, biologically speaking, the second class may contain cells that have just transitioned from a diploidy state to a ploidy one. Such cells may appear indistinguishable physically, be it externally or internally. Therefore, we expect this cell class to be the least resolved among the three classes.
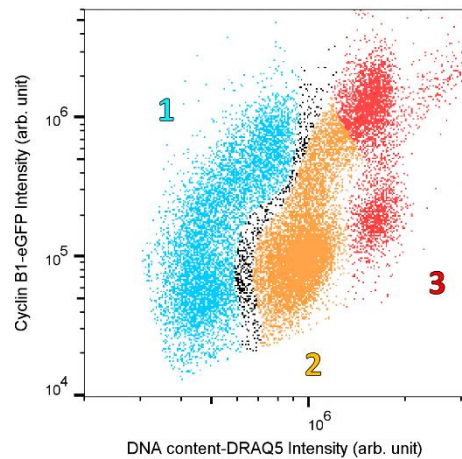
Figure 5-6: Scatter plot of the cell cycle readout for a 24-hour drug-treated with 48-hour sample showing the three cell classes used for the label-free ML classification. The cell classes are 1. Diploid cells (blue), 2. G0/G1 and S polyploid cells (orange), and 3. G2/M+ polyploid cells (red). Highlighted in black are the diploid cell populations that will not be used in the sampling of the training set, due to their close proximity the polyploid cell population.

In defining our ground truth for the ML model training, we have followed three criteria. The first is to verify that the ML classifiers show low bias towards cells

belonging to different experiment days. This is checked via running preliminary ML classification tests on control cells from either experiment, where we look for chance-like performance. The second is to have the selection process avoid cells from the uncertain subpopulations previously defined in section 5.2.2.2 (populations highlighted in black in Figure 5-6). The third is to ensure that the training set has equal proportions of each population fed into it wherever possible (summarised in Figure 5-7). For example, the diploid cell class was built with nearly equal proportions of G0/G1, S, and G2/M diploid cell populations. On the other hand, the G2/M+ polyploid cell class consists of 50% of the G2/M polyploid cells and 50% of the >4N polyploid cells (40%, 20% and 40% of which are G0/G1, S and G2/M cell phases, respectively), because the latter type is under-represented in the original samples. In addition, the size of the training set (600) was selected based on the abundance of the least enriched cell phase (e.g. polyploid cells in the S phase in the third class, see Figure 5-7).

Lastly, the training set is checked for any apoptotic-looking cells that may have escaped the initial image quality screening (e.g. cells with small degrees of plasma membrane blebbing that were not detected via the cell quality check steps on IDEAS, see Supplementary Figure B-2 in Appendix B).
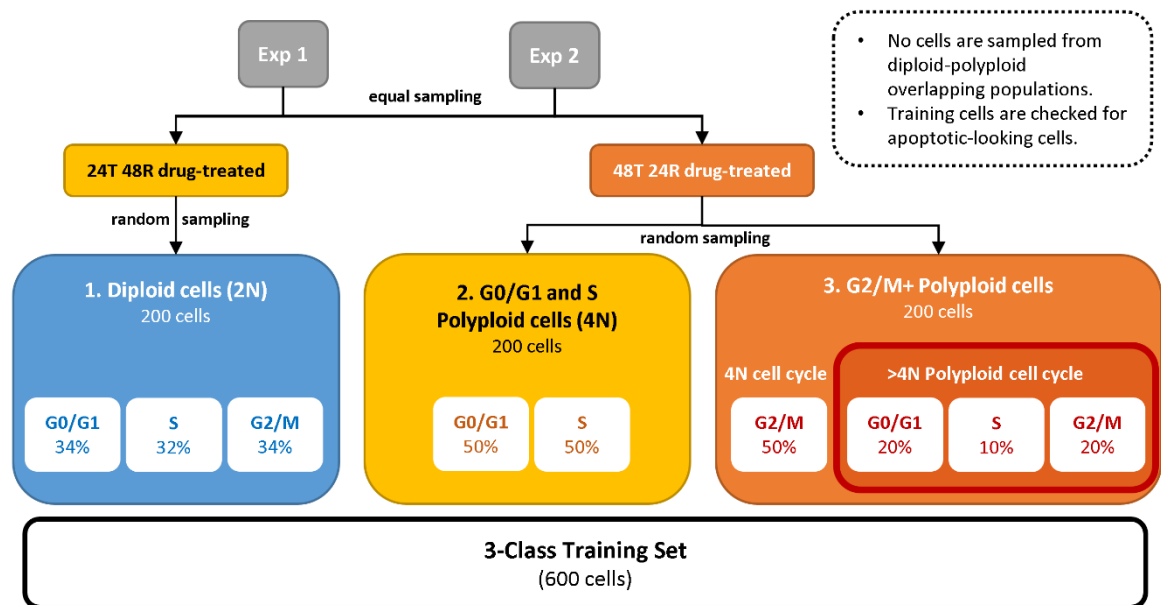
Figure 5-7: Diagram summarising how the 3-class training set is sampled for our label-free machine learning classification of the cell-cycling polyploids (i.e. G2/M+ polyploid cells) in Osteosarcoma.

5.2.4.2 Label-free classification and results evaluation

In our ML classifiers, we have tested the same ML algorithms used by H. Hennig et al. [2] in their label-free classification of mitotic cells using CellProfiler Analyst, which are the Random Forest (RF) and the Gradient Boosting (GB) classifiers. Both tree-based algorithms are appropriate for this type of analysis, and have performed well in previous works [1], [2]. We also test the performance of four more ML classifiers supported by CellProfiler, which are: Support Vector Machine (SVM, herein abbreviated as SV), Logistic Regression (LR), K-nearest Neighbours (KNN, herein abbreviated as KN), and Multi-layer Perceptron (MLP). The classifiers were set to ignore cell features that are irrelevant to the analysis such as those related to cell location in the montage, and cell orientation.

The classification results are evaluated by looking at the recall values for the overall (or per class) classification of each model. The recall value assesses the classification sensitivity, and it is similar in principle to the yield measurement used to assess a sorting performance. The recall is calculated from the confusion matrices of each classification, which refer to tables summarising the prediction performance of each class against its true label. In the classification of a given class, true positives (TP) and true negatives (TN) refer to the accurately predicted positives and negatives (where positives are data that belong to the said class, while negatives are those that do not). Similarly, false positives (FP) and false negatives (FN) are the inaccurately predicted data. The recall metric for a class looks at the prediction power in assigning the correct label to each cell (recall = TP/[TP+FN], i.e. true positives/actual total of the label). Other metrics that are commonly reported in ML classifications such as the precision (purity) or accuracy are not reported here because they are dependent on the size of the true classes in the data set.

Model evaluation also includes commenting on the potential significance of input features in the classification. This is done through the leave-one-out method, where a feature module is omitted from the model training and data scoring each time. In ensemble decision tree models (such as RF and GB), there are several approaches to estimate the importance of input features after the model is trained and the trees are built. Inherently, the tree branching in these models is based on a measure of impurity at each node, so it is common to find a readout on overall feature importance reported

along a trained tree-based model. While it is the easiest to retrieve, this readout tends to overestimate features with high cardinality, thus it will not be reported either.

5.3   Results and Discussion

Label-free brightfield and darkfield images of the live U-2 OS osteosarcoma cell samples have been collected from two-experiment repeats on ImageStream, and the results are discussed in the following subsections.

Th first section looks at the reported cell sizes from the brightfield images, as area-related features show promising potential for the label-free identification of our target cells. In addition, cell size or shape can naturally correlate with other feature measurements such as those of intensity, which has been shown to have the capability of identifying the target polyploid cells in mixed samples of diploid cells (see Chapter 4).

We discuss in the second section the results of the 3-class classification of the brightfield and darkfield images for all collected cell samples. We then report on the performance of the six tested ML models and their ability to find our target G2/M+ polyploid cells against diploid cells in samples.

5.3.1   Osteosarcoma cell size measurements via brightfield images

We report here the observed sizes of the studied cells at different stages of the cell cycle. Similar to the previous chapter, comparisons will be made on the whole diploid and polyploid populations, as well as their subpopulation (G0/G1, S, and G2/M cell cycle phases). All the measurement in this section have been performed on IDEAS using our user-defined mask for the brightfield images (see section 5.2.2.1). The data from two experiment repeats are combined then size measurements are plotted and compared in either box plots or histograms. Cells that fall in uncertain gates (described in section 5.2.2.2) are included based on the general cell cycle population to which they were assigned. Cell sizes will be expressed here in either area or diameter measurements. On IDEAS, area is calculated first by adding up the pixels belonging to a cell's brightfield image mask. This is then converted into squared micrometres based on the used magnification (where it is 1 pixel $= 0.25\ \mu m^2$ for a magnification of

40x). The diameter measurement is calculated in the software from the area measurement by assuming a circular shape for the masked area.

5.3.1.1   Cell size measurement of diploid cells

Using a box plot (Figure 5-8), we compare here the interphase cell cycle populations in diploid cells from drug-free samples (green), and the diploid cells from 24-hour drug-treated with 48-hour recovery samples (blue). The plot helps immediately spot the increase in the diameter medians (horizontal line inside) in the drug-stressed cells when compared to the cells with no drug treatment. For the drug-free diploid cells (nearly 24K cells in total), the diameter medians for G0/G1, S, and G2/M phases were 16±1, 17±1, and 18±1 µm, respectively. For the drug-stressed diploid cells (nearly 11K cells in total), the diameter medians for both G0/G1 and S phases was 19±2 µm, and 21±3 µm for G2/M phase. The G0/G1 and S phases in drug-treated samples (be it for diploid or even polyploid cells) having highly overlapped distributions has also been observed with the scatter intensity measurements in the previous chapter.
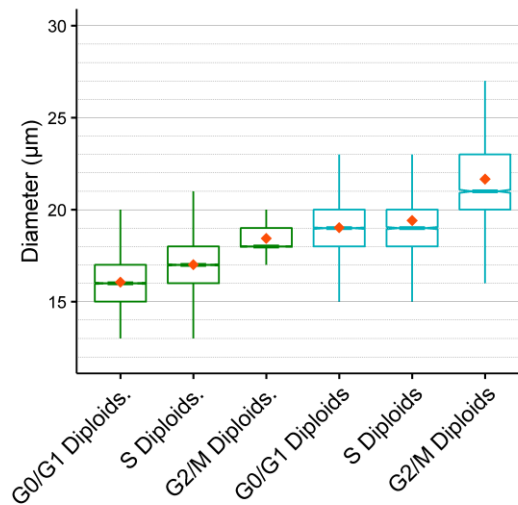
Figure 5-8: Box plot of the cell diameter measurements extracted from the brightfield images of the osteosarcoma diploid cells at different cell cycle phases (G0/G1, S, and G2/M) for a drug-free control (green), and a 24-hour drug-treated with 48-hour recovery samples (blue). Both samples are DRAQ5-labelled and consist of combined data from two experiment repeats on ImageStream. The x-axis is categorial, showing repeated set of the cell cycle phases for each sample in a different colour, and the y-axis refers to the measured diameters expressed in distributions along the y-axis. The line inside the box refers to the median, the red diamond indicates the mean, and the box's height encloses 50% of the events.

As noted from the cell diameter measurements, the diploid cells that have been exposed to the drug environment (then allowed to recover) do exhibit larger cell sizes compared to drug-free diploid cells. This has been seen in the previous chapter through both the forward and side scatter intensity readings. Therefore, it suggests that cell size shifts in the drug-stressed diploid cell is one of the factors affecting the previously reported increased intensity values.

The importance of this observation comes from how it could affect any subsequent analysis on the cells, be it for sorting or classification purposes. For example, this result suggests that if we were to train a ML model to classify diploid and polyploid cells using a training set sampled from a drug-free sample, then chances are the model may mislabel many diploid cells in drug-treated samples and classify them instead as polyploid cells. In fact, this may also extend to polyploid cells in drug-free samples. In other words, such polyploid cells may get mistaken for drug-stressed diploid cells as they also exhibit different cell sizes than drug-induced polyploids. However, in the case of the latter, the rarity of polyploid cells the drug-free samples (especially cell-cycling polyploid cells) makes it difficult to draw any meaningful analysis on them. Nonetheless, it remains true that a gradual increase in cell size is observed in drug-free samples as the DNA content increases with the change of cell position in cell cycle (green boxes in Figure 5-8). Polyploid cells in drug-free samples can then be assumed to reflect similar gradual increase in cell size (as also observed with drug-induced polyploid cells later in section 5.3.1.2.2). Therefore, the rest of the analysis in the following sections will be based on diploid versus polyploid cells in drug-treated samples.

5.3.1.2   Cell size measurement of polyploid cells

In this section, cell size variation in diploid and polyploid cells from drug-treated samples are reported. We first look at both populations as a whole, then we observe the changes through cell progression in the cell cycle. The latter should also highlight our target cells and how the compare, size-wise, to the rest of the sample.

5.3.1.2.1 Size measurement of overall diploid versus polyploid cells

Figure 5-9 summarises cell size measurement in terms of diameter (a) and area (b) for the overall populations of diploid (blue, nearly 11K cells in total) and polyploid (red, nearly 27K cells in total) cells in drug-treated samples. Both measurements show shifts of cell size in polyploid cells towards higher values.
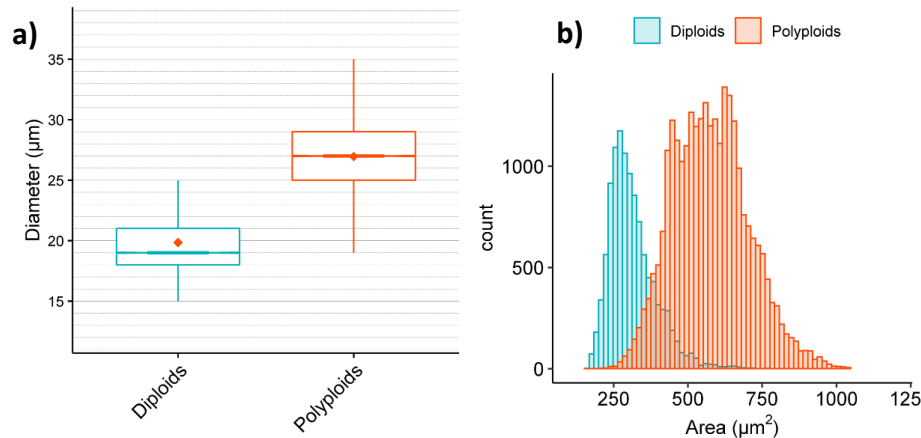
Figure 5-9: Cell diameters box plot (**a**) and area histogram (**b**) of the osteosarcoma diploid (blue) and polyploid (red) cells from drug-treated samples (DRAQ5-labelled). These were extracted from the combined brightfield image data collected over two days on ImageStream. For the box plot, the x-axis is categorial, showing each phenotype in its assorted colour, and the y-axis refers to the measured diameters expressed in distributions along the y-axis. The line inside the box refers to the median, the red diamond indicates the mean, and the box's height encloses 50% of the events.

At the intersection of the area measurement histograms, approximately 88% of the polyploid cells could be seen to report higher area measurements than approximately 85% of the diploid cells. Using the peak resolution formula from the previous chapter would give us $R = 0.68$, which is not far off the values we found previously with the scatter intensity measurements. It is worth mentioning, that this resolution metric seems to be sensitive to the size of the polyploid population and its drug-treated sample type. For example, if only the polyploid population from the 24-hour drug-treated sample is compared to its diploid population (nearly 8K cells), the resolution will be $R = 0.52$. This is not truly representative of the overall polyploid population, because the 24-hour drug-treated samples tend to mostly have cells in the G0/G1 phase of the

first ploidy cell cycle. Therefore, the polyploid cells from the 48-hour drug-treated samples are needed for a proportionate representation.

If a circular shape is assumed for the suspended cells, this will give us diameter medians of approximately 19±2 and 27±3 µm, for diploid and polyploid cells, respectively.

These results suggest that the supervised ML models will most likely utilise the area-related or area-dependant features to classify the diploid population with high precision (because it belongs to its own class in the classification).

5.3.1.2.2   Size measurement of cell-cycling polyploid cells in drug-treated samples

Continuing from the previous section, we examine here the cell size variation in subpopulations within drug-treated samples. This includes examining how the target cells, i.e. cell-cycling polyploid cells (at G2/M+ phase), fair in size against diploid cells of the same sample, thus addressing the feature's potential in the classification of cell target images.

Similar to the previous two sections, the cell size for different subpopulation of the diploidy (blue) and polyploidy (red) cell cycles from drug-treated samples is expressed via diameter measurement box plots in Figure 5-10. The diameter medians for G0/G1, S, and G2/M phases of diploid are the same as those reported in section 4.2.2. For the polyploid cell population, the diameter medians for G0/G1, S, and G2/M+ phases were 25±3, 27±3 and 29±2 µm, respectively.

These measurements support the chapter's hypothesis as they show (via cell size here) how label-free (brightfield) images provide significant information in identifying our target cell-cycling polyploid cells against diploid cells. They also suggest that a size-based threshold could be implemented (in a comparable manner to the previous chapter's) to identify the target cells against the diploid ones. However, this is based only on direct area or diameter measurements on the brightfield images, and not on the bigger ensemble of features that can be extracted from either brightfield or darkfield data. If we are to allow the decision making to branch further with the rest of the

features in a ML model, then we could expect the predictions' power to increase for our target cells.
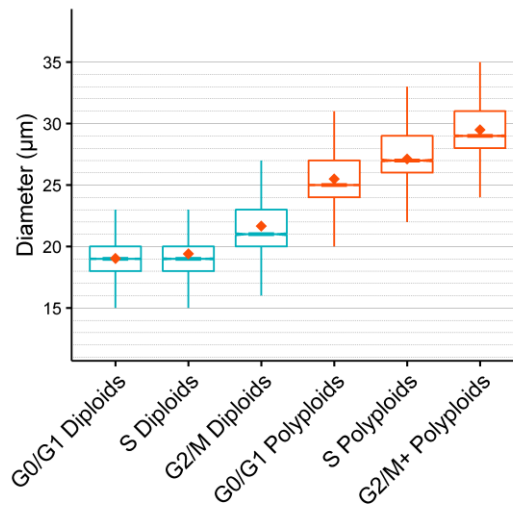
Figure 5-10: A box plot of the cell diameter measurements extracted from the brightfield images of the osteosarcoma diploid (blue) and polyploid (red) cells at different cell cycle phases (G0/G1, S, and G2/M) for a 24-hour drug-treated with 48-hour recovery sample. These were extracted from combined brightfield image data of DRAQ5-labelled samples collected over two days on ImageStream. For the box plot, the x-axis is categorial, showing each phenotype in a different colour, and the y-axis refers to the measured diameters expressed in distributions along the y-axis. The line inside the box refers to the median, the red diamond indicates the mean, and the box's height encloses 50% of the events.

In Figure 5-10, an increase in the RCV values can be observed in cell phases starting from drug-stressed diploid cells at the G2/M phase moving upward, suggesting increased size variability (and potentially classification noise) among the drug-treated polyploid cells. The target cell population, G2/M+ polyploid cells, has the largest cell size in the sample, with more that approximately 75% of its cell having larger diameters than diploid cells. Consequently, we can expect a supervised ML model to detect less diploid contamination in our target cells based on such area-related features. On the other hand, the polyploid cells at the G0/G1 and S phases are likely to be the least resolved cell classes, with more chances to be mistaken as the target polyploid cells than the smaller diploid cells. These observations do help in expecting the minimum prediction power of a supervised ML model trained and tested on our label-free cell image data, which are shown next. These findings are also similar to those of scatter from the previous chapter, except in ML classification we cannot control the output yield (or recall here) by the learners.

Lastly, if thresholds based on size (or diameter here) are to be used, we can see that between the overall diploid cell population with median at approximately 19±2 µm and the target cell's at 29±2 µm, we would need a size detection resolution of at least 2 µm to gate 75% of the original target cell with minimum contamination from the original diploid population (less than 0.2%, by setting a threshold at the diploid's median+3RSD, i.e. approximately 25 µm here).

5.3.2   Image-based label-free machine learning classification of cell-cycling polyploid cells

The hypothesis of this chapter is tested here via running supervised ML models on nearly 280 numerical features (see section 5.2.3) extracted from the brightfield and darkfield cell images. As previously stated in the methods (section 5.2.4), the data are first trained using six different ML algorithms to look for three phenotypes: diploid cells, G0/G1 and S polyploid cells, and our target cells, the G2/M+ polyploid cells. The ML models have been trained with cells (a total of 600 cells), then used to classify cells from both the training set and new unseen data of all three sample types in our work (a total of approximately 100K cells). In the following subsections, we first discuss the results of when features from both images are used together in the classification. Next, we discuss when either image type is used on its own, along with the potential effect of each feature module on the classification.

5.3.2.1   Classification results using all features of the brightfield and darkfield cell images

Figure 5-11 summarises the classification results for the six tested models based on the label-free features extracted from the brightfield and darkfield images.

When classifying the cells using brightfield and darkfield images together, all six models scored above 75% in the overall classification recall (Figure 5-11, a), with KN and LR being the weakest performers with nearly 6% difference from the top performer, the GB model. However, these overall scores are less informative regarding our topic of interest, which is the segregation power between the diploid cells versus the cell-cycling polyploid cells. When it comes to the classification of both populations,

all ML models had correctly labelled an average of approximately 70% of the target cell-cycling polyploid cells (Figure 5-11, b), and 90% for the diploid cells (Figure 5-11, c). Even though the prediction recall for target cells was variable between the weakest and strongest ML performer (from approximately 68% to 75%), the mislabelling of diploid cells as target cells was only approximately 0.7% or 0.3%, respectively, of the original diploid population for all models (Figure 5-11, b). The same is true for the predicted diploid cell class, where mislabelling target polyploid as diploid cells was as low for all six models (Figure 5-11, c).
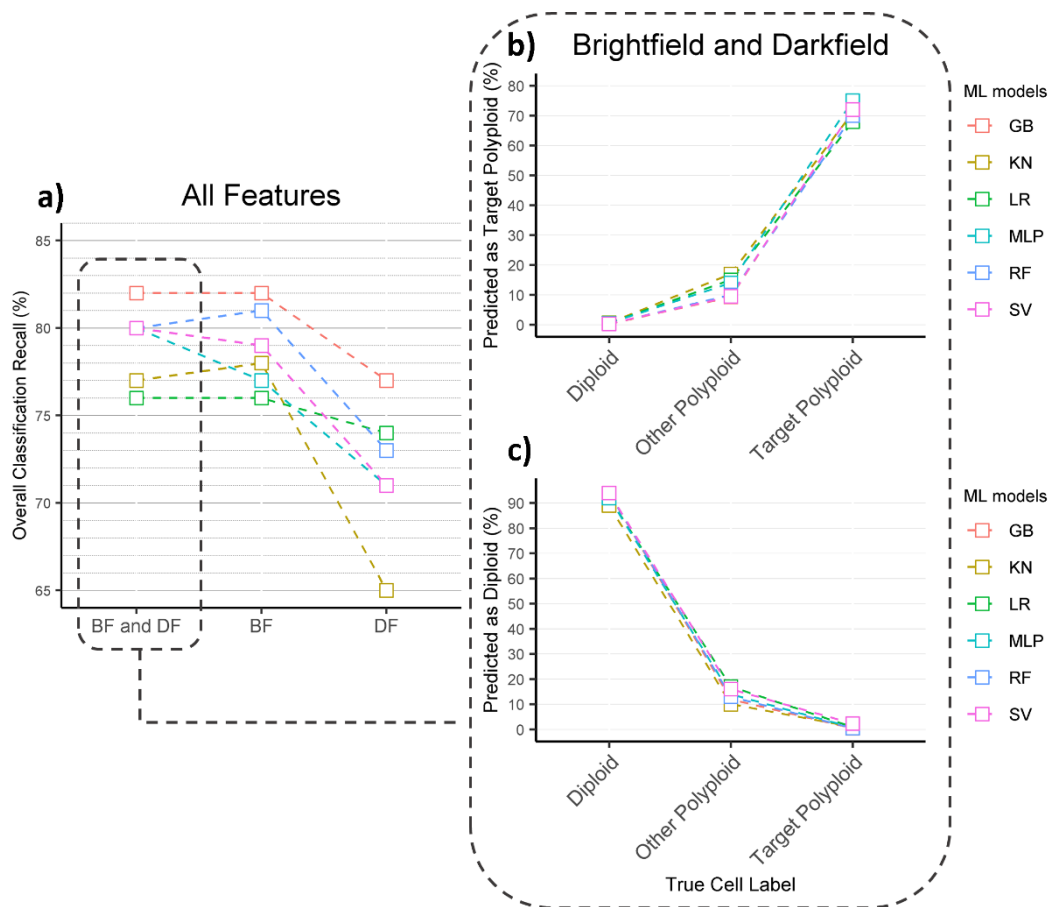
Figure 5-11: Plots summarising the classification performance using the brightfield (BF) and darkfield (DF) images via the reported recall scores. Plot **(a)** compares the average recall scores of the overall cell classifications using all extracted features from the BF and DF images together or individually. For classification with all features from both BF and DF images, the figure also shows the predicted numbers for our populations of interest: **(b)** cell-cycling polyploid cells and **(c)** diploid cells. The used six supervised ML models are the Gradient Boosting (GB), K-nearest Neighbours (KN), Logistic Regression (LR), Multi-layer Perceptron (MLP), Random Forest (RF), and Support Vector Machine (SV).

These results show that classification with brightfield and darkfield images with any of the models can significantly recognise the difference between these cell two populations, thus supporting the chapter's hypothesis. Any mislabelling of other types of polyploid cells as the target cells (Figure 5-11, b) is tolerable in our experiment design, so the performance of these ML models can be considered successful for our objectives. Nonetheless, we recommend the tree-based GB and RF algorithms in the label-free classification of the target cells due to their consistent good performance, especially the GB when it comes to the darkfield images, which is consistent with results from previous works on similar applications [2].

5.3.2.2   Classification results using either brightfield or darkfield cell images and input features importance

When it comes to classification using either only the brightfield or darkfield images, the overall perofmance for all model (Figure 5-11, a) saw nearly no effect when the darkfield images were omitted. On the other hand, omitting brightfield images resulted in a drop of approximately 5% in the overall recall values acrosss most models, with a drop of more than 10% for the KN model, while the LR one showed the least change. Other than the type of the used image, one of the main differences between the features we extracted from the brightfield and darkfield images is the cell's overall area and shape measurments. This feature module is directly related to the cell body described by brightfield images. Thus, it is not computed for darkfield images as they show instead a granular pattern without a definitive cell boundary (see Figure 5-3). However, before we discuss the impact of the indivisual feature modules, we first address how well the models perfom when the brightfield and darkfield images are used indivisually to predict our populations of interest.

In Figure 5-12, dropping either of the darkfiled or the brightfield images from the classifcations showed no significant changes in the capablity of all models to recognise that the defining features of the majoirty of the cell-cycling polyploid cells are different from those of the diploid cells, and vice versa. This, once again, supports the research hypothesis on that regard. One notable difference, however, can be seen in the increase of the mislabelling of other polyploid cells as target polyploid cells when only the darkfield images are used (Figure 5-12, b). This explains the previously mentioned

drop in the overall recall values for the darkfield compared to the other image sets (Figure 5-11, a). This agrees with the side scatter performance from the previous chapter, knowing that darkfield images are the same in principle except that they are two-dimensional.
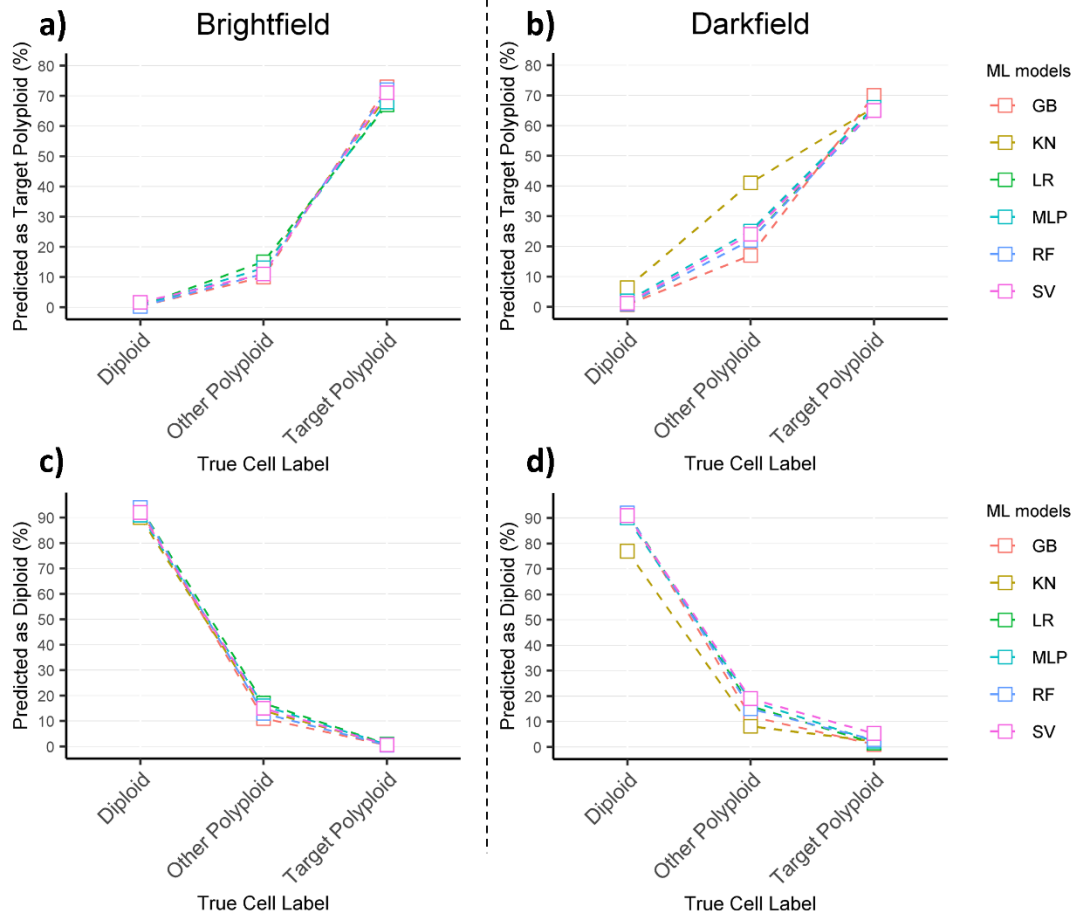
Figure 5-12: Plots showing the classification performance in terms of predicting the diploid and target cell-cycling polyploid cells, when all features extracted from either the brightfield (BF) **(a, c)** or the darkfield (DF) **(b, d)** images are used. The used six supervised ML models are the Gradient Boosting (GB), K-nearest Neighbours (KN), Logistic Regression (LR), Multi-layer Perceptron (MLP), Random Forest (RF), and Support Vector Machine (SV).

Next, we use the leave-one-out approach to report on feature importance in Figure 5-13. When classifying with brightfield images (Figure 5-13, a), there appears no major effect on the overall prediction performance when either feature module is dropped. This applies to their area and shape features, which were the only module missing from the darkfield images. However, in the previous results section, we saw how cell size is a good estimator for the studied cells. This suggests that there is a high

corelation between the brightfield feature modules, or at least for some of them. Intuitively, brightfield captures transmitted light and some forward scatter. For cells that are highly transparent in visible light, it can be seen how measurements of intensity for example can correlate with cell size. A small dip in the performance of some models is seen when the intensity feature module is dropped, but it does not appear significant when compared to the one observed in darkfield images next.
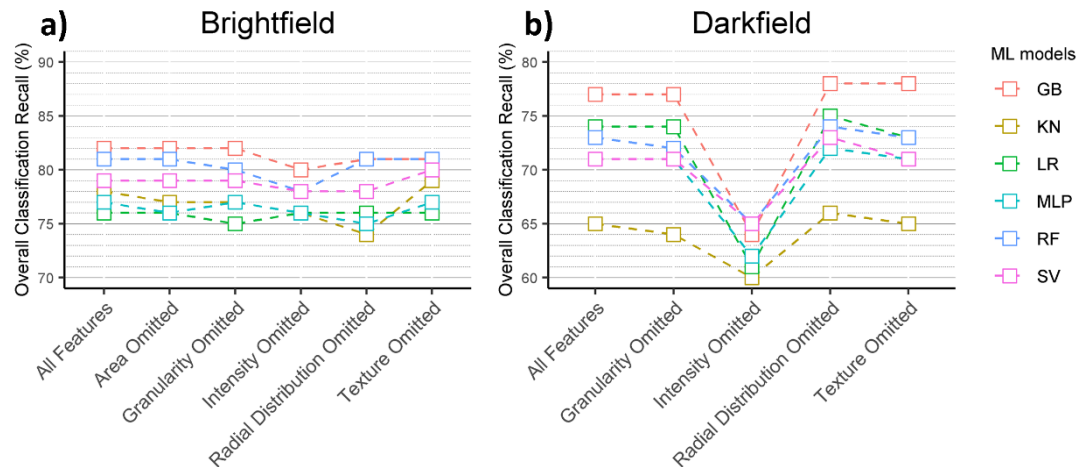
Figure 5-13: The overall classification recall scores for the brightfield (BF) (a) and the darkfield (b) images against the type of feature modules used as a input. The used six supervised ML models are the Gradient Boosting (GB), K-nearest Neighbours (KN), Logistic Regression (LR), Multi-layer Perceptron (MLP), Random Forest (RF), and Support Vector Machine (SV).

For darkfield images (Figure 5-13, b), we can see a noticeable drop in the classification performance across all models only when the intensity feature module is omitted. Radial distribution does measure intensity as well, but omitting it had no such negative effect. These suggest that the darkfield image classification is highly reliant on the overall intensity of the cells rather than the intensity distribution within. This then implies that the significant intensity shifts observed in the side scatter of diploid and polyploid cells (Chapter 4) are most likely the result of a proportionate increase in the side scattering organelles within the polyploid cells, rather than it being a localised change in intensity unique to the polyploid cells. It is worth mentioning that increasing the step size for the texture measurements on CellProfiler (1.5 µm in the reported results) appeared to negatively reflect on the classification performance. Nonetheless, these findings do not exclude the potential significance of texture or granularity related

features that may not have been well-represented by the low detection resolution of the machine.

Lastly, we would like to clarify that some of the features calculated by the CellProfiler could appear repetitive when it comes to analysing images of healthy cells of the same cell line, especially their brightfield images. For example, measuring the intensities at the edge of cells could be meaningful in other applications with different cell lines or with FL images. However, for our application, this feature is highly correlated with the perimeters measured by the area and shape module. In other words, ML methods do not necessarily pick on intuitively meaningful features in the input data. The high collinearity in the features could also explain how the LR model was on the weakest side of the performers when it comes to brightfield images, but one of the highest performers in darkfield images (which have less collinearity).

5.4   Summary

In this chapter, we have experimentally validated the second hypothesis of the thesis that states that label-free brightfield and darkfield images can be used in the classification of cell-cycling polyploid cells in osteosarcoma against diploid cells. This has been done through supervised machine learning (ML) models using six algorithms on the cell image data collected via ImageStream. The models were set to classify cells collected from two-experiment repeats into three classes: diploid cells, G1 and S phase polyploid cells, and G2/M+ phase polyploid cells (i.e. the target polyploid cells). All models show high classification resolution between the two classes of interest, where less than 1% of either diploid or cell-cycling polyploid cells were mistaken for each other's predicted class. In terms of overall classification, the brightfield images scored higher recall values than the darkfield images. The results also show that classification of darkfield images can be negatively affected when the intensity feature module is omitted. For the used ML models, the Gradient Boosting (GB) tended to be the overall strongest performer when using brightfield and darkfield images together or separately. The chapter also reported from the brightfield images the observed cell size differences between the target cells (a median diameter of 29±2 µm) and the drug-stressed diploid cells (19±2 µm). Based on which, a size detection resolution of at least 2 µm was recommended to detect approximately 75% of the original target cells with

minimum diploid contamination. The overall results suggest that that area and intensity-based features in images can be adequate in identifying the target cells from the diploid cells. However, ML based analysis can provide good assessment of the hundreds of extracted image features when it comes to meeting the desired classification objectives. They could also minimise the misclassification of non-target polyploid cells, where less than 20% of such cells in the original sample were misclassified as target. Lastly, the chapter highlighted in the methods the limitation of the standard labelling protocol of DRAQ5 for cell cycle analysis on ImageStream, and presented our proposed steps to optimise it.

5.5   References

[1] T. Blasi *et al.*, 'Label-free cell cycle analysis for high-throughput imaging flow cytometry', *Nat. Commun.*, vol. 7, p. 10256, Jan. 2016, doi: 10.1038/ncomms10256.

[2] H. Hennig *et al.*, 'An open-source solution for advanced imaging flow cytometry data analysis using machine learning', *Methods*, vol. 112, pp. 201–210, Jan. 2017, doi: 10.1016/j.ymeth.2016.08.018.

[3] P. J. Smith, M. Wiltshire, and R. J. Errington, 'DRAQ 5 Labeling of Nuclear DNA in Live and Fixed Cells', *Curr. Protoc. Cytom.*, vol. 28, no. 1, Apr. 2004, doi: 10.1002/0471142956.cy0725s28.

[4] P. J. Smith, M. Wiltshire, S. Davies, L. H. Patterson, and T. Hoy, 'A novel cell permeant and far red-fluorescing DNA probe, DRAQ5, for blood cell discrimination by flow cytometry', *J. Immunol. Methods*, vol. 229, no. 1–2, pp. 131–139, Oct. 1999, doi: 10.1016/S0022-1759(99)00116-7.

[5]   'CellProfiler   4.0.7   documentation'.   https://cellprofiler-manual.s3.amazonaws.com/CellProfiler-4.0.7/index.html (accessed Jun. 5, 2022).

Chapter 6 :

Overall thesis discussion, conclusions, and future work

6.1   Introduction

In this thesis, we presented our experimental and analytical work that discussed the detection of cell-cycling polyploid cells in osteosarcoma through their label-free parameters on flow cytometry. The thesis started with stating the research problem, hypotheses, and objectives in Chapter 1. In Chapter 2, we gave an overview on flow cytometry, which is the high throughput single cell analysis technique used in the current work, along with a reference to our results to help explain the important concepts and any method limitations. This was followed by Chapter 3 where we established the control environment of the thesis experiments, while Chapter 4 and Chapter 5 discussed our characterisation results of the cell label-free parameters on the CFC and IFC, respectively. In the following sections, we present the thesis overall discussion, conclusions and main findings, as well as the future work and recommendations.

6.2   Overall thesis discussion and conclusions

6.2.1   Thesis hypotheses and primary objectives: results highlights and discussion

Through our experimental and analytical results, we have validated both of our thesis hypotheses via showing that the label-free scatter or imaging features can be used to identify cell-cycling polyploid cells against diploid cells.

For the first hypothesis and primary objective, we have shown that the high forward and side scatter intensities of target polyploid cells can be used to identify them in drug-treated samples with very high yield (approximately 85-100%) and low diploid contamination (approximately 0.02-0.4% of the original diploid population, depending on the thresholds). This was done through analysis on data collected from experiments on CFC instruments, where overall results showed that more than 75% of the cell-cycling polyploid cells had higher forward and side scatter intensities than the diploid cells (area measurement on the FACSVerse). Scatter thresholds were then constructed based on these results (along with those of scatter control references) to perform scatter-based automated clustering to find the target cell population. The clustering

results showed that the purity of the target cells can be optimised by the side scatter of the cells, while forward scatter (area measurement) can be used for general polyploid detection. In addition, the results also showed that the difference in the observed intensity between the diploid and polyploid populations is mainly a consequence of the width and peak measurements of the forward and side scatter signals, respectively.

For the second hypothesis and objective, supervised machine learning (ML) classification with six different algorithms showed that the brightfield and darkfield images can be used in the classification of target cells against diploid cells. The image data were collected from IFC experiments conducted on ImageStream, while the feature analysis and classification were done through CellProfiler. All tested ML models were able to classify the target cells with a high prediction recall (yield) (approximately 68-75%), where nearly 0.3-0.7% of the diploid cells in the original samples were mislabelled as target cells. While both image types were effective where the hypothesis is concerned, the results showed that the brightfield images had a better overall classification performance than the darkfield images. The classification results also showed that brightfield images were minimally affected by omission of individual feature modules, whereas the predications of darkfield images were affected when the intensity feature module was ignored. From the brightfield images, the measured median diameters were 29±2 µm for the target cells, and 19±2 µm for the overall diploid population in the drug-treated samples, which suggests that a minimum detection resolution of 2 µm would be needed to detect 75% of the target cells with approximately 0.2% misclassifications of diploid cells as target. The findings do suggest that direct measurements related to cell size or overall intensity from the images could provide sufficient ground for the target cell detection without the complexity of the ML analysis. Nonetheless, the said analysis can help evaluate the overall performance of the large number of extracted features, as well assess any assumptions about the contributions of the various features to the desired detections. It can also minimise the classification overlap of the target cell with other non-target cells (to nearly 20%).

In conclusion, all the four explored label-free parameters showed their capability of detecting the target cell-cycling polyploid cells against diploid cells. We presume that both forward scatter and the brightfield images benefited from their capability to assess

the size changes between the cells, along with the changes in transmitted light for brightfield images. For side scatter and darkfield images, both showed sensitivity to the overall intensity measurement between the two population of interest, which seemed to scale with the perceived size changes between them. The findings also suggest that polyploid cells only assume such high scattering features or significant size changes after they commit to a polyploidy cell cycle. The sources for which are assumed to scale from their previous diploidy state, be it in terms of cell content or overall shape. Nonetheless, the findings are limited by the resolution of the used detectors (especially the CCD cameras), so it is possible for unique scatter patterns to exist between the two phenotypes in higher resolution. It is also possible that the observed scatter sources within the cell may be biologically different between the phenotypes, despite sharing the same overall scatter intensity distributions.

It should be noted that both hypotheses were assessed through DRAQ5-labelled samples to help recognise each population for the scatter analysis, but the labelling was found to introduce overall scatter intensity shifts. Nonetheless, we expect the analysis results to apply to label-free samples, as well. This is because the overall relative shifts observed between the different sample types when they were labelled, were also observed between them when they were label-free. Another point to highlight is that the study assessed the polyploid detection against drug-stressed diploid cells. While a shift between the scatter and sizes of the drug-free versus drug-stressed diploid cells was observed, we similarly assume here that our findings (regarding the ploidy-diploidy resolution) will still apply in drug-free samples, except with some possible overall shifts.

6.2.2   Thesis secondary objectives: results highlights and discussion

Beside the main research objectives, the work in this research investigated several aspects relating to the experimental control measures, results reproducibility, and the robustness of the analysis methods. These were presented in the form of four objectives, which are discussed below.

Regarding the first objective, we have presented in Chapter 3 (section 3.4) a general guideline for the handling of the cell samples to ensure reproducibility of the biology

across the different platforms. We also showed how sensitive scatter measurements can be to potential instrumental noise or the illumination power, and presented our approach regarding assessing meaningful intensity shifts (section 3.3.2.2). Some optimisation steps to the standard DRAQ5 labelling protocol were necessary to make our biology readout achievable on ImageStream (see section 5.2.1.3 in Chapter 5). DRAQ5 is commonly used in ImageStream to visualise the nucleus of live cells rather than quantify the DNA. Some works that employed it for cell cycle analysis did so in low concentrations [1], which were not sufficient for cell cycle analysis on heterogeneous samples of actively-cycling human diploid and polyploid live cells (i.e. the studied cell sample). Nonetheless, more optimisation experiments on the proposed protocol are recommended to serve wider applications. More specifically, the tolerance degree of the DRAQ5 labelling concentrations could be optimised in terms of providing sufficient and stable DNA labelling for cell cycle labelling without affecting the instrument through cross contamination. In section 5.2.1.3 in Chapter 5, we also discussed other instrument-based limitations that hindered the cell cycle readout reproducibility on ImageStream, thus requiring for each of its experiments a parallel one on a CFC instrument.

For the second objective, we have unified the gating analysis approach on the data collected by CFC instruments through supervised automated gating tools. This was done through a customised version of DAFi [2] gating script on R updated by this work, which introduced more flexibility such as handling non-transformed gating for cell cycle analysis, as well as accepting FCS files from either analogue or digital CFC instruments (see section 3.2.3.2 in Chapter 3).

For the third objective, we have identified the drug-free cell samples and a set of 4-25 µm polystyrene beads (see section 4.3.1 in Chapter 4 for more) as scatter reference samples for the detection of cell-cycling polyploid cells. We have noted, however, that beads as scatter sample standards are more suited for CFC instruments because of the high pixel saturation in the darkfield images on ImageStream (see Chapter 2 section 2.2.2.2.1). We also noted that the peaks of the side scatter intensity for the reference polystyrene beads can be perceived differently, relative to the cells, by the different CFC instruments depending on their optics (see section 4.3.1.1). The spread in their intensity distributions was also suspected to be sensitive to their travelling position

within the focused stream. All of which should be taken into consideration when conducting scatter experiments across platforms of different systems in terms of optics, fluidics, or electronics. In literature, there have been efforts to define a standardised scatter unit to calibrate the scatter channels in different flow cytometers, such as the scatter cross section [3], [4]. However, this approach is suited for objects in the submicron regime because large particles (even those uniform in composition such as beads) do not exhibit linearity between size and scattered light over all angles. Even if linearity were to be applicable, objects with sizes comparable to the laser beam widths or larger would have variable scatter cross sections due to the uneven exposure. For cells, this can be complex to compute due to the presence of intracellular structures and their variability between the different phenotypes. Uneven exposure of large cells could also make it less practical to use tuneable hydrogel particles to create optical replicates of target cells (in terms of overall refractive index or embedded organelles) [5]. For one reason, this would require a deeper investigation on the scatter sources within the target cell and their distributions (see section 6.3.1 in future works). For another, replicating possibly unique organelle distributions within cells may not be feasible, thus may require different manufacturing procedures of said particles per instrument. Therefore, in our application, we generally recommended the referenced polystyrene bead sizes kit as a starting control set for the side scatter detection of cells, subject to optimisation with different bead sizes per instrument.

The last objective is concerned with defining the specifications for a scatter-based or image-based label-free detection of the target cells on any platform. From our findings and the previous discussions, we echo our remarks that such specifications can highly be dependent on the instrumentation of said platform. For example in terms of optics, this can be affected by the illumination exposure (e.g. light wavelength, spot size, and power), collection angles or the resolution of the detectors. In terms of fluidics, this can be the size of the focused core or its tolerance to any turbulence sources. The thesis discussed an example of the effect of such variations in the instruments through the aforementioned beads controls, where the scatter intensity peaks of cells appeared shifted relative to beads scatter between two CFC instruments, the FACSCalibur and FACSVerse (see section 4.3.1.1). Overall, we have observed that the measured label-free parameters on FC instruments for diploid and polyploid cells result in continuous distributions with a single or double peaks, depending on the enrichment degree of

either phenotype (see Figure 4-2 in Chapter 4). For detection specifications aiming for the least diploid mislabelling within the target cell population on any platform, we recommend the following reference samples or resolution targets for the studied label-free parameters. For forward scatter, we highlight the width measurement, rather than peak intensity, for the target cell detection. For such measurement, the detectors should be able to resolve a size difference of at least 2 µm for cells (based on which, the cell offset, perpendicular to the optical path, should be less than 5 µm from the detection centre). Side scatter generally produces weaker signals compared to forward scatter for particles above the submicron (less than the forward's by a factor of 10^3 [6]). For it to be able of target cells detection, we suggest a system with optics and detectors capable of resolving the side scatter peak intensities of the tested 4-25 µm polystyrene beads set (see beads specification in section 3.2.1 in Chapter 3, Figure 4-12 in Chapter 4, and the discussion in section 4.3.1.1). For either of the brightfield or darkfield images, the special resolution of the detection camera affects the resolution of the measured size or total intensity. Therefore, once again, we recommend a resolution of at least 2 µm/pixel (where the system calibrated images on ImageStream had a resolution of 0.5 µm/pixel in the current work).

6.3   Future work and recommendations

We present here directions for potential future works along with our recommendations. These are listed in three sections based on the area of interest that they could serve.

6.3.1   Scatter sources in osteosarcoma cells

The thesis conclusions suggest that the observed side scatter differences between diploid and polyploid cells are likely a consequence of a size scaling of the latter. Nonetheless, further investigation may provide better insight on this readout. The presented work was limited by the low resolution of the IFC. As a result, some information regarding the distributions of the observed granularity within the cell darkfield images may have been lost. Therefore, we recommend imaging the cells with high resolution microscopy. General image inspections also suggested that regions occupied by the nuclei (whenever it was found in focus, see Figure 5-3 in Chapter 5) tended to be dimmer than other locations in the cells. Further inspections with careful

region masking could provide results that help expand on the overall topic of side scatter sources in cells. Naturally, attention should be paid to the used illumination light characteristics (e.g. wavelength, power distribution, or polarisation), as the resulting scatter could vary in response.

6.3.2   Cell data analysis in flow cytometry

Based on the tested analysis methods, we discuss here two possible directions for cell data analysis optimisation in flow cytometry. The first is concerned with the handling of CFC data from different instruments. We believe that centralised gating analysis could benefit from automated gating tools such as DAFi with further improvement. For example, the number of input configuration files could be minimised on DAFi by including compatible normalisation steps before gating. This could be in the form of automated cluster recognition and gates shift correction based on a reference control. The second is regarding supervised ML analysis on IFC data. The published analysis pipeline for cell classification via CellProfiler Analyst [1], [7] provides a helpful open-source tool that could serve various application. However, its data management can be less user friendly, and its analysis output provides limited flexibility to evaluate the classification performance. For example, the pipeline entails some expertise in SQL and Regular Expression to manage the databases and help clean the input for the classifiers. There also exists some restrictions in handling the input features for the classifier tool. Therefore, we do suggest for works interested in deeper analysis to utilise different tools or directly scripting their ML model tests for better control on the outcomes.

6.3.3   Label-free detection or sorting of polyploid cells

As part of the thesis objectives (section 6.2.2), we have proposed a couple of measures that could assist in the detection or sorting of the cell cycling polyploid cells on any platform. The studied label-free detection of polyploid cells could also be attractive for less complex detection or sorting platforms such as microfluidic devices, provided that they carry well-defined detection systems that can meet the minimum recommended detection resolutions (see section 6.2.2). In microfluidic devices, light detection can be performed with on-chip integrated sources and detectors [8], [9]. For

some of these systems, cell focusing (or sorting) may utilise negative dielectropherises instead of hydrodynamic focusing. This method could require extra steps to optimise the relevant parameters, such as the relative electrical permittivity of cells and media (not covered by the thesis).

When it comes to experimental sorting of rare cells such as the cell-cycling polyploid cells, one generally needs to consider steps for enriching their low numbers for a better sorting yield, recovery, and purity. This can be done through increased drug dose, as was followed in the current work. However, we do recommend reducing the intended sorting yield for the target cells to minimise the diploid cell contamination in processed 24-hour drug-treated samples (see section 4.3.2 in Chapter 4 for more). Sorting experiments should also consider potential diploid contamination from doublet events. The thesis results showed that doublets exclusion via scatter could be less resolved in diploid-polyploid heterogeneous samples (compared to resolving them via the DNA label). Therefore, the quality and the handling of the prepared samples should be observed to minimise cell aggregations (see also the recommended sample handling protocol in section 3.4 in Chapter 3).

For the sort of live osteosarcoma cells, preliminary experiments are recommended to confirm that the live cell suspensions do maintain their cell phase status throughout the sorting. While putting cells on ice could slow their biology, these cells may experience some inevitable changes in the environment temperature (e.g. inside the sorter or during DNA labelling incubation post sort). The suspension media or buffer may also experience concentration changes. All these factors could affect any subsequent sorting validation. Sorting on fixed cells instead of live ones could be performed first, to test the capability of the used instrument in performing the label-free sort with the intended yield, recovery, or purity rates (without worrying about the changing biology). For such case, however, any morphological changes in the cells due to the fixing process should be noted beforehand and characterised for the used instrument.

For benchtop flow cytometry sorters with fluorescence detection, adding a non-toxic labelled parameter to the scatter sorting thresholds, such as cyclin B1-eGFP, could help target the cell-cycling polyploid population. However, this could impact purity by excluding other potentially interesting polyploid cells (e.g. cells at the G0/G1 phase

of the second ploidy cell cycle), while possibly still preserving the original diploid contamination (as they tend to be cells at the G2/M phase of the diploidy cell cycle, see Figure 4-14 in Chapter 4).

6.4 References

[1] T. Blasi et al., 'Label-free cell cycle analysis for high-throughput imaging flow cytometry', Nat. Commun., vol. 7, p. 10256, Jan. 2016, doi: 10.1038/ncomms10256.

[2] A. J. Lee et al., 'DAFi: A directed recursive data filtering and clustering approach for improving and interpreting data clustering identification of cell populations from polychromatic flow cytometry data: DAFi', Cytometry A, vol. 93, no. 6, pp. 597–610, Jun. 2018, doi: 10.1002/cyto.a.23371.

[3] L. de Rond, F. A. W. Coumans, R. Nieuwland, T. G. van Leeuwen, and E. van der Pol, 'Deriving Extracellular Vesicle Size From Scatter Intensities Measured by Flow Cytometry', Curr. Protoc. Cytom., vol. 86, no. 1, p. e43, Oct. 2018, doi: 10.1002/cpcy.43.

[4] L. Rond, F. A. W. Coumans, J. A. Welsh, R. Nieuwland, T. G. Leeuwen, and E. Pol, 'Quantification of Light Scattering Detection Efficiency and Background in Flow Cytometry', Cytometry A, vol. 99, no. 7, pp. 671–679, Jul. 2021, doi: 10.1002/cyto.a.24243.

[5] J. Kim, O. Liu, J. Agresti, and A. T. Nguyen, 'Hydrogel particles with tunable optical properties', US10942109B2, Mar. 09, 2021 Accessed: Jun. 30, 2022. [Online]. Available: https://patents.google.com/patent/US10942109B2/en

[6] N. Catsimpoolas, Ed., Cell analysis. New York: Plenum Press, 1982.

[7] H. Hennig et al., 'An open-source solution for advanced imaging flow cytometry data analysis using machine learning', Methods, vol. 112, pp. 201–210, Jan. 2017, doi: 10.1016/j.ymeth.2016.08.018.

[8] R. Thomas, A. Harrison, D. Barrow, and P. M. Smowton, 'Photonic integration platform with pump free microfluidics', Opt. Express, vol. 25, no. 20, p. 23634, Oct. 2017, doi: 10.1364/OE.25.023634.

[9] A. D. Goater et al., 'Laser micromachining of optical biochips', San Jose, CA, Feb. 2007, p. 645909. doi: 10.1117/12.700899.

Appendices

Appendix A

A python script to automate the conversion of the filtered populations text file by DAFi to FCS 3.0 files

```
# STEP 0: import needed libraries
import cytoflow as flow
import pandas as pd
import tkinter as tk
from tkinter.filedialog import askopenfilenames
import os

# STEP 1: open a window dialog for the user to choose the desired DAFi txt file(s)
root = tk.Tk()
root.withdraw()# To hide the root window

# Each element in the list represents one file path.
files_path_list = askopenfilenames(filetypes = [("TXT Files",".txt")], initialdir="/",
                    title='Please select the DAFi txt output file(s) to be split and converted:')

# STEP 2: read the txt file(s) as pandas dataframe, each element in the list represents one data set.
files_raw_list = [pd.read_csv(file_path, delimiter="\t") for file_path in files_path_list]

#Optional: check that you have imported the first or first two elements correctly by viewing their first 5 lines
for i in range(len(files_raw_list)):
    print(files_raw_list[i].head())
    print("==========================================================================")

#STEP 3: Next we loop to create individual subpopulations.
files_data_subsets_dicts = {}
for i in range(len(files_raw_list)): #this takes each original txt file.
    #use the max value of Population col to decide how many populations u r generating first
    range_max_pop_ID = files_raw_list[i]['Population'].max() + 1 #This allows copying original sample as pop0
    subset_name = ""
    data_subsets_dicts = {}
    for ID in range(range_max_pop_ID):
        #plus 1 above ensures max as a value is included, otherwise it does 0 to max-1
        subset_name = "pop{}".format(str(ID))
        print(subset_name) #this is just to check the key names for each subset is working.
        #first exclude the whole sample (aka pop0)
        if ID == 0:
            data_subsets_dicts[subset_name] = files_raw_list[i]
        else:
            #now filter for all subsets, where an event is part of a subset when TRUE = 0 (FALSE = 1, that's based
on DAFi)
            data_subsets_dicts[subset_name] = files_raw_list[i][files_raw_list[i][subset_name] == 0]
    files_data_subsets_dicts[i]= data_subsets_dicts
    data_subsets_dicts = {} #so on exist it is an empty temp dict.

#optional:  check you imported the first or first two elements correctly by viewing their first 5 lines
for file_key in files_data_subsets_dicts:
    for subset_key in files_data_subsets_dicts[file_key]:
        print(files_data_subsets_dicts[file_key][subset_key].head())
        print("==========================================================================")
    print("==========================================================================")
```

```
#Applying modifications on the values of the subset_keys (i.e. subsets) for each file_key (i.e. file) in the files
dict
for file_key in files_data_subsets_dicts:
    #get pop1 col index from any subsets in it
    pop1_col_index = files_data_subsets_dicts[file_key]["pop1"].columns.get_loc("pop1") #just calls for the
1st subset
    for subset_key in files_data_subsets_dicts[file_key]:
        #choose all cols before pop1 for the current subset_key, then re-write the current subset_key in one line
        files_data_subsets_dicts[file_key][subset_key] = files_data_subsets_dicts[file_key][subset_key].iloc[:,
0:pop1_col_index]

# check you imported the first or first two elements correctly by viewing their first 5 lines
# This helps check the size of the filtered subpopulations where size represented as (# of events, # of
channels in FCS)
for file_key in files_data_subsets_dicts:
    for subset_key in files_data_subsets_dicts[file_key]:
        print(files_data_subsets_dicts[file_key][subset_key].head())
        print(type(files_data_subsets_dicts[file_key][subset_key]))
        print("shape/size of ", subset_key, " is ", (files_data_subsets_dicts[file_key][subset_key]).shape)
        print("======================================================================")
    print("======================================================================")




# STEP 4: create multiple empty Cytoflow experiments datatype (for each file) with the desired (empty)
channels.

# we are making individual experiments instead of tubes, to allow for possibility of some files having
different channels.
files_data_ex_dict = {}
for file_key in files_data_subsets_dicts:
    files_data_ex_dict[file_key] = {subset : flow.Experiment() for subset in files_data_subsets_dicts[file_key]}

# Next you need to add channels with no data to the empty experiments, so they can be smoothly added by
to the metadata
#for i in range(len(data_list)):
#    for channel in data_list[i].columns:
#        ex_list[i].add_channel(channel, data=None)

for file_key in files_data_subsets_dicts:
    for subset in files_data_subsets_dicts[file_key]:
        for channel in files_data_subsets_dicts[file_key][subset].columns:
            files_data_ex_dict[file_key][subset].add_channel(channel, data=None)

#Optional: check that you have created the correct number of empty experiments in the list.
for file_key in files_data_subsets_dicts:
    if len(files_data_subsets_dicts[file_key].keys()) == len(files_data_ex_dict[file_key].keys()) :
        print("Everything is fine! You can keep going!")
        # If this was one script, I would put here the rest of the code.
    else :
        print("ERROR! The length of the empty experiments list does NOT equal the number of imported csv
files!")
```

```
# STEP 5: You have to add conditions that define the tube(s) as it is a required argument in the export
function.

for file_key in files_data_subsets_dicts:
    for subset in files_data_subsets_dicts[file_key]:
        files_data_ex_dict[file_key][subset].add_condition("Subset", "object")

# STEP 6: add the events of your tube to the experiment.
# you also will fill each conditioned tube with a value for its condition(s) to identify it.

for file_key in files_data_subsets_dicts:
    for subset in files_data_subsets_dicts[file_key]:
        files_data_ex_dict[file_key][subset].add_events(files_data_subsets_dicts[file_key][subset], {"Subset" :
subset})




# Optional: check how conditions mostly are defined in one of the data files here.
print(files_data_ex_dict[0]["pop1"].data.dtypes)

#Optional: print the data to double check they filled up fine
# also checks the current metadata (which isn't much except for the channels' names as keys)
for file_key in files_data_ex_dict:
    for subset in files_data_ex_dict[file_key]:
        print("shape/size of ", subset, " is ", (files_data_ex_dict[file_key][subset].data).shape)
        print(files_data_ex_dict[file_key][subset].data.head())
        print("************************")
        print(files_data_ex_dict[file_key][subset].data.tail())
        print("************************")
        print(files_data_ex_dict[file_key][subset].metadata)
        print("=======================================================================")
    print("=======================================================================")
    print("=======================================================================")

#STEP 6: prepare the file names and saving location for the FCS files generated next

# Make a list of original file names:
filenames_list = [os.path.splitext(os.path.basename(files_path_list[i]))[0] for i in range(len(files_path_list))]

# Optional: print to check name example to be used below
print(filenames_list)

# Do the same but for directories (in this example script, this does not do much because we only call one
directory).
directory_list = [os.path.split(files_path_list[i])[0] for i in range(len(files_path_list))]
print(directory_list)

#Also list the desired DAFi gate names for ease
#NOTE: This also assumes same gating hierarchy for all input files.
subsets_names = ["FULL", "Whole_Cells",
        "Singlets_D5",
        "GFP", "No-GFP",
        "Diploids",
        "Polyploids",
        "G0-G1_Diploids", "S_Diploids", "G2-M_Diploids",
        "G0-G1_Polyploids", "S_Polyploids", "G2-M_Polyploids", "G0-G1_Polyploids2"
        ]
```

""

We need Next to add at least the standard metadata to our FCS file to be created.
The scripts of the Cytoflow functions named "export_fcs" and "fcswrite", show that they already count for some of the standard keywords such as $BEGINSTEXT, $ENDSTEXT, $BEGINANALYSIS, $ENDANALYSIS, $BEGINDATA, $ENDDATA, $BYTEORD, $DATATYPE, $MODE, $NEXTDATA, $TOT, and $PAR. In other words, we don't need to input them manually, thus they are commented off in the codes below. The functions are set to always choose $DATATYPE: F, $BYTEORD as '4,3,2,1', and $MODE as L. If one wishes to change these for a good reason, then they need to be changed from within the script of the "fcswrite" function.

In addition, based on the data imported from the input csv file and the metadata range you input yourself for the channels, these functions will also automatically fill in some other FCS standard metadata such as PnN, PnB, PnR.

Even though the functions are made to basically cover writing the main keywords in the accepted format, we may need to input any other FCS file keywords (e.g. $BTIM) in the metdata dictionary key "fcs_metadata", in case the functions gave errors.

Some extra notes:
In the current versions of FlowJo, remember if you opened the DAFi transformed FCS file along with its original one (where both share the same channel names), FlowJo may assume the default display of the axis range based on the first FCS file you import. For example, if the first FCS file has a range of 2^12, that's what will be displayed regardless of whether the second file is higher (e.g. 2^18) or not.

For next step, the current code treats each sample/tube as being almost identical in terms of supposed range etc, but you can fix it.

This is the same for the input metadata, where some are recommended to be input individually in a list above or extracted from the original file names.
"""

```
# STEP 7: add required metadata to each file's experiment!
# Set the channel range according to the input data

ch_r = 4096.0 #set at 4096.0 with FSC3 for original DAFi (1024.0 with FCS2 in the thesis customised version)

for file_key in files_data_ex_dict:
    counter = 0
    for subset in files_data_ex_dict[file_key]:
        files_data_ex_dict[file_key][subset].metadata['ignore_v'] = [] # metadata that are added by Cytoflow's
import files, so done manually here.
        files_data_ex_dict[file_key][subset].metadata['name_metadata'] = '$PnN' # same as previous comment.
        for t, channel in enumerate(files_data_ex_dict[file_key][subset].channels):
            files_data_ex_dict[file_key][subset].metadata[channel] = {'type': 'channel', 'fcs_name': channel,
'range': ch_r}
        files_data_ex_dict[file_key][subset].metadata['fcs_metadata'] = {files_path_list[file_key]:
                                        {'$EXP': 'Basmah Almagwashi (BA)',
#                                         '$BTIM': '13:11:00', '$DATE': '22-Oct-2019', '$ETIM': '13:13:00',
                                         'CF_File': 'These are DAFi-filtered subsets converted to FCS3.0 via
python - BA',
                                         'Subset ID': subsets_names[counter]
                                        }}
        counter = counter + 1

#Optional: show that our changes have been added successfully on the first experiment as an example.
print(files_data_ex_dict[0]["pop1"].metadata)


# STEP 8: export now as FCS3.0 files, which is based on the conditions you had for each sample/tube.

# This defines the sample type; here it assumes all imported files (for a set of txt DAFi files) are same sample
type
sample_name = "24T48R_CON_D5"


#Specify in "by" the conditions you are choosing and in "subset" their values

for file_key in files_data_ex_dict:
    counter = 0
    for subset in files_data_ex_dict[file_key]:
        try:
            flow.ExportFCS(base =
("{0}_"+sample_name+"_{1}_{2}").format(filenames_list[file_key],"{0:0=2d}".format(counter),subsets_name
s[counter]), path = directory_list[file_key],
                    by = ["Subset"], subset = ('"{0}" in
Subset').format(str(subset))).export(files_data_ex_dict[file_key][subset])
            counter = counter + 1
            pass
        except:
            #exception is often raised by flow's export when there is an empty subpopulation in DAFi
            #so count the loop and continue to next iteration without any fcs3 exports
            counter = counter + 1
            continue

# This outputs an FCS 3.0 file for each subpopulation in the DAFi text input.
#There is a possibility that these files won't open on some specific BD software like Diva because they may
have their own required keywords.
```
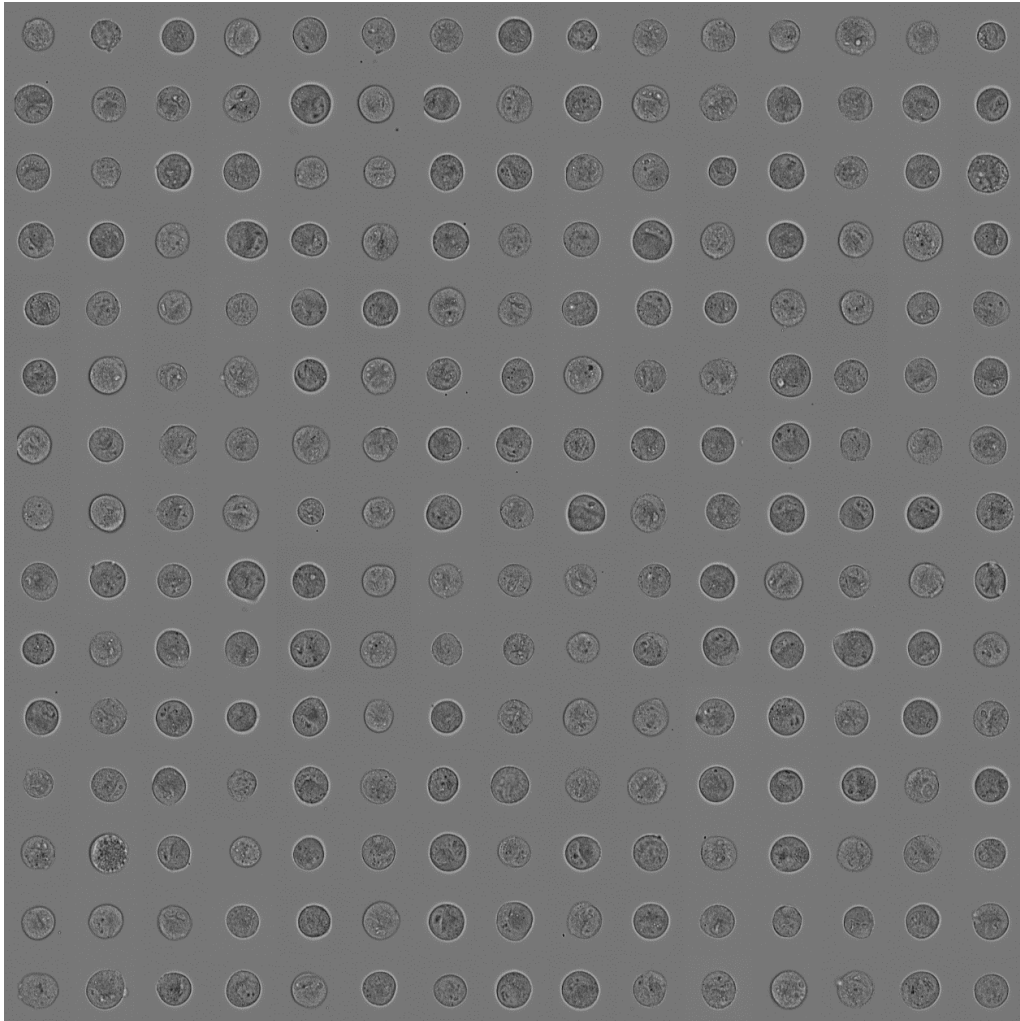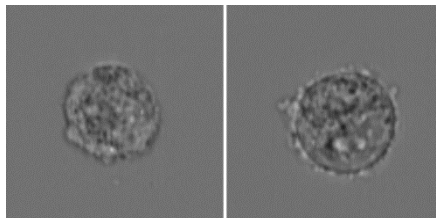
Appendix B

Supplementary Figure B-1: Example of brightfield cell images tiled (15 x 15) via MATLAB for analysis on CellProfiler. The example shows polyploid cells at the G2/M phase.

Supplementary Figure B-2: Two brightfield image examples of unhealthy cells that are excluded upon visual inspection from the selection of the machine learning training set. Both examples are polyploid cells at the G1 phase.