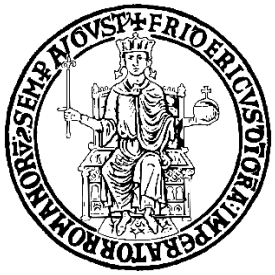
University of Naples "Federico II"



Ph.D. Thesis in
INDUSTRIAL PRODUCT AND PROCESS ENGINEERING - XXXIII
Department of Chemical, Materials and Production Engineering (DICMaPI)
2020-2021

A SINGLE-CELL ATLAS OF BREAST CANCER CELL LINES TO STUDY TUMOUR HETEROGENEITY AND DRUG RESPONSE

Supervisor:
Prof. Diego di Bernardo

Candidate:
Gaetano Viscido

Dedico il mio lavoro di dottorato a Zio Giandonato.

Grazie per esserci sempre vicino.

Table of Contents

SUMMARY 5

CHAPTER 1 - Introduction to breast cancer 7

1.1 – Molecular subtyping of breast cancer 7

1.1.1 – Luminal A and B breast cancer subtypes 7

1.1.2 – HER2 positive breast cancer subtype 8

1.1.3 – Triple negative breast cancer subtype 9

1.2 – The potential of single-cell RNA sequencing in personalized breast cancer therapy 10

CHAPTER 2 - Technologies for single-cell RNA sequencing 12

2.1 – Main technologies for single-cell RNA-seq 12

2.1.1 – Single-cell isolation technology 12

2.1.2 - Illumina next generation sequencing technology 13

2.2 - Droplet-based microfluidic for single-cell transcriptomics. 15

2.3 - Drop-seq platform for high throughput transcriptome profiling of single cells 18

2.4 - Comparison of Droplet-based single-cell RNA sequencing platforms 20

CHAPTER 3 - Implementation of the Drop-seq microfluidic platform 22

3.1 - Drop-seq microfluidic system implementation 22

3.1.1 – PDMS microfluidic device for droplet generation 23

3.1.2 – Magnetic stirrer system 24

3.2 – Barcoded bead functionality test: poly(A) RNA capturing in bulk mode 26

3.3 – Testing the microfluidics implementation 28

3.3.1 – Doublets events are technical artifact that confound scRNA-seq data analysis 29

3.3.2 – Experimental workflow optimization and cell doublet estimation 29

3.4 – Improvement of the microfluidic platform 34

3.4.1 - Implementation of a spiral channel microfluidic device for barcoded bead ordering 34

3.4.2 – Performance test with deterministic barcoded bead encapsulation 35

3.4.3 – Implementation of a new passive filter for barcoded beads 36

CHAPTER 4 - Single-cell transcriptome profiling of breast cancer cell lines (CCLs) for automated cancer diagnosis 38

4.1 – Single cell RNA-seq of breast cancer cell lines. 38

4.1.1 – Selection of comprehensive panel of breast cancer cell lines. 38

4.1.2 – Single-cell RNA-seq of CCLs 38

4.2 – Sequencing reads alignment and gene expression quantification 39

4.3 - Breast cancer cell line single-cell atlas construction 40

4.4 – Biomarker analysis of Breast Cancer Cell lines in the atlas. 41

4.5 – Single-cell RNA sequencing captures the expression of clinically relevant signatures across CCLs 42

4.6 – The breast cancer single-cell atlas for automated cancer diagnosis 44

CHAPTER 5 - Intrapopulation gene expression heterogeneity of cancer cell lines *47*
5.1 – scRNA-seq shows intrapopulation heterogeneity within CCLs **47**
5.2 – HER2 expression state is dynamically regulated in MDA-MB-361 cells **48**
5.3 – Cell cycle analysis **49**
5.3.1 – Cell cycle *in silico* prediction of the MDA-MB-361 cell line 51
5.3.2 – DNA staining protocol optimization for cell cycle analysis 51
5.3.3 – Nocodazole and HBSS protocol optimization for cell cycle arrest and analysis 53
5.3.4 – Assessment of the cell cycle contribution to the HER2 of the MDA-MB-361 cell line 54
CHAPTER 6 - Impact of gene expression heterogeneity on drug response *56*
6.1 – Drug sensitivity correlates with intrapopulation drug target heterogeneity **56**
6.2 – DREEP: a bioinformatics tool to predict drug response at the single cell level **59**
6.3 – Experimental validation of DREEP drug sensitivity prediction **61**
6.4 – Mathematical model to explain the HER2 state dynamic interconversion **64**
6.5 – Experimental validation of the modelling prediction **67**
6.6 – Normalized growth rate inhibition to assess the afatinib effect on growth rate **70**
CHAPTER 7 - Conclusions *72*
References *73*
APPENDIX A - Materials and Methods *82*
Cell culture **82**
Drop-seq platform set-up **82**
Cell clustering and identification of marker genes **82**
TCGA BC bulk expression dataset and deconvolution into BC cell-lines **82**
Spatial sequencing data **83**
Embed new cells into the BC atlas and prediction of the cancer type **83**
Single-cell drug sensitivity prediction **83**
Differential drug sensitivity prediction between HER2+ and HER2- cells in the MDA-MB-361 cell line **83**
Validation of drug sensitivity prediction **84**
Prediction of cell cycle phase from scRNA-seq **84**
HER2 antibody staining procedure for flow cytometry analysis **84**
HER2 expression dynamics experiment **84**
Drug sensitivity assay **85**
APPENDIX B – Mathematical model of the HER2 interconversion dynamics *86*
APPENDIX C – Supplementary Figures *87*
APPENDIX D – Supplementary Tables *93*

SUMMARY

Breast Cancer (BC) patient stratification is driven by receptor status and histological grading and subtyping, with about 20% of patients for which absence of any actionable biomarkers results in no clear therapeutic intervention. Clinical decision for breast cancer patients still relies primarily on the expression status of three biomarkers of therapeutic agents: the estrogen and progesterone receptors (ESR1 and PgR, respectively), and the aberrant expression/amplification of the epidermal growth factor receptor 2 (HER2/ERBB2). However, current clinical approaches for the diagnosis of such biomarkers do not account for the whole transcriptional landscape of the cell and the intrapopulation gene expression heterogeneity of tumors, that may be responsible for drug resistance in cancer patients. It is therefore necessary to discover and establish new predictive and prognostic biomarkers for patient stratification and personalized medicine that take into account tumor heterogeneity.

Here, I evaluated the potentiality of single-cell RNA-sequencing (scRNA-seq) for automated diagnosis and drug treatment of BC. To this end, I implemented Drop-seq in the lab, a droplet-based microfluidic platform that enables to measure the gene expression profile in single-cell for thousands of cells. By means of Drop-seq, I transcriptionally profiled 35,276 individual cells from 32 cell lines covering all BC subtypes, showing that with scRNA-seq we successfully measured the expression of clinically relevant receptors. This breast cancer single-cell atlas can be used to computationally map single cell transcriptional profiles of patients' tumor biopsies to the atlas to determine their composition in terms of cell lines. By this approach, I found that each tumor is heterogeneous and composed of multiple cell lines mostly, but not exclusively, of the same subtype. I observed that in most cell lines there is a high degree of heterogeneity in the expression of BC receptors. I focused on whether such heterogeneity impacts a cell line's overall drug sensitivity. By correlating the percentage of cells expressing a given drug target (e.g. HER2, etc.) to the known toxicity of the relevant drug across the 33 cell lines, I observed a significant negative correlation (the higher the % of cells, the higher the toxicity). I then focused on the MDA-MB-361 cell-line of the luminal B subtype with a gain in genomic copy number of the locus containing the *ERRB2* gene coding for HER2. Despite HER2 amplification, scRNA-seq showed that only about 70% of cells express its mRNA. To investigate the origin of this heterogeneity, I performed fluorescence-activating cell sorting (FACS) to isolate HER2 expressing cells (HER2+) from non-expressing cells (HER2-) in the MDA-MB-361 cell population. After approximately three weeks, both subpopulations re-established the original heterogeneity, thus showing that heterogeneity in HER2 expression in these cells is dynamic and not regulated by genetic mechanisms. This observation led us to the development of a bioinformatic approach named DREEP (DRug Estimation from Expression Profiles) to automatically predict responses to more than 450 anticancer agents starting from scRNA-seq and confirmed the validity of the approach using published large-scale studies on drug sensitivity.

Application of DREEP to the MDA-MB-361 cell line identified drugs able to selectively inhibit the growth of the HER2- subpopulation. Etoposide was predicted to selectively inhibit the growth of the HER2- cells but not HER2+ cells. I experimentally validated the DREEP prediction of the effect of etoposide on the HER2- subpopulation. However, DREEP predicted afatinib, a specific and selective HER2 inhibitor, to be equally effective on both subpopulations, even though HER2- cells do not express the target of afatinib. Surprisingly, the experimental validation that I performed confirmed this counter-intuitive prediction. We thus developed a mathematical model

to explain this counterintuitive result, in which we show that the afatinib treatment has the same effect on both subpopulations if the interconversion time between the two HER2 states is comparable to the cell cycle duration. Finally, I experimentally validated the model prediction by testing the interconversion dynamics of the HER2 state upon afatinib perturbation in MDA-MB-361 cell line.

In Chapter 1, I summarize the current molecular stratification of breast cancer, highlighting the main molecular features of each subtype with a brief discussion of some therapeutical strategies. I motivate the potentiality of single-cell RNA sequencing as a powerful method for cancer diagnosis.

In Chapter 2, I focus on the relevant technologies for single-cell RNA sequencing. I illustrate the next generation sequencing (NGS) with Illumina technology and droplet-based microfluidic as a powerful tool for single-cell RNA sequencing. I introduce Drop-seq, a droplet-based microfluidic platform that enables single-cell transcriptome profiling of thousands of cells, with low costs. Finally, I compare the Drop-seq performance with other droplet-based microfluidic platforms.

In Chapter 3, I describe in detail the Drop-seq platform that I implemented in the lab, the main components of the system, and all tests I carried out to set an optimized experimental procedure to perform single-cell RNA sequencing of breast cancer CCL. In addition, I show the improvement that I operated to the microfluidic implementation.

In Chapter 4, I show the single-cell RNA sequencing of a panel of 32 breast cancer cell lines and the generation of a breast cancer single-cell atlas. I aslo show the potentiality of the atlas for automated breast cancer diagnosis.

In Chapter 5, I focus on the intrapopulation biomarker heterogeneity within cancer cell lines, and specifically on the heterogeneous expression state of HER2 in the MDA-MB-361 cell line. I also show that the HER2 heterogeneity is driven by non-genetical mechanisms. Finally, I investigate a possible role of the cell cycle as a driver of the HER2 heterogeneity in MDA-MB-361.

In Chapter 6, I show that the intrapopulation heterogeneity of a drug target (i.e. HER2) affects drug response against that drug target inhibitors. I describe DREEP (DRug Estimation from Expression Profiles), an algorithm that I contributed to develop and that is able to automatically predict the drug response to more than 450 anticancer agents starting from scRNA-seq. By applying DREEP to the MDA-MB-361 cell lines, I demonstrated that etoposide has a specific effect on HER2- cell subpopulation.

In Chapter 7, I draw final considerations of the possible outcomes of my work.

CHAPTER 1 - Introduction to breast cancer

Breast cancer (BC) is one of the most frequently diagnosed cancer in women worldwide [1]. Currently, therapeutic improvements have led to increasing chances for a cure in about ~70% of early breast cancer patients, however advanced breast cancer with distant organ metastases is considered incurable with currently available therapies [1]. Moreover, BC is a highly heterogeneous disease composed by multiple subtypes [2]. Immunohistochemical (IHC) biomarkers, together with traditional clinicopathological variables including, tumor size, tumor grade and nodal involvement, are conventionally used for patient prognosis and management [2], [3]. In this Chapter, I summarize the molecular classification and subtyping of breast cancer and breast cancer cell lines (CCLs).

1.1 – Molecular subtyping of breast cancer

The current breast cancer stratification relies on the systematic detection of the expression status of clinically relevant biomarkers, in particular the estrogen and progesterone receptor (respectively ESR1 and PgR) and overexpression or aberrant expression of the epidermal growth factor receptor 2 (HER2/ERBB2).

Gene expression profiling of breast cancer has identified two biologically distinct ESR1 positive subtypes of breast cancer, defined luminal A and luminal B [3], which are stratified according to the HER2 status. Luminal A breast cancer is characterized by the sole positivity for ESR1 and/or PgR, while the luminal B subtypes shows in addition positivity for HER2 [2]. Aberrant expression of HER2 with ESR1 and PgR negativity characterizes the HER2 positive (HER2+) breast cancer while tumors with no expression of those three biomarkers define the triple negative breast cancer (TNBC). Table 1.1 and Figure 1.1 summarize this classification.
Each breast cancer subtype differs for risk factors, clinical grade, histopathological features, outcome, and response to systemic therapies [4].

	Immunoprofile			
Subtype	**ESR1**	**PgR**	**HER2**	**Other Names**
Luminal A	+	+/-	-	**Luminal**
Luminal B	+	+/-	+	
HER2 positive	-	-	+	
Triple Negative A	-	-	-	**Basal A**
Triple Negative B	-	-	-	**Basal B or claudin-low**

Table 1.1 – **Breast cancer subtyping**, adapted form Dai et al. (2017). ESR1 = estrogen receptor; PgR = progesterone receptor; HER2 = ERBB2 = epidermal growth factor receptor 2.

Luminal tumors are the most common subtypes of breast cancer, with luminal A constituting the majority of the cases [2], [5]. ESR1 plays a crucial role in breast carcinogenesis, whose inhibition forms the mainstay of breast cancer endocrine therapy [4]. ESR1 positive tumors are largely well-differentiated, less aggressive, and associated with better outcome after surgery than ESR1 negative ones Luminal B cell lines are, in principle, more invasive and consequently more aggressive than luminal A cells, as HER2 overexpression is shown to be associated with ESR1 downregulation [2], [6].

Luminal features of breast cancer include expression of luminal cytokeratins 8/18 (KRT8/18) [7], and transcription factors like FOXA1 and GATA3, that has been shown to be involved in the expression regulation of ESR1 target genes and associated with favourable prognosis [8], [9].

Overall, luminal cancer cell lines are comparably more differentiated to the other subtypes, and have less propensity for migration due to tight cell-cell junctions, consistent with that at the tumor level [2]. However, luminal A tumors have higher expression of ESR1-related genes and lower expression of proliferative genes than luminal B [10], [11], with luminal B tumors characterized by higher grade and proliferation and poorer prognosis than luminal A tumors [3].

Luminal breast cancer patients benefit from endocrine therapy, that can be administered for 5-10 years, such as tamoxifen or aromatase inhibitors therapy [12].

1.1.2 – HER2 positive breast cancer subtype

The HER2 positive subtype is present in 13-15% of breast cancers [1] and is characterized by overexpression of HER2 (chromosome 17) caused by gene amplification, as assessed by immunostaining or fluorescence in situ hybridization (FISH) [13]. Cell lines that are classified in this subtype are heterogeneous and encompass both luminal and basal features. HER2 positive tumors are more aggressive and show higher cell migration behaviors than luminal, since HER2 over-expression is associated with the breakdown of cell-cell junctions [2], [6]. Overall, HER2 positive tumors are characterized by a poorer prognosis than luminal ones, due to a higher risk of early relapse in case of no complete eradication of tumor cells [14]. HER2 positive tumors are sensitive to anthracycline and taxane-based neoadjuvant chemotherapy [10], [15], [7]. Besides the chemotherapy backbone, targeted therapy is available for HER2 positive breast cancer patients [12]. Targeted therapies for HER2 positive cancer patients include the monoclonal antibody trastuzumab, directed against HER2, that demonstrated a reduction in the rate of recurrence [16]. However, many HER2 positive tumors show trastuzumab resistance. For example, the HER2 positive cell line JIMT1 has been studied for the resistance against trastuzumab and lapatinib [17]. Other studies show that PTEN loss [18] and CXCR4 upregulation [14] are implicated in trastuzumab resistance. It has been reported that MEK (S217/219), ESR1, TYK2, FASN, GRB7, and MAPK1/3 (Thr202/Tyr204) strongly correlate with trastuzumab response, while SFN, CAV2, GRB2, RB1, and FLNA associated with resistance, highlighting that upregulation of genes involved in insulin/MAPK signaling predicts response to trastuzumab, whereas the mTOR pathway, Toll-like receptor pathway, N-Glycan biosynthesis, and inositol-phosphate signaling are associated with resistance [19].

1.1.3 – Triple negative breast cancer subtype

The TNBC subtype is the most heterogenous type of breast cancer and is diagnosed when no expression of the three clinically relevant biomarkers is detected, for example by IHC (ESR1-, PgR-, HER2-). TNBC subtype accounts for the 15-20% of breast cancers [20]. Dai et al. [2] reviewed the literature to categorized TNBC, including different subtypes, in triple negative A (TNA) and triple negative B (TNB), based on their respective molecular features. TNBC is characterized by high Ki67 and PCNA proliferation markers, and expression of EGFR [2]. TNA cell lines are enriched FOR basal markers like KRT5/6 and KRT14/17, integrins (ITGA6, ITGB4/6), LAMB3, LAMC2, TRIM29, S100A2, SLPI, ANXA8, COL17A1, BNC1, CD10/14/58/59, MET, LYN, CD133, GABRK, VTCN1, BST2, FABP7 [2]. TNB shows expression of mesenchymal features and has been reported to designate the mesenchymal cluster or normal-like/claudin-low [2]. TNB is characterized by gene signatures for extracellular matrix remodeling (COL1A2, COL5A1/2, SPARC, FN1, LOX, TIMP1/3, MMP2/14) and cytoskeletal modification (VIM, MSN) to enable cell migration, expression of collagens (including COL3A1, COL6A1/2/3, COL8A1), indicative and epithelial-to-mesenchymal transition (EMT) process, and other markers of aggressive features and drug resistance (AXL), such as PLAT, TGFB1, TGFBR2, CTSC, PLAU, PLAUR, SERPINE1/2, HAS2, PRG1, as well as stemness features like $CD44^{+}/CD24^{-}$ [21], [22], [12], [23], [2]. Although both TNA and TNB are comparatively more aggressive than the other subtypes, TNA shows more differentiated features than TNB, that phenotypically appear more mesenchymal-like and are more likely invasive.

Overall, TNBC patients have shown poor prognosis when compared to hormone receptor-positive tumors, with increased likelihood of distant recurrence and death within 5 years of diagnosis [24]. Currently, since the ESR1-, PgR-, HER2- status of TNBC and the lack of other molecular targets, for this subtype there is no effective targeted therapy [24], with chemotherapy the only therapeutic strategy for breast cancer patient clinical care. It has been suggested that EGFR could be a possible target for TNBC targeted therapy, however, early phase clinical trials failed to demonstrate a significant activity of EGFR-targeted monoclonal antibodies as well as tyrosine kinase inhibitors [25].

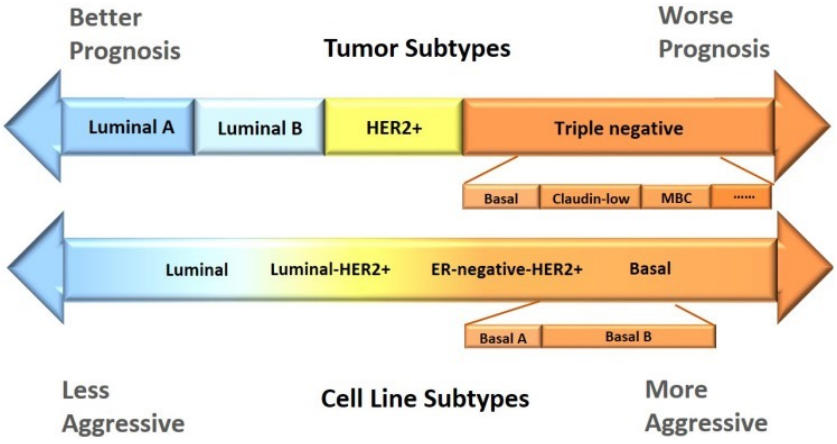


Figure 1.1 – Main features of the different BC subtypes.

1.2 – The potential of single-cell RNA sequencing in personalized breast cancer therapy

One of the main roadblocks to personalized medicine of cancer is the lack of biomarkers to predict outcome and drug sensitivity from a tumour biopsy. Systematic methods for diagnosis and classification of breast cancer patient to drive clinical decisions still relies on the identification of a few biomarkers. For example, ISH or *in situ* hybridization (FISH) are the standard for the HER2 status determination of HER2 protein expression of a cancer biopsy, or assessment of HER2 gene amplification [26]. However, these methods do not account for the whole transcriptional landscape, that may reveal the presence of drug resistant subpopulations and the likelihood of cancer relapse [27]. Expression-based biomarkers measured from bulk RNA-sequencing of a tumor biopsy have been shown to be the most powerful predictors of drug response in vitro [28]–[30]; one limitation, however, is that the population average gene expression measurements (so called *bulk* gene expression profiles) are performed in samples containing mixed populations of cells, and thus the intrinsic heterogeneity of cancer samples cannot be quantified, such as rare tumor subpopulations and subclones that contribute to cell diversity [31].

Multigene assays such as MammaPrint3, Oncotype DX4,5, and PAM506 can classify breast cancer (BC) tumor types and risk of relapse [1]. However, their clinical utility is limited to the prediction of sensitivity to chemotherapy in a subset of high-risk estrogen receptor positive breast cancers [1], [32].

Overall, genomic and transcriptional biomarkers of drug sensitivity have been found only for a restricted number of drugs [28], [29], [33]. As a consequence, BC patient stratification is still mainly driven by receptor status and histological grading and subtyping [1], with about twenty percent [34] of patients for which paucity of actionable biomarkers limits the potential for the development of personalized therapies. Moreover, even when a targeted treatment option is available, drug resistance may arise [1] partly because of rare drug-tolerant cells characterized by distinct transcriptional or mutational states [35]–[38]. Recently, advanced in technologies based on microfluidics, have made measurements of single-cell transcriptomics (scRNA-seq) possible. scRNA-seq yields a molecular profile of each individual cells and thus takes tumour heterogeneity into account, opening up new avenues of investigation, as schematically shown in Figure 1.2. Single-cell transcriptomics offers rapid and comprehensive molecular phenotyping of the tumour at affordable costs, thus making it a prime candidate for routine clinical applications. Understanding the underlying subpopulation and marker expression diversity within the population has the potential to unravel resistance mechanisms that are masked at the *bulk* population level [39].

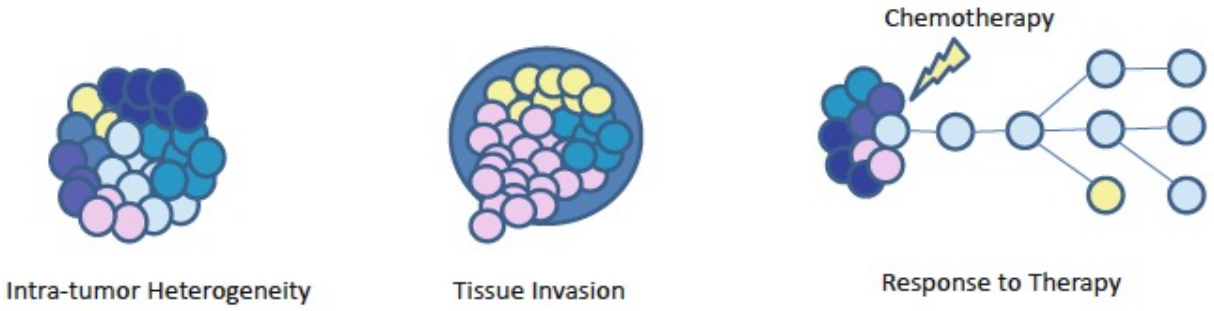


Figure 1.2 – Application of single-cell RNA-seq. Single-cell transcriptome profiling enables to study the intra-tumor heterogeneity that can drive metastasis formation and drug resistance.

CHAPTER 2 - Technologies for single-cell RNA sequencing

In the past decade, *bulk* RNA sequencing (RNA-seq) technologies have been widely used as standard assays to measure the gene expression profile of cell populations and to identify differences between conditions. However *bulk* methods average the gene expression profile across all the cells in the sample, basically assuming that the average response is representative of each cell. Complex cell samples, such as cancer cells, contain heterogeneous populations that include many cell types with different gene expression programs. Single-cell RNA sequencing (scRNA-seq) provides the possibility to explore this gene expression variability at the single-cell level, thus unmasking the existence of different subpopulation with unique behaviors and answering key biological questions such as cell heterogeneity and differentiation.

Conventionally, *bulk* RNA-seq protocols include RNA extraction from samples, that is converted into a cDNA library with sequencing adapters for the so-called Next Generation Sequencing (NGS) technology, which enables massive parallel sequencing of the sample and reconstruction of the transcriptome in term of genes expressed and the relative level. Compared to *bulk*, scRNA-seq includes additional steps, such as the isolation of the RNA from individual cells and tagging the cell-of-origin of the captured RNA by labeling it with a specific cell identifier (i.e. cell barcode). Here, I illustrate the main technologies proposed so far to perform single-cell RNA sequencing of mammalian cells. I show that scRNA-seq platforms differ for the method of cell isolation and the throughput (i.e. the number of cells that a platform has the potential to yield). In what follow, I focus on the droplet-based technologies, and I describe in detail the Drop-seq platform for scRNA-seq. Finally, I conclude with a comparative analysis of Drop-seq with other droplet-based scRNA-seq technologies.

2.1 – Main technologies for single-cell RNA-seq

Single-cell sequencing of RNA, that is scRNA-seq, requires the combination of two main technologies: (i) single-cell isolation technologies to capture the cell transcriptome and generate cDNA libraries for sequencing; and (ii) Next Generation Sequencing technologies for massive parallel sequencing of the libraries obtained from the previous step.

2.1.1 – Single-cell isolation technology

Isolation of single cells from a complex population is the first step of all scRNA-seq methods. Over time, many sensitive and accurate scRNA-seq platforms have been introduced, improving from low to high throughput platforms, defined as the number of cells that a scRNA-seq platform is able process. The number of cells processed reflects the technological advances and it has rapidly increased over the years, from few cells up to hundreds of thousands of single cells [40], (Figure 2.1). The most common implementations of scRNA-seq are well-based, microfluidic-based and droplet microfluidics-based methods [41]. In well-based methods, single cells are deposited manually, or automatically by Fluorescent Activation Cell Sorter (FACS), or within microfluidic chips, into wells that contain cell lysis buffer, oligos with different barcodes and other reagents able to convert the captured RNA into cDNA. For example, Smart-seq2 [42] is a well-based

methods that generates full-length cDNA from FACS isolated cells deposited in a standard 384 well-plates, but yields low cell throughput. Higher throughput methods have been reported susch as Seq-Well [43], a well-based low-cost platform in which cells are deposited by pipetting the cells onto a polymer chip patterned with thousands of micro-wells able to trap single cells. Microfludics-based applications include the commercial Fluidigm C1 [44] platform consisting of an automated microfluidic-based system that can capture and process up to 96 individual cells, generating full-length cDNA; cell capture, lysis, reverse transcription, and cell multiplexing occur in an integrated fluidic circuit chip [45]. Finally, droplet-microfluidics based methods have been recently introduce that have the potential to process thousands of single-cells with high throughput. In Section 2.3, I will describe in detail the Drop-seq technology [46], which is the one I implemented in this thesis, and I will briefly compare it to the InDrop [47] and 10x Chromium (10X Genomics Chromium, 10X Genomics, Pleasanton, CA) technologies.

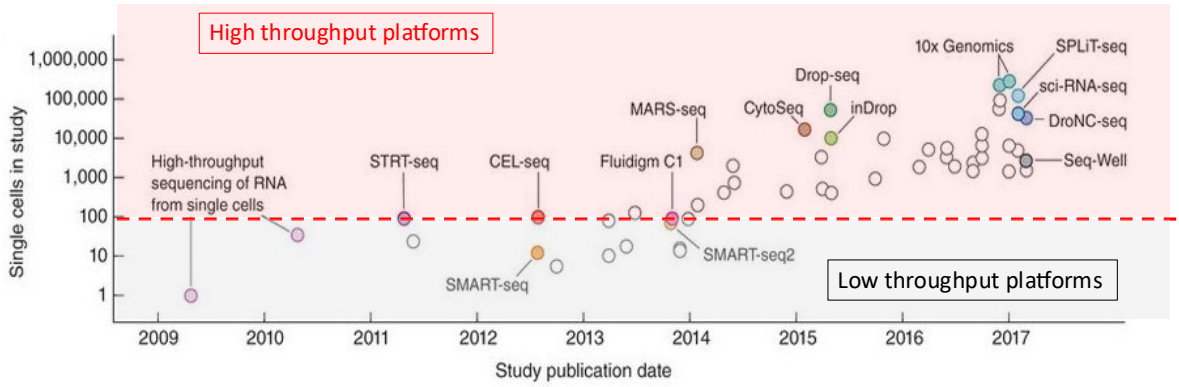


Figure 2.1 – **Technologies for single-cell isolation, developed over the years**. x-axis: publication date; y-axis: the number of cells reported in the study. Adapted from Svensson et al. – 2018

2.1.2 - Illumina next generation sequencing technology

The Illumina technology is currently the standard and the most widely used technology for the sequencing of libraries generated from scRNA-seq platforms.
Overall, Illumina NGS involves three main steps: library preparation, sequencing, and data analysis [48]. Sequencing libraries are typically generated from RNA by first retrotranscribing these to cDNA and then adding Illumina adapters to both ends of cDNA fragments. The cDNA cleavage in short fragments and the addition of adapters relies on several protocols, of which one of the most used is the tagmentation process. In this process, transposase enzymes (tagmentase) are employed to simultaneously cleave and tag with adapters the double-stranded cDNA fragments. Depending on the protocol, this step yields cDNA short fragments with an average length from 200 to above 600 bp. Finally, a limited-cycle PCR enables library amplification with complete adapter sequences.

The library of adapter-ligated short fragments is then loaded onto a solid support (glass slide), defined 'flow cell'. The fragments bind to the flow cell via hybridization of the Illumina

adapters (referred as P5/P7) with complementary oligonucleotide sequences onto the flow cell surface [27], as depicted in Figure 2.2. Subsequently, each cDNA fragment that have bound the flow cell surface is amplified (clonal amplification) to generate clusters composed of thousands of identical copies of the same fragment.

Once each fragment has been clonally amplified, the next step is the sequencing process. The Illumina technology relies on the sequencing by synthesis (SBS) technique. In SBS, a single strand of the cDNA fragments acts as template for the activity of a polymerase, which introduces chemically modified nucleotides to synthetize the complementary strand. Each modified nucleotide contains a fluorescent tag and a reversible terminator. The fluorescent tag consists of a nucleotide-specific fluorochrome that indicates which nucleotide has been added (base call). The role of the reversible terminator is to block incorporation of the next base and therefore further polymerization. In this way, the SBS process consists of multiple consecutive steps, defined 'cycles'. During each cycle, after nucleotide incorporation, first unincorporated nucleotides are washed away, and then the flow cell is imaged by total internal reflection fluorescence (TIRF) microscopy using either two or four laser channels (two color or four color chemistry), for base call. In the four color chemistry, the imaging system is capable of detecting 4 different fluorochromes (each of the 4 nucleotides carries a different fluorescent tag), while in the two color chemistry the system identifies the incorporated nucleotide from 4 different possibilities: detection of the combination of the 2 color, color 1, color 2 or no color. Each of such possibilities correspond to a specific nucleotide. Since each fragment has been amplified in thousands of identical copies (cluster), the incorporation of a fluorescently labelled nucleotide, results in a signal sufficiently above the background noise, to determine which base has been incorporated. However, a quality score is assigned to each base, that indicates how confident is the assignment of each base call by the sequencer.

At the end of each cycle, the reversible terminator is cleaved and so the next base can bind and the sequencing process go on until all cycles are completed. The result is a string of ACGT characters, defined 'read', that represents the nucleotide sequence of a specific transcript. The sequencing process can be accomplished in two modes: sequence only the forward strand (read1) or both forward and reverse strand (read1 and read2). The latter mode is called 'paired-end' mode, that offers several benefits during the bioinformatic analysis. In scRNA-seq the paired-end mode is required to capture both the information from the transcript (usually read2) and the information of a barcode (read1) that specifically indexes the paired transcript, and allow to correctly assign it to the cell from which it come from (see below) [48]. Indeed, the length of the sequenced read1 and read2 depends on the number of cycles during the sequencing process, that in turn depends on the Illumina reagent kit utilized.

Different Illumina sequencing machines provide varying levels of throughput, defined as the total amount of reads that the system is capable of sequencing, including the MiniSeq, MiSeq, NextSeq, NovaSeq and HiSeq models. The MiniSeq provides 7.5 Gb with 25 million reads/run at 2x150bp reads. The MiSeq can perform 2x300 bp reads, 25 million reads for an output of 15 Gb. The NextSeq can provide 120Gb with 400 million reads at 2x150 bp read length [49]. The recently developed NovaSeq 6000 sequencer further improves the sequencing performances, providing up to 500Gb with 2x150 bp kit for the S1 flow cell and up to 3,000 Gb with the 2x150bp read length for the S4 flow cell.

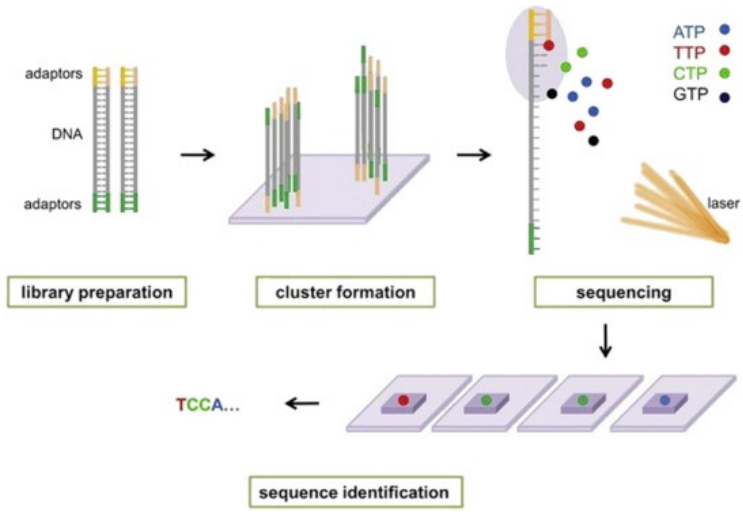


Figure 2.2 – **Main steps of the Illumina sequencing.** The NGS library is prepared by adding adapter sequences to both ends of the double-stranded cDNA fragments. Then, the library fragments bind the solid support of the flow cell, where sequencing occurs. Modified nucleotides are added to the flow cell for the sequencing. These nucleotides have a reversible 3' fluorescent blocker so the DNA polymerase can only add one nucleotide at a time onto the DNA fragment during each sequencing cycle. Wavelength detection of the fluorescent tag by a computer allows to identify what base was added. The process continues until the full DNA molecule is sequenced.

2.2 - Droplet-based microfluidic for single-cell transcriptomics.

Microfluidics technology enable precise formation and handling of small fluid volumes dispersed as droplets that contain just a few fLs to nLs [50]. Droplet-based microfluidic enables to isolate and capture cells with high throughput (up to thousands of cells) in aqueous droplets, to perform large-scale gene expression profiles at the single-call level for transcriptomic studies, as schematically shown in Figure 2.3. Droplets can be generated in the channels of a microfluidic device with high frequency (Hz–kHz) by pressure-driven flows, producing an emulsion, that consists of two immiscible fluids, one of which is dispersed as droplets in the continuous phase of the other. Typically, a surfactant (surface active agents, which mainly act at the oil/water interface by reducing surface tension) is essential to stabilize the droplets against coalescence as they are thermodynamically metastable [50]. In droplet-based microfluidic transcriptomic application, the droplet dispersed phase is an aqueous suspension and the continuous phase is an immiscible inert oil (water-in-oil emulsion). In this way, each droplet behaves as an individual micro-reactor or micro-chamber to encapsulate biological samples, such as cells [50], [51]; compared to conventional cell culture vessels, a single droplet can accommodate up to 10^3–10^9 times less volume, making droplet-based microfluidic a powerful tool to increase the throughput. Droplet generation can be achieved within a microfluidic device channels by active production (that involves the use of valves or electric fields) or passive production methods. Passive production can be achieved using pressure-driven flows and a specific geometry of the microfluidic channels, as shown in Figure 2.3, such as T-junction, flow-focusing, or co-flowing geometry [51]. In the T-junction geometry, droplets are generated when the aqueous phase flow is orthogonally sheared

by oil and thereby generates droplets. The flow-focusing geometry produces droplets by shearing the aqueous stream from two directions. In the co-flow geometry, the aqueous phase is forced through a channel, which is placed co-axially inside a bigger channel, through which immiscible oil is pumped.

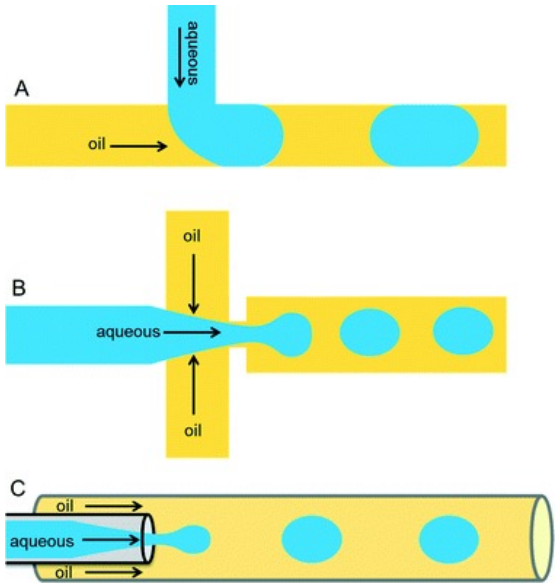


Figure 2.3 – **Different droplet-generating geometries**. (A) In the T-junction geometry, the perpendicular flow of the aqueous phase is sheared by oil and thereby generates droplets. (B) The flow-focusing geometry produces droplets by shearing the aqueous stream from two directions. (C) In the co-flow geometry, the aqueous phase is forced through a capillary, which is placed co-axially inside a bigger capillary, through which immiscible oil is pumped. Adapted from (Shembekar et al. – 2016)

Droplets can be generated by several mechanisms, among which squeezing, dripping and jetting [52]. In the squeezing mechanisms the emerging dispersed phase obstructs the flow of the continuous phase, causing the pressure to rise, that in turn, allows the continuous phase to squeeze on the dispersed phase, forming a droplet. The dripping mechanism occurs when shear stresses overcome the interfacial tension, and drop breakup is caused by the shearing of the dispersed phase by the continuous phase. The jetting mechanism is characterized by the formation of long threads in the dispersed phase, which are broken due to the Plateau–Rayleigh instability, in which liquids, by virtue of their surface tension, tend to minimize their surface area [53].

The flow-focusing geometry in the droplet-generating junction offers a stable dripping mode for a certain range of flow rates, as studied by Moon et al. [54], that optimized the flow rates for droplet generation with a flow-focusing geometry, using water as dispersed phase and oil as continuous phase (Figure 2.4). The flow-focusing geometry allows the formation of a cylindrical thread of the aqueous stream. Importantly, the ratio between the water and the oil flow shapes the thread features at the flow-focusing geometry, such as thickness and breakup behaviors, determining the generation of either a dripping or jetting mechanism as well as either a stable or unstable droplet formation (depending on the consistency of the breakup frequency and uniformity of the resulting droplets). Droplet generation with dripping mode is widely used for cell

encapsulation in droplets [51], [55], that can be achieved by operating a comparatively small water flow rate with a wide range of oil flow rates.

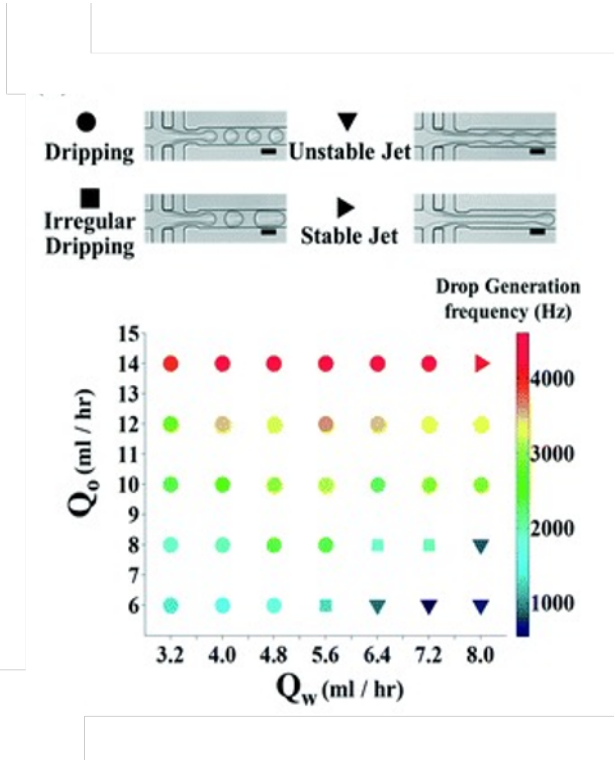


Figure 2.4 – Image showing droplet-generation modes (top) and diagram showing the influence of flow rates of the dispersed water phase and carrier oil phase (Qw and Qo) on the droplet-generation mode and frequency. Adapted from (Moon et al. – 2017).

Soft lithography is one of the most common techniques to fabricate microfluidic devices for droplet-based single cell sequencing [56]. Overall, a microfluidic device consists of micro-channels and ports for input and output material, and, in droplet-based microfluidics, one of the specific channel geometries, described in Figure 2.3. Generally, the microfluidic chip contains at least two layers: a substrate layer, which is usually made of glass or polydimethylsiloxane (PDMS), and another layer with the channel network [51]. In droplet-based microfluidics, the chip material also has to be highly hydrophobic to ensure efficient wetting of the channel walls by the carrier phase, while preventing surface interactions of the aqueous droplets [51]. Channel coating with hydrophobic chemicals, allows to achieve a high degree of hydrophobicity. Some examples are silanes [57], [58] and Aquapel (PPG Industries) [59].

2.3 - Drop-seq platform for high throughput transcriptome profiling of single cells

Drop-seq technology is a droplet-based microfluidic method, first developed in 2015 by Macoscko et al., which enables highly parallel genome-wide expression profiling of individual cells [46]. The Drop-seq method is based on the co-encapsulation of single cells together with barcoded beads in aqueous nL-scale droplets, formed by precisely combining aqueous and oil flows in a microfluidic device (Figure 2.5).

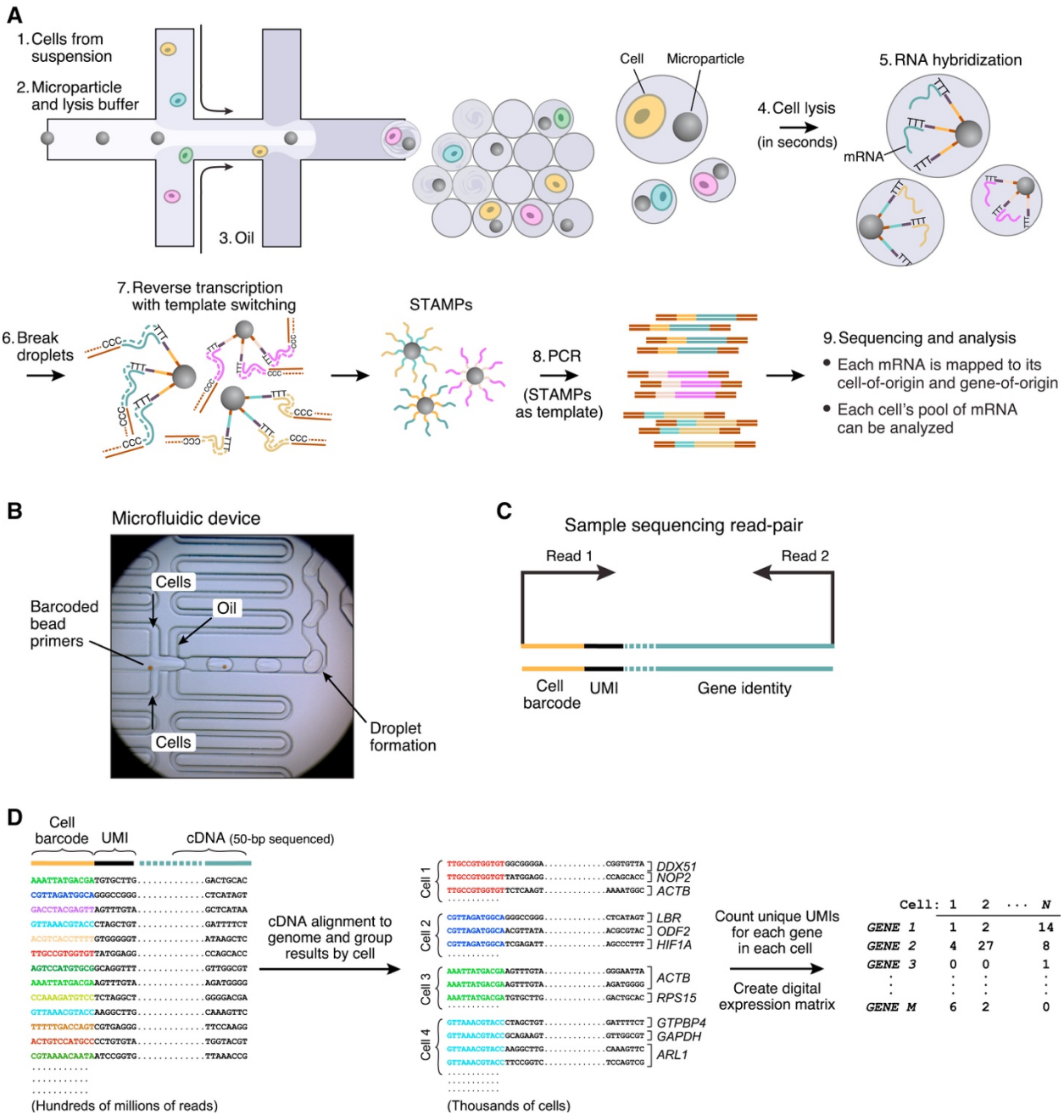


Figure 2.5 – **Overview on the Drop-seq single-cell transcriptome capturing and processing.** (A) Cells and barcoded beads co-flow in the microfluidic devices and get encapsulated in aqueous droplets. Within the droplet environment, cell lysis occurs since the presence of the lysis buffer. Then, the cell transcriptome is captured by the primers onto the barcoded bead surface. Following emulsion breakage, Barcoded beads are recovered, and the transcriptome reverse transcribed in *bulk*, yielding uniquely barcoded STAMPs. Each STAMP represents the transcriptome of the cell of origin. STAMPs are then amplified and prepared for next generation sequencing. (B) Main features of the Drop-seq microfluidic device, highlighting the droplet generation at the flow-focusing geometry. (C) Illustration of the paired-end sequencing mode: read1 covers UMI and barcode, while read2 includes the treanscript. (D) Sequencing reads are aligned to the reference genome, then are organized by their barcode (to assign transcripts to the cell of origin), for each gene in a cell transcript are counted through UMI count, and the digital expression matrix is created. Adapted from (macoscko et al. – 2015)

The Drop-seq microfluidics chip consists of three main channels, one containing a highly diluted cell suspension, another individually barcoded beads in a lysis buffer and a third channel with oil. By means of the flow-focusing geometry of the microfluidic device, millions of nanoliter aqueous droplets-in-oil are generated per hour. Thousands of generated droplets contain exactly one barcoded bead and one cell, whereas the majority contains either no beads, or no cell. Indeed, cells are randomly distributed when arriving at the droplet-generating junction and get encapsulated randomly in droplets according to Poisson distribution (assuming that the droplet volume is much greater than the barcoded bead and cell volume) [51], [60], [61]:

$$P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Where P is the fraction of droplets that will contain k cells (e.g. k =1 is the fraction of droplet that contain 1 cell) and λ is the mean number of cells per droplet and is calculated by multiplying cell concentration by the droplet volume. It should be emphasized that, regardless of λ, the majority of droplets will not contain single cells. Highly diluted cell suspension (e.g. $\lambda = 0.05$) yields only ~5% of droplets with one cell (k =1) but enables to largely reduce the droplets with two or multiple cells (~0.1%, for k =2), at the price of an increased fraction of empty droplets. Hence, setting the cell concentration is crucial to find the optimal trade-off between the occurrence of multiple cells and throughput. The number of cells and barcoded beads that get encapsulated at the droplet-generating junction is described by two independent Poisson variables. The probability of the occurrence of i barcoded beads and j cells per droplet is given by [63]:

$$P(k_b = i, k_c = j) = e^{-(\lambda_b + \lambda_c)} \frac{\lambda_b^i \lambda_c^j}{i!j!}$$

Where λ_b and λ_c are respectively the average number of barcoded beads and cells per droplet.

In droplets containing both a cell and a barcoded bead, cell lysis occurs, and cell's poly(A)-RNA is captured by the ~100 million oligonucleotides attached to the bead surface. These beads are then collected, reverse-transcribed in *bulk* to form STAMPs (i.e. Single-cell Transcriptomes Attached to Microparticles), and cDNA amplified for sequencing. In Drop-seq technology, the barcoded beads consist of 30µm diameter resin microparticles, which bound onto the surface DNA oligonucleotides (primers). The primers on all barcoded beads contain a common sequence (PCR handle) to enable PCR amplification after STAMP formation. All individual primers onto the same

barcoded bead share the same "cell barcode" (12 bp) but have different unique molecular identifiers (UMIs; 8bp). UMIs are molecular tags that are used to detect and quantify unique mRNA transcripts, enabling mRNA transcripts to be digitally counted, and to avoid double-counting sequence reads that arose from the same mRNA transcript. A 30 bp oligo dT sequence is present at the end of all primer sequences for capture of mRNAs. The cell barcode is a strategy to infer the cell of origin of each transcript. In the bioinformatic analysis, all transcripts that share the same barcode, are assigned to the same cell of origin.

The Drop-seq procedure to sequence single-cell transcriptomes consists of collecting all barcoded beads after the droplet generation and reverse transcribe the captured transcriptome in STAMPs. Then, an exonuclease I reaction is performed to chew back primer that have not captured any transcript. STAMPs are amplified and then NGS single cell library are generated for high throughput next generation sequencing. During sequencing, the first read (read1) yields the cell barcode and UMI. The second, paired-read (read2) interrogates sequence from the cDNA. Following data pre-processing, single-cell transcriptomes are reconstructed by computational pipelines. Sequencing reads are aligned to a reference genome to identify the gene-of-origin of the cDNA. Next, reads are organized by their cell barcodes, and individual UMIs are counted for each gene in each cell. The result, is a 'digital expression matrix' in which each column corresponds to a cell, each row corresponds to a gene, and each entry is the integer number of transcripts detected from that gene, in that cell.

2.4 - Comparison of Droplet-based single-cell RNA sequencing platforms

Other strategies employing droplet-based microfluidics have been developed for transcriptome profiling of single cells with high throughput. Currently, there are other two droplet-based systems for high throughput scRNA-seq in addition to Drop-seq, namely the inDrop technology and the commercial 10x Chromium technology. [46], [47], (Figure 2.6). All of these droplet-based platforms are based on the same barcoded bead and cell co-encapsulation methods, and have been demonstrated to be robust at generating NGS single-cell libraries in single-cell RNA sequencing experiments. However, they are based on different barcoded bead manufacturing approaches, barcode design, and cDNA amplification and thus have different experimental protocols. The DNA sequences of barcoded bead primers share a common structure, containing a PCR handle, cell barcode, UMI, and poly-T. However, the beads are fabricated with different materials. The beads used in 10X and inDrop systems are made of hydrogel, while Drop-seq uses resin microparticles [62]. In Drop-seq, the resin microparticles are small hard beads, the encapsulation step follows the Poisson distribution. The capture rate of one bead and one cell within a single droplet follows the Poisson distribution (as shown above; Section 2.3), therefore yielding a large number of empty droplets [62]–[64]. On the contrary, for both InDROP and the 10x Chromium system the hydrogel beads are soft and deformable, closely packed in the microfluidic channel, and their encapsulation can be synchronized to achieve a super-Poissonian distribution (Figure 2.6). This highly affects the cell capture rate and the throughput of the plastform. The capture efficiencies have been reported to be 2-4%, 75% and 50% respectively for Drop-seq, inDrop and 10x chromium. The input cell material is also very different: Drop-seq requires >200,000 cells, from 1,000-2,000 to 10,000 cells for InDrop, and >1,000 cells for the 10x Chromium system [41], [62]. Interestingly, Zheng et al. (2019) in their study show that, although 10x Chromium captured the highest average number of genes per cell (~3,000), Drop-seq captured a comparable number

of genes (~2,500), even higher than InDrop (~1,250) [62]. However, although Drop-seq performs slightly worse than the 10X Chromium system, it is substantially cheaper, making it an attractive choice when the sequencing of a very large number of samples is required. Nevertheless, both Drop-seq and InDrop require operator expertise in the microfluidic field. Thus, the implementation of such scRNA-seq platforms may not be accessible to all laboratories [64].

To conclude, the choice of the most suitable droplet-based microfluidic platform, strictly depends on the research requirements and the field of study. 10x Chromium allows to capture more genes, and, as well as InDrop, to process samples with very few cells, such as biopsy sample. On the other hand, Drop-seq enables low-cost sample processing and high throughput sequencing, at the price of large sample as input and, as for InDrop, microfluidics expertise.

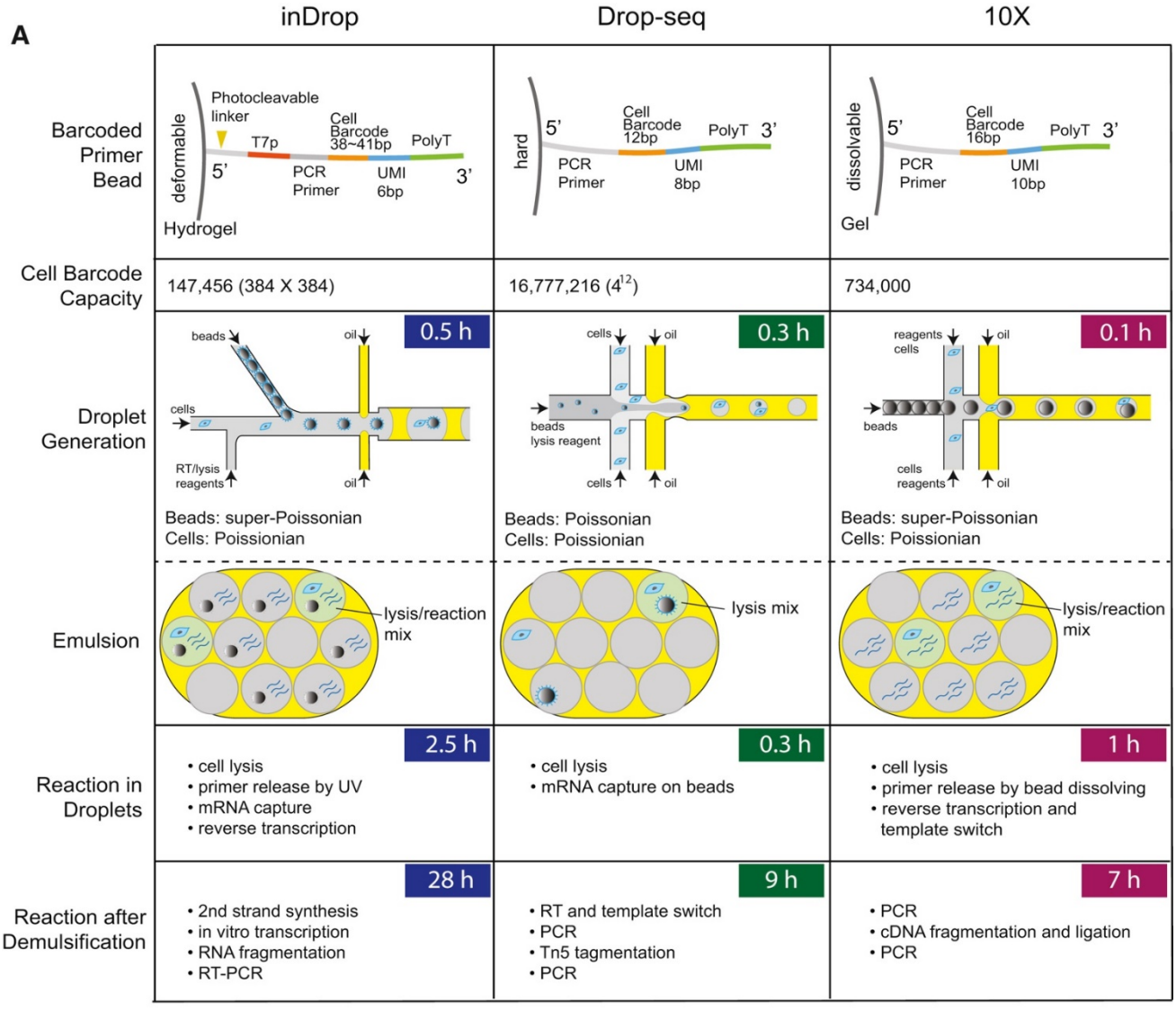


Figure 2.6 – **Schematic and comparison of experimental features of the three systems**. Adapted from (Zhang et al. – 2019)

CHAPTER 3 - Implementation of the Drop-seq microfluidic platform

The Drop-seq microfluidics platform was first developed in 2015 [46] to perform highly parallel transcriptome profiling of individual cells from a complex cell suspension. A detailed description of this technology in reported in Chapter 2, section 2.3. Here, I describe in detail, the components and the functioning of the Drop-seq microfluidics set-up that I implemented in the lab, as well as the protocol I optimized to successfully perform scRNA-seq experiments. To this end, I performed several experimental tests to refine experimental conditions and in addition I carried out a human-mouse mixture experiment in order to estimate the cell doublet rate, defined as the occurrence of two cells together with a barcoded bead in specific conditions. Finally, in collaboration with Gianmarco Nocera, PhD, we optimized the performance of the microfluidic device, by implementing a spiral channel for barcoded bead ordering. We observed an improvement of the filtering efficiency of the microfluidic device, and a reduction in reagent loss during the experiments.

3.1 - Drop-seq microfluidic system implementation

As described in Macosko et al., I implemented the Drop-seq set-up in the lab from scratch. As shown in detail in Figure 3.1A, the microfluidics set-up consists of the following devices:

- Polydimethylsiloxane (PDMS) microfluidics device
- Three Syringe pumps
- Flexible PTFE tubes
- Magnetic stirrer system
- Inverted optical microscope

I employed three syringe pumps to drive three fluids: (I) carrier inert oil connected to port 1 in Figure 3.1B; (II) a highly diluted cell suspension connected to port 2 in Figure3.1B, and (III) barcoded bead in lysis buffer suspension connected to port 3 in Figure .3.1B. Syringe pumps are set to apply constant pressure that drives the flows of cells, barcoded beads and oil from syringes to the PDMS microfluidic device under specific flow rates. I used PTFE flexible tubes to link aqueous flows and oil flow from syringes to the respective microfluidics device inlet ports.
I used syringe pumps in two different orientations: horizontal for cell suspension and oil syringe pump, and vertical for the barcoded bead syringe pump to uniformly distribute the barcoded beads in suspension in the syringe.

3.1.1 – PDMS microfluidic device for droplet generation

The Drop-seq PDMS microfluidic device (Figure 3.1B) consists of the following main components:

- Four ports: 3 inlets and 1 outlet
- Passive filters
- Flow-focusing geometry

The PDMS microfluidic device is composed of three inlet ports, respectively for oil, cell and barcoded bead suspension and one outlet port for collecting droplets, while passive filters implemented to each inlet port help to prevent channels from clogging. The cross-sections of the rectangular channel are 125µm x 100µm (depth x width). Aqueous and oil flows enter the microfluidic device through the respective inlet port; cell and barcoded bead suspensions co-flow and converge with the oil flow in the flow-focusing geometry where single cells co-encapsulation with barcoded beads in aqueous droplets occurs and the water-in-oil emulsion is generated (Figure 1C,D). The droplet outflow in the oil continuous phase then exits the microfluidic device channel through the outlet port in a PTFE tube and is collected in an ice-cold reservoir. A serpentine-shaped channel improves the mixing of the lysis buffer with cell suspension within the droplet, before exit from the outlet port, enabling correct cell lysis for barcoded bead transcriptome capturing. The microfluidic device is placed on an inverted microscope stage: this allows to continuously check the correct droplet generation in the device. The main feature indicative of a good droplet generation is the *triangle* formation (Figure 1E) at the flow-focusing junction where all the flows (beads, cells, and oil) come together to form the droplets. A triangle with well-defined outlines is key to perform a good quality Drop-seq experiment because it implies uniform droplets production and allows to troubleshoot any flow problems that might be contributing to poor droplet quality. Downstream of the triangle, the emulsion outflow should appear blurred without any flickering of the flow, indicative of good quality droplet generation, while well-defined outflow outlines, as well as flickering triangle tip, are both indicative of non-uniform or poor quality droplet generation.

The microfluidics device was produced applying the replica molding technique to fabricate PDMS devices from a silicon wafer (see Methods). Briefly, I treated the Drop-seq custom master-mold under a fume hood for 5 minutes with trimethylchlorosilane (TMCS); then I poured a 10:1 PDMS-curing agent mix onto the master, rest 2 hr in a vacuum chamber to remove bubbles and then baked at 80 °C for 2 hr for PDMS reticulation and polymerization. After the peeling-off the reticulated PDMS, channels and ports are reproduced on the PDMS surface. I used a biopsy punch to drill all ports and subsequently I cut and washed in 2-propanol each single PDMS device. In addition, I washed glass slides with acetone, then water and eventually 2-propanol. After overnight incubation in a vacuum chamber, I performed oxygen plasma irreversible bonding on glass slides, to produce complete PDMS microfluidic devices.

3.1.2 – Magnetic stirrer system

Barcoded beads consist of 30μm mean resin microparticle and this feature causes rapid precipitation in the syringe. I experienced that this represents a weakness of the system since precipitated barcoded beads increase the chance of obstructing the microfluidics system by forming clogs both in syringe/PTFE tubes and in the channels of the device. To solve this issue, as shown in Figure 1F,G, I installed a magnetic stirrer system in close proximity to the bead pump. The system includes two main components: (I) magnetic tumble stirrer with neodymium iron boron magnet, (II) magnetic stir disc (5mm diameter, 1.7 mm thick) placed in the barcoded bead syringe. The magnetic tumble stirrer generates a magnetic field coming from the neodymium iron boron magnet, causing an 'up and down' movement of the magnetic stir disc along the syringe. Thereby, barcoded beads are continuously stirred and resuspended in the lysis buffer medium, preventing precipitation during microfluidic experiments. In addition, the vertical mode of the syringe pump allows to best fit the magnetic stir disc movement along the syringe.

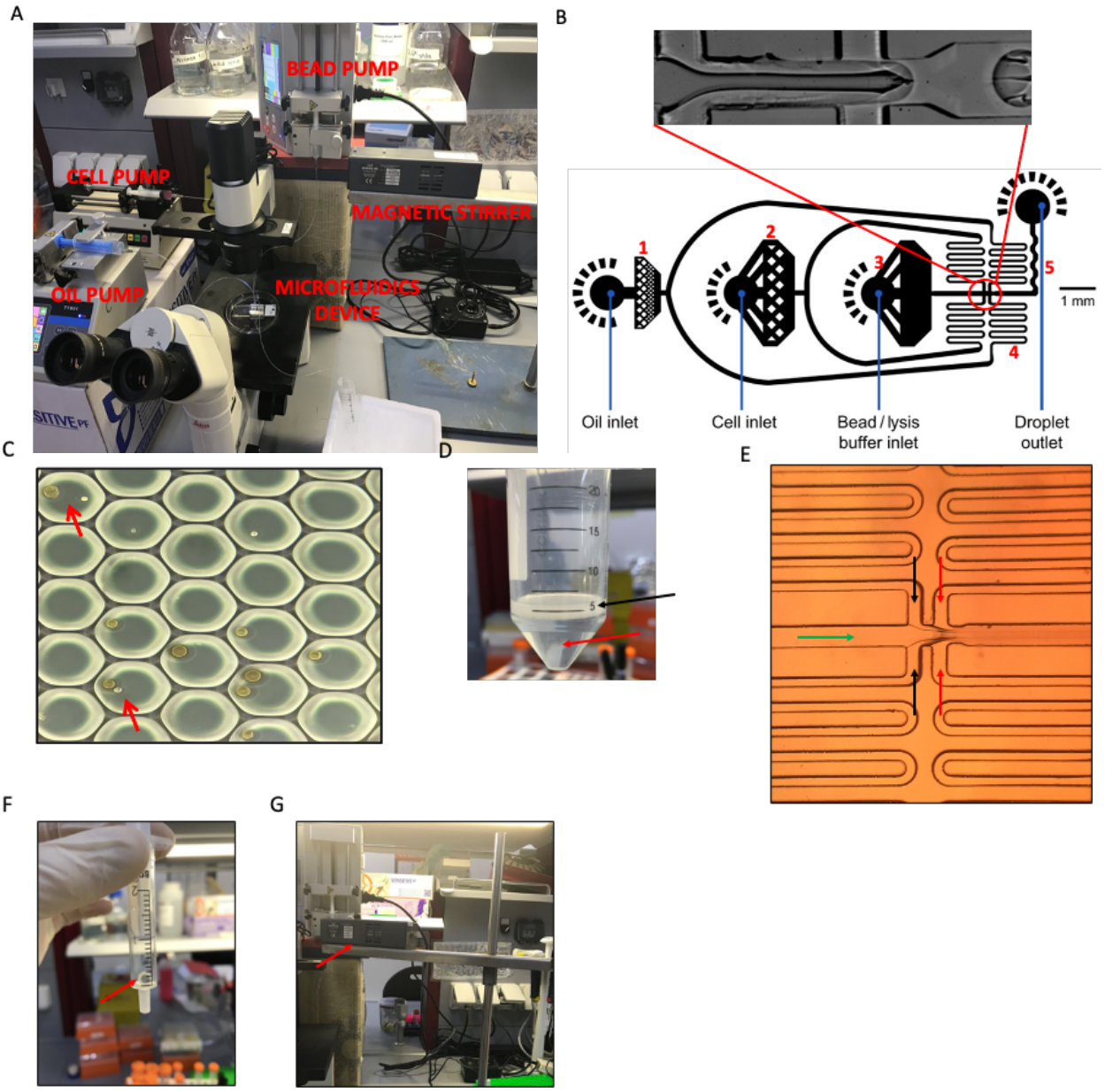


Figure 1. (A) Main components of the DROP-seq microfluidic system implemented in the lab. Aqueous flows and oil flow move from pumps to the microfluidics device, were droplet generation occurs. Droplet are then collected in a tube one ice to preserve RNA integrity. The microscope allows to check the droplet generation. The magnetic stirrer system continuously keeps in suspension the barcoded beads in the syringe in vertical mode. (B) Microfluidic device with a detail on the flow-focusing geometry, adapted from Macoscko et al. 1,2,3: passive filters with pillars at each inlet port. 4: flow resistors to dump mechanical instabilities of the syringe pump. 5: serpentine channel to improve fluid mixing within the aqueous droplet. (C) The picture shows a water-in-oil emulsion where cells (smaller with spheres) are encapsulated with barcoded beads in aqueous droplets. The great majority of droplets are empty, since cell and barcoded microparticle encapsulation follows the Poissonian distribution. Red arrows highlight the occurrence of a cell with a barcoded bead in the same droplet. Here, I did not use lysis buffer in order to show cells. (D) How water-in-oil emulsion appears after collection. The black arrow highlights the droplets (aqueous phase), while the red arrow highlights the underneath oil phase. (E) Triangle formation at the flow-focusing junction, adapted from Macoscko et al. Arrows respectively show channel and direction of the barcoded bead flow (green), cell suspension flow (black) and oil flow (red). The emulsion outflow appears as a continuous blurred stream. (F) magnetic stir disc place into the barcoded bead syringe. (G) magnetic tumble stirrer placed near the barcoded bead syringe to allow magnetic stirring.

3.2 – Barcoded bead functionality test: poly(A) RNA capturing in bulk mode

Following the Drop-seq platform setup, I decided to test the barcoded bead functionality before performing any experiment with the platform. To this end, I mixed barcoded beads with purified RNA in batch mode (i.e. in a tube) to perform RNA poly(A) tail binding, along with poly(A) RNA capturing with oligo(dT)$_{20}$ primer as positive control (Figure 2A) and barcoded bead without any RNA addition as negative control (mock). Oligo(dT)$_{20}$ primer is a string of 20 deoxythymidylic acid residues that hybridizes to the poly(A) tail of mRNA triggering enzymatic RNA reverse transcription. In brief, the test consists of the following main steps: (I) poly(A) RNA capturing; (II) reverse

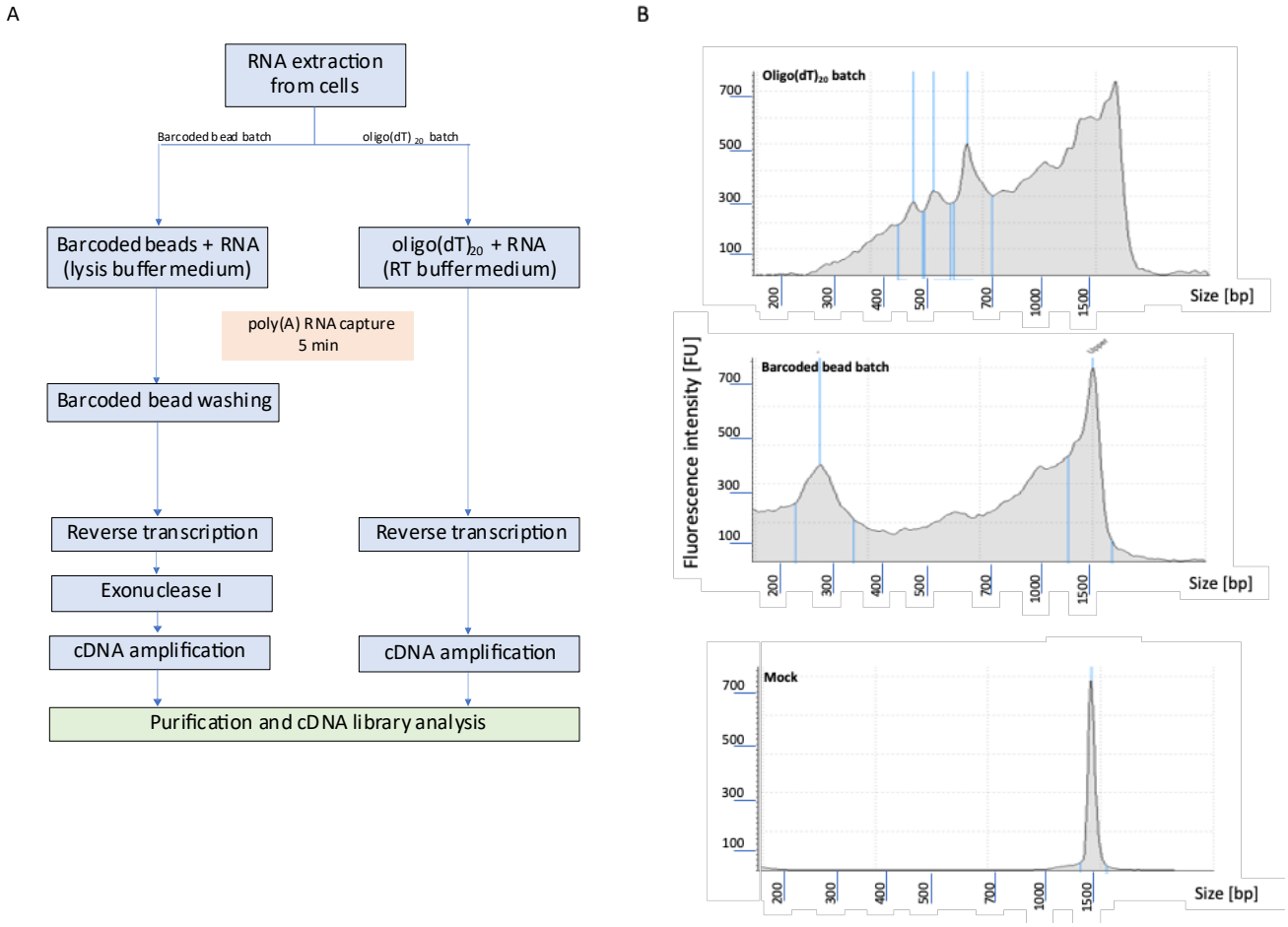


Figure 2. (A) Schematic workflow for the barcoded bead and oligo(dT)$_{20}$ *bulk* RNA-capturing test. I performed both test in parallel. Extracted RNA from cells was captured by barcoded beads and oligo(dT)$_{20}$ in two different batch. The RNA capturing medium was the reverse transcription (RT) buffer for oligo(dT)$_{20}$, on the contrary barcoded beads captured the RNA in lysis buffer medium 1:1 with PBS 1x. I washed barcoded beads several times before reverse transcription reaction, while oligo(dT)$_{20}$ directly underwent reverse transcription. After PCR amplification, I analyzed purified cDNA with the TapeStation chip D1000 high sensitivity. (B) Analysis of the cDNA from barcoded bead batch and oligo(dT)$_{20}$ batch. cDNA signal is detectable both from barcoded bead batch and oligo(dT)$_{20}$ batch, while no signal from the mock negative control (same protocol of barcoded bead but without any RNA sample). For the barcoded bead batch little RNA degradation is detectable, however do not affect the goal of this test

transcription; (III) cDNA amplification and analysis. All the experimental parameters I used for both tests are reported in Table 1.

First, I optimized the medium condition for the barcoded bead hybridization with poly(A) RNA. During a Drop-seq microfluidics run, lysis buffer and PBS co-flow in the channel and mix 1:1 in the droplet; I reproduced this condition by preparing the capturing medium with lysis buffer and PBS in 1:1 ratio, thus keeping the poly(A) capturing environment similar to the droplet. For the oligo$(dT)_{20}$ primer protocol, I decided to use reverse transcription buffer as capturing medium. This decision was driven by the fact that N-lauroylsarcosine (sarkosyl), the cell lysis agent of the lysis buffer used in Drop-seq, is a powerful denaturating agent (see methods). Thus, for the oligo$(dT)_{20}$ primer protocol, I performed the reverse transcription step directly in the capturing medium, bypassing the problem of sarkosyl that would have inactivated the reverse transcription enzyme in the subsequent reverse transcription step. I set the capturing medium volume relatively small, approximately to 21 µL for both tests, to improve the RNA chance to successfully hybridize.

I then combined 400ng of total RNA with either 20,000 beads or 500ng oligo$(dT)_{20}$ primer for 5 min. Then, for the barcoded beads batch, I performed several washes to remove the lysis buffer denaturating condition before going on with reverse transcription, while for the oligo$(dT)_{20}$ primers batch I added directly in the capturing medium the reverse transcription enzyme (see methods).

Both oligo$(dT)_{20}$ and barcoded bead batch yielded successful RNA capture, reverse transcription, and amplification, as shown in Figure 2B, although the barcoded bead batch displayed RNA degradation (shorter cDNA fragments); this was not a problem because I performed this test with the purpose of testing the bead functionality. I analyzed results with TapeStation chip D1000 high sensitivity, but I evaluated this device not suitable to perform cDNA fragment population analysis from transcriptome, since the expected distribution spreads in a wide range around a peak at 1300-1500 bp that is near the upper detection limit of this device. Thus, for the next experiments I decided to switch to the Agilent Bioanalyzer high sensitivity chip, that better fits the bp range of cDNA fragments in my experiments.

	Barcoded bead batch	**Oligo$(dT)_{20}$ batch**
Amount	20,000 beads	500 ng
RNA amount	400 ng	400 ng
Capturing medium	Lysis buffer	RT buffer
Medium volume	21 µL	21 µL
Capturing time	5 min	5 min

Table 1. Summary of all parameters and settings in the barcoded bead functionality test. RT = reverse transcription.

3.3 – Testing the microfluidics implementation

Once I implemented the microfluidics system and assessed barcoded bead functionality, I set to establish a customized protocol to optimize scRNA-seq experiments. Specifically, may aims were:

(I) Customize a tailored protocol for cDNA and NGS library preparation with Drop-seq;

(II) Experimentally estimate the number of sequenced cells, in specific condition (i.e. cell concentration) and microfluidic run duration;

(III) Estimate the doublet rate.

The number of cells obtained (exposed to a barcoded bead) over the total number of cells processed in droplet during the microfluidic run (input cells), has been expressed by Moon et al. (2018) [54] as cell yield; specifically, it is defined as the fraction of cells encapsulated one-to-one with a barcoded bead for all cells encapsulated in droplets, expressed as percentage, and calculated by:

$$cell\ yield = \frac{cell + bead\ \#}{total\ cell\ \#} \times 100$$

Practically, I found more useful the barcoded bead yield:

$$bead\ yield = \frac{cell + bead\ \#}{total\ bead\ \#} \times 100$$

This because during a Drop-seq experiment I work with barcoded beads rather than cells, since cell lyse because of the lysis buffer, to allow barcoded bead capturing of poly(A)-RNA, and hence they cannot be observed. On the other hand, after the microfluidic run, the total amount of recovered barcoded beads can be counted, and from the barcoded bead yields it is possible to estimate the proportion of cell transcriptomes captured (STAMPs), since it expresses the percentage of barcoded bead exposed to a cell among the recovered barcoded beads. This is very helpful to decide how many single cell transcriptomes to sequence over the total barcoded beads obtained from the microfluidic run.

The cell doublet estimation consists of performing a mixed-species experiment: this experiment involves mixing cells from two different species in a 1:1 ratio, typically human and mouse species; captured transcriptomes from isolated cells are sequenced, and then is checked the percentage of species-specific transcripts assigned to each unique barcode. A barcode can be considered species-specific when ≥99% (depending on the protocol), of assigned transcripts match the same genome, either human or mouse genome; on the contrary, a barcode with assigned a mixed pool of human-mouse transcripts (i.e. matched both human and mouse genome) is considered a no species-specific barcode since captured transcripts come from two different species from the occurrence of one barcoded bead with two cells (one human and one mouse cell); such condition represents a cell doublet.

3.3.1 – Doublets events are technical artifact that confound scRNA-seq data analysis

Cell doublets (or multiplets) consist of two or more cells encapsulated with a barcoded bead in the same droplet, causing the bead unique barcode to tag more than one cell. In this condition, the transcriptome coming from multiple cells is recognized as if it belongs to a single cell (or barcode). In droplet-based single-cell methods, the cell doublet rate is defined as the proportion of two or multiple cells occurring together with a barcoded bead in the droplet and therefore tagged with the same barcode. The cell suspension concentration affects the chance of two cells occurrence the droplet; highly diluted cells are encapsulated in droplets according to the Poisson distribution, and increasing the cell concentration, proportionally increases the number of cells processed in droplets, but proportionally increases also the cell doublet rate. From time to time, I also experienced that cell aggregates in suspension contribute to increasing the cell doublet rate. Indeed, cells can stick together in suspension because of DNA release in the medium from dying cells may cause cells to clump together, due to its sticky nature. This is a critical point since high degree of aggregation prevents a successful single cell sequencing experiment.
Barcode doublets or multiplets arise when a cell get encapsulated with two or more barcoded beads and in such event cell transcripts are captured by both barcodes.

3.3.2 – Experimental workflow optimization and cell doublet estimation

In order to perform scRNA-seq experiments, I set to optimize the experimental workflow with the implemented microfluidic setup. Here, in what follows I refer to cell and barcoded bead concentrations as the ones loaded in the syringe. Indeed, in the droplet, each concentration is the half of the concentration loaded in the syringe, since in the microfluidic channel cell and barcoded bead flows mix 1:1 before encapsulation at the droplet generating junction. For example, 100 cell/μL concentration loaded in the syringe yields a cell concentration in the droplet volume of 50 cell/μL. Macoscko et al. reported sequencing data for a set of cell concentrations, and provide a cell reference concentration of 100 cell/μL in the Drop-seq experimental procedure, with a barcoded bead yield of ~5% (percentage of barcoded beads that have been exposed to a cell) and a doublet rate of 1.9% on a total of 1,020 sequenced cells (Drop-seq lab protocol version 3.1); December 28, 2015 I decided to test a further cell concentration of 500 cell/μL while I kept barcoded bead concentration to 120 bead/μL, as protocol. I set flow rates for aqueous suspensions (cells and barcoded beads) to 66.6 μL/min, while oil flow rate to 250 μL/min.

In order to experimentally estimate the cell doublet rate with 500 cell/μL, I performed the mixed-species human-muse mixture experiment. I mixed MDA-MB-453 human breast cancer cell line with NIH-3T3 mouse fibroblast cell line in a 1:1 ratio; cells from the two species were randomly resuspended at the cell loading concentration of 500 cell/μL, in order to obtain in the droplet volume 250 cell/μL final concentration.

I performed cell and barcoded bead droplet encapsulation in the microfluidic device for poly(A) RNA capturing for 15 minutes. Since barcoded beads flow at 66.6 μL/min, with a concentration of 120 bead/μL, ~8,000 barcoded beads get encapsulated per min, and ~120,000 in 15 min. After droplet breakage and enzymatic steps to obtain STAMPs (see Methods), I counted the barcoded beads with a hemocytometer. The resulting total number of barcoded beads I obtained was ~105,000 (average of three counts).

I decided to go on with the PCR amplification with 40,000 barcoded beads, in order to have a barcoded bead reservoir backup in case of amplification failure. From the 40,000 barcoded bead sample, I aliquoted ~2000 beads in PCR tubes, then each aliquot underwent PCR, with the following program:

95 C 3 minutes
4 cycles of:
98 C 20 s
65 C 45 s
72 C 3 min
11 cycles of:
98 C 20 s
67 C 20 s
72 C 3 min
Then:
72 C 5 min
4 C forever

The amplified cDNA population was expected to have the bp average ranging from 1000bp and 1500bp. To purify the cDNA population from the PCR master mix, I used Ampure XP beads for cDNA library clean-up and size selection. I purified the amplified cDNA library using a volume of Ampure XP beads 1.8x the cDNA library sample volume. However, at this ratio, the Ampure XP beads were not capable of excluding primer dimers from the cDNA library, which appear as a strong peak at approximately 100bp (Figure 3A). In order to overcome primer dimer carryover, I decided to change the Ampure XP beads volume to 0.6x the cDNA library sample volume; this ratio resulted suitable to selectively purify the cDNA population while excluding primer dimers. With this purification condition, the resulted cDNA population appeared smooth and with no contaminant and/or primer dimer peak (Figure 3B). The final cDNA library, after Ampure XP beads 0.6x clean up, resulted with the bp average at 1428 bp, and concentrated 11.84 ng/µL.

In the subsequent step, I generated Illumina indexed NGS library from the cDNA library (see Methods for details). I generated four NGS libraries, each with 600 pg of the cDNA library sample (four independent reactions) in order to ensure that the NGS library would have enriched for all captured transcripts. Nevertheless, in other tests, I observed that even one single reaction is enough to this purpose, but I kept four reactions for all experiments I performed. Then, I carried out the PCR as reported in the Drop-seq experimental procedure, to amplify and complete Illumina adapter addition to each of the four tagmented libraries. After PCR, I pooled together the reactions. Subsequently, I purified each NGS library with 0.6x Ampure XP beads, as above. However, since this ratio did not exclude primer dimers from the sample, I directly removed adapter dimers by agarose 2% gel purification, immediately after the 0.6X Ampure XP beads clean up (see Methods). In brief, I run the NGS library on 2% agarose gel, then using a scalpel, I specifically cut and recovered the library population to purify while excluding the primer dimers. Figure 3C shows the NGS library that I obtained with the bp average at 425 bp and concentrated 2.46 ng/µL.

NGS library was now ready for sequencing. Next generation sequencing was performed with the Illumina NextSeq 500 sequencer for high-capacity parallel sequencing. In paired-end mode, shown in Figure 3E, the first read (read1) yields the cell barcode and UMI (12bp barcode + 8bp UMI), while the second read (read2) includes the paired transcript sequence from cDNA. I set

20 bp for base calling of the read1 barcode and UMI (12 bp barcode + 8 bp UMI), while 64 bp for read2 to call the paired transcript sequence. In addition, 8 bp included the read1 index (Figure 3D and Methods).

In collaboration with Gennaro Gambardella, PhD, a computer scientist in the lab, we analyzed the sequencing data from the Drop-seq human-mouse experiment. Sequenced reads were aligned both to the human genome (hg19) and mouse genome (mm10). Cellular barcodes (cell-assigned barcodes), arising from STAMPs, were called and background barcodes excluded. Background barcodes tagged no cells (in empty droplets), however, can be contaminated with ambient RNA incorporated into droplets, that is the pool of mRNA molecules that have been released in the cell suspension, likely from apoptotic cells or damaged cells that have leaked out RNA [65]. Usually, cell barcodes have associated significantly more transcript counts than the background barcodes (e.g. that sampled ambient RNA). To select cell-assigned barcodes, we thus filtered out all barcodes with less than approximately 10,000 UMI total counts, corresponding to the steep slope inflection point in the barcode-UMI plot, as shown in Figure 3G, recovering 4918 transcriptomes from cell-assigned barcodes.

The scatter plot in Figure 3G shows the cell doublet analysis of the human-mouse mixture experiment. Each dot represents a transcriptome (cellular barcode) with the number of associated human and mouse transcripts respectively on the x and y-axis. Blue dots indicate human-specific transcriptomes (average of 99% human transcripts), while green dots indicate transcriptomes that were mouse-specific (average of 99% mouse transcripts). Red dots represent transcriptomes with a significant proportion of transcripts associated with both the human genome and mouse genome and represent cell doublet (i.e. barcodes that occurred with two or more cells in the droplet). We obtained a cell doublet proportion of 14.9% working with the cell concentration of 500 cell/μL.

Overall, we processed 40,000 barcoded beads obtaining 4,918 cells, from a microfluidic run with 500 cell/μL, with a cell doublet rate of 14.9% from the human-mouse mixture experiment. Since 4918 represents the number of barcoded beads that have been exposed to a cell in the 40,000 barcoded bead sample, the barcoded bead yield was:

$$bead\ yield = \frac{4918}{40{,}000} \times 100 \approx 12\%$$

This value resulted greater than the value reported in the Drop-seq experimental procedure of 5%, with the reference cell concentration of 100 cell/μL (doublet rate of 1.9%). Because of the higher cell doublets rate, we decided that the 500 cell/μL concentration was not suitable to perform scRNA-seq experiment. Hence, we reduced the cell concentration to 200 cell/μL for all subsequent experiments, in order to reduce the cell doublet rate.

To estimate the cell yield, I had to do some approximations: 40,000 barcoded beads yielded 4,918 cells, therefore I approximated that 105,000 barcoded beads (the total amount I obtained) would yield ~12,800 cells, although this is an approximation since this value could fluctuate depending on the bioinformatic filtering (that is influenced by the quality of the data). During the microfluidic run, cells flowed at 66.6 μL/min at the concentration of 500 cell/μL, yielding ~500,000 cells in droplets in 15 min:

$$cell\ yield = \frac{12{,}800}{500{,}000} \times 100 \approx 2.5\%$$

This means that, on overage, the platform is able to capture the ~2.5% of cells. This highlights that one of the drawbacks of the Drop-seq platform is that this it is not suitable for processing small sample such as biopsy sample or rare primary cells. It is worth noting that a similar cell capture rate was reported also by Ziegenhain et al. (2018) [41], where they show that Drop-seq allows to capture from 2 to 4% of input cells, and the need of a relatively high number of starting cells compared to other droplet-based techniques for scRNA-seq.

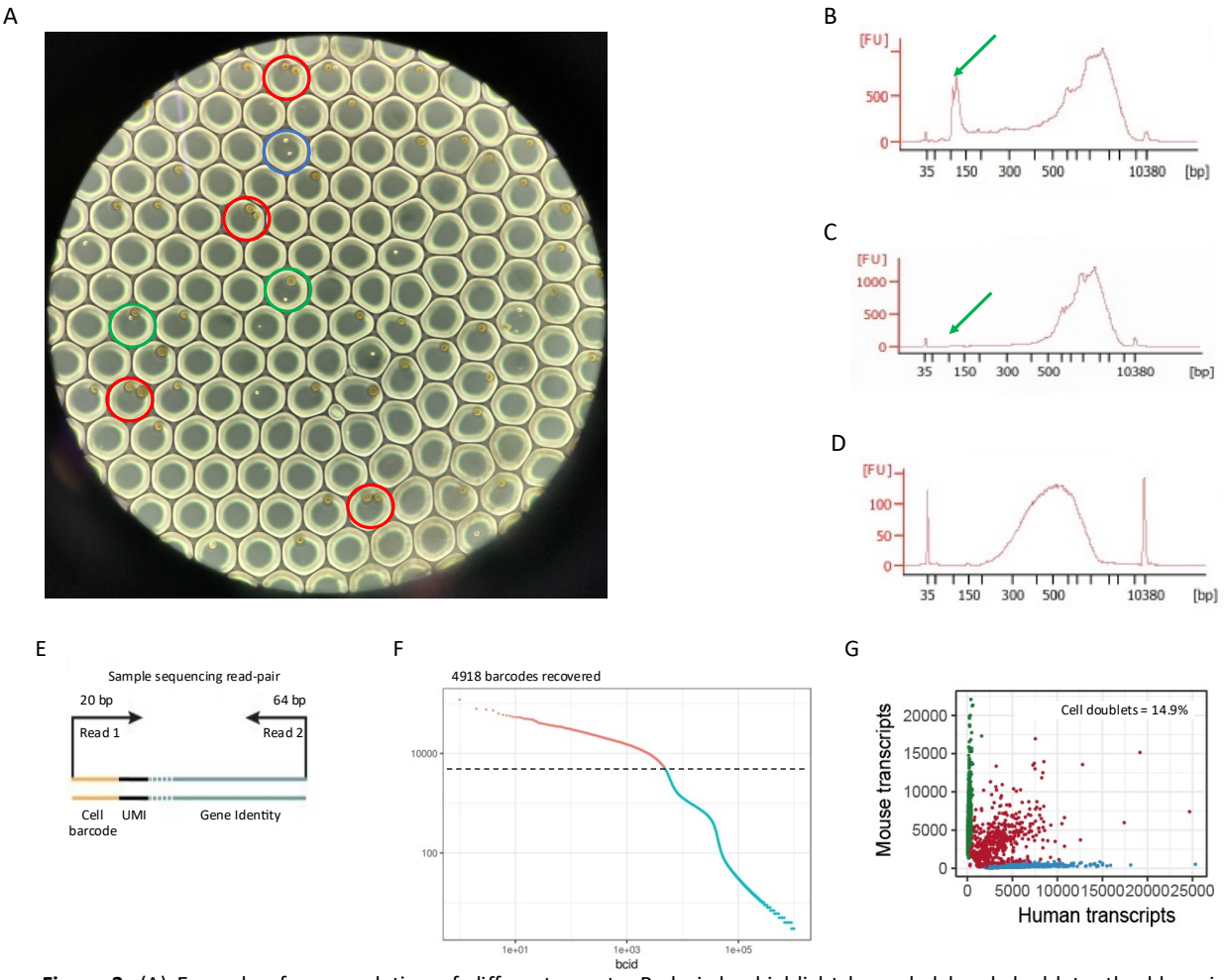


Figure 3. (A) Example of encapsulation of different events. Red circles highlight barcoded bead doublets; the blue circle highlights a cell double event; green circles highlight the occurrence of a cell together with a (B) Bionalyzer high sensitivity analysis of the cDNA library purified with Ampure XP beads in a ratio of 1.8X with the sample volume. This ratio do not exclude primer dimers, that appear as a strong peak at approximately 100 bp (highlighted by a green arrow). (C) cDNA library cleaned from primer dimers with an Ampure XP bead ratio set to 0.6X and; the cDNA distribution appear smooth with a bp average at approximately 1300 bp. The green arrow shows the absence of the primer dimer peak. (D) NGS library distribution analyzed by bioanalyzer high sensitivity chip, with a bp average around 500 bp. No primer dimer peak is detectable after agarose 2% gel purification. (E) schematic representation of the paired-end sequencing for the human-mouse mixture experiment. Read1 covers unique barcode (12 bp) and UMI (8 bp), and together account for 20 bp. Read2 covers the transcript sequence paired to read1; up to 64 bases were called for human-mouse mixture experiments. (F) Barcode-UMI plot. The steep inflection point (dashed line) is the threshold to exclude cell barcodes from background barcodes. We recovered up to 4918 barcodes after filtering. (G) Cell doublet rate for the human-mouse mixed experiment resulted 14.9%. Barcodes with associated $\geq$ 99% of transcripts from the same species (i.e. align on the same genome, either human or mouse), were considered to come from the occurrence of one single cell with a barcoded bead. Below this percentage, barcodes were considered non-species specific and classified as doublet.

3.4 – Improvement of the microfluidic platform

The Drop-seq platform is a powerful tool for automatic, high-throughput scRNA-seq. However, the main drawback is the need of highly diluted cell and barcoded bead suspension to avoid high rate of multiple encapsulation events (i.e. cell doublets, barcoded bead doublets). This results in the great majority of droplets being empty (relatively low throughput), since cell and barcoded bead encapsulation at the droplet generation junction is stochastic and follows the Poisson distribution

3.4.1 - Implementation of a spiral channel microfluidic device for barcoded bead ordering

Recently, Moon et al. reported a modified version of the Drop-seq microfluidic device to overcame random distribution of the barcoded beads in droplets by implementing a spiral channel at the barcoded bead port [54]. The spiral channel, through inertial effect, orders highly concentrated barcoded beads to form a train (Figure 4A). Thus, ordered barcoded beads are equally spaced when entering the flow-focusing junction, resulting in deterministic encapsulation events. This modification of the original Macoscko microfluidic device improves the capture rate of cells, as it increases the fraction of cells encountering a single bead. Indeed, the use of the spiral channel allows to increase concentration of beads as these will be orderly spaced thus increasing the number of droplets containing a single bead. Cells are still randomly encapsulated, whereas the barcoded bead will be deterministically encapsulated. Therefore, I implemented the Moon version of the Drop-seq microfluidic device to improve the microfluidic setup in the lab. Moon et al. optimized the aqueous and oil phase flow rate values to generate droplet in a stable dripping mode at the flow-focusing geometry. I set the cell suspension and barcoded bead flow rate to 3.2 mL/hr and the oil flow rate to 12 mL/hr, by selecting from a range of flow rate values provided in the paper. As shown by Moon et al., the flow rate I selected were suitable to generate a stable dripping mode for droplet generation.

I then selected the cell and barcoded bead concentration. I set the cell loading concentration to 250 cell/μL, as reported in the paper. Moon et al. analyzed a range of barcoded bead concentration from 100 and 1250 bead/μL, showing that the microfluidic device yields very low barcoded bead doublets (<0.5%) even at the highest barcoded bead concentration with throughput and cell yield increasing over barcoded bead concentration. Moon et al. performed the human-mouse mixture experiments with 1000 bead/μL, however, I decided to not work with that concentration, because, from time to time, I experienced that a high barcoded bead concentration increases the chance to form clogs not only in the microfluidic device channels, but also in the syringe needle, PTFE tubes and inlet port. I set the barcoded bead concentration to 250 cell/μL, that is reported in the paper to yield 5% and ~500 cell/min of respectively cell yield and throughput. With this setting I operated droplet generation using the Moon microfluidic device (flow rates and concentration for this experiment are summarized in Table 2). Barcoded bead stirring was the same as the Macoscko microfluidic device.

3.4.2 – Performance test with deterministic barcoded bead encapsulation

In order to estimate the performance of the Moon microfluidic device in terms of cell and barcoded bead yield, I generated droplets using the Moon microfluidics device with the setup (syringe pumps, microscope, etc…) implemented for the Drop-seq platform. For this test I resuspended cells in a buffer (as described in Moon et al.) with no sarkosyl, in order not to lyse cells and thus to count intact cells in droplet. I collected droplets in a 60 mm plastic dish, rather than in a tube, to distribute them in a monolayer onto a larger surface. Then, I captured random fields in order to count bead and cell occupancy in the droplets (example in Figure 4B, full data in Appendix C – Supp Figure C1) under an optical microscope. I counted 27 fields, with a total of:

- 88 droplets with one barcoded bead
- 35 droplets with one cell

8 droplets contained one barcoded bead and one cell together. In my hands, the microfluidic device yielded no droplets with more than one barcoded bead, suggesting an efficient barcoded bead ordering in the spiral channel. Surprisingly, in this test, no cell doublet occurred as well. I compared my results with results in the paper, by calculating the cell yield:

$$cell\ yield = \frac{8}{35} \times 100 = 22\%$$

And the barcoded bead yield:

$$bead\ yield = \frac{8}{88} \times 100 = 9\%$$

The cell yield I obtained was 22%, resulting much higher than the reported value in the Moon et al. publication of ~5%, in the same cell and barcoded bead conditions.
Barcoded bead yield is a suitable parameter to work with, since the barcoded beads get recovered after droplet breakage, irrespective of cells that lyse in the droplets. The resulted barcoded bead yield I obtained, was 9%; this result was comparable with the value I found with the Macoscko microfluidic device with 500 cell/µL (12%). Moreover, 9% of barcoded bead yield was higher than the barcoded bead yield reported in Macoscko et al. with the reference settings and cell concentration reported in the paper (~5% barcoded bead yield, with 100 cell/µL), with the advantage that the Moon microfluidic device and settings significantly reduced doublet events, such as bead doublets, that contributes to errors in cell barcoding.

3.4.3 – Implementation of a new passive filter for barcoded beads

Over time, I experienced that one of the main challenges during a Drop-seq microfluidic run is to uniformly input barcoded beads in the channel and avoid clog formation. Debris flowing in the barcoded bead port can obstruct the channel causing barcoded beads to aggregate and clog (Figure 4C). For the cell suspension I experienced much less clog events; this is particularly due to the smaller size of the cells (15 μm average) compared to the barcoded beads (30 μm average). In addition, the passive filter at the cell suspension inlet port has been designed with closer pillars than the barcoded bead inlet port filter, preventing more efficiently debris from passing the filter barrier; indeed, this is due to the smaller size of cells. Nevertheless, in some cases, I noticed that primary cells or starved cells that I processed with the Drop-seq platform, showed high sticky features that led cells to stuck in the filter and blocking in turn the other cells flowing through the inlet port eventually decreasing the number of cells encapsulated over time at the droplet generation junction. All of this events often force to switch microfluidics device for continuing the generation of droplet, and indeed I experienced that, in this issue, the barcoded beads filter represents the component with the higher chance of clog formation.

However, the Macoscko microfluidic device consists of an exhaustive passive filter at the barcoded bead inlet port compared to the Moon passive filter. The latter consists of only three pillars and the port ends with an angle of approximately 90° to the spiral channel for (Figure 4D); I experienced this design to be very inefficient in filtering, and from test to test I was forced to switch microfluidic device almost every time the barcoded bead port was clogged, with no few possibilities of recovering the microfluidic device. In order to solve this problem, I decided to implement the Macoscko passive filter of the barcoded bead inlet port to the Moon microfluidic device. The wafer master mold was fabricated by Gianmarco Nocera, PhD using standard photolithography protocol. The new passive filter implemented in the Moon microfluidic device is shown in Figure 4E. This new device resulted efficient as much as the Macoscko passive filter in preventing channels from clogging.

	Cells	Barcoded beads	Oil
Flow rate	3.2 mL/hr	3.2 mL/hr	12 mL/hr
Concentration	250 cell/μL	250 bead/μL	-

Table 2. Concentrations and flow rates set for the Moon microfluidic device test.

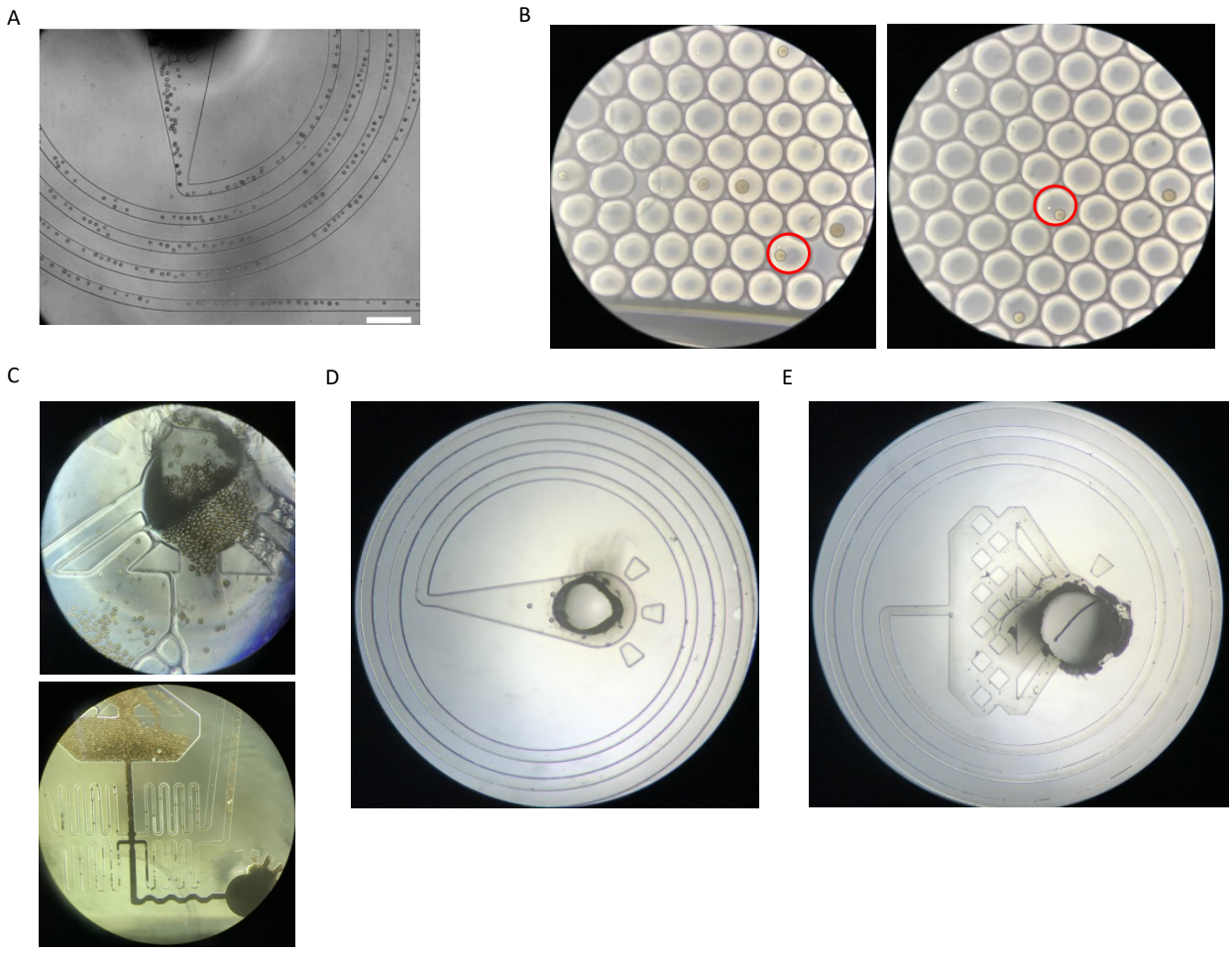


Figure 4. (A) Example of inertial ordering of barcoded beads in the spiral channel; adapted from Moon et al. (B) Two representative fields for counting cell and barcoded bead occupancy in droplets (20x magnification). Red circles highlight the occurrence of one cell with one barcoded bead. (C) Example of barcoded bead clog in the Macoscko microfluidic device. Upper image 10x magnification, lower image 5x magnification. (D) Passive filter in the Moon microfluidic device (10x magnification). (E) Passive filter we implemented to decrease the chance of bead clog formation.

CHAPTER 4 - Single-cell transcriptome profiling of breast cancer cell lines (CCLs) for automated cancer diagnosis

By means of the Drop-seq technology that I described in Chapter 2 and implemented in Chapter 3, I performed the next generation sequencing of single cells from a panel of 32 breast cancer cell lines (CCLs). Here, I describe the sequencing of the 32 cell lines, and the subsequent data processing to generate a comprehensive single-cell transcriptomics atlas enabling automated cancer diagnosis. Data processing and computational analysis was performed in collaboration with Gennaro Gambardella, PhD.

4.1 – Single cell RNA-seq of breast cancer cell lines.

4.1.1 – Selection of comprehensive panel of breast cancer cell lines.

A panel of 32 cell lines was selected for single-cell sequencing, of which 31 breast cancer cell lines (CCLs) and one additional non-tumorigenic CCL from fibrocystic breast disease. To fully explore the whole transcriptome landscape of breast cancer at the single-cell resolution, the selected panel covered all breast cancer subtypes as detailed below:

- 9 luminal A
- 2 luminal B
- 5 HER2+
- 8 triple negative A
- 7 triple negative B

I collected data from the literature and public databases for each CCL, including the expression status of clinically relevant biomarkers, the growth condition, and the derivation site (Appendix D - Supp. Table D1). Most of the CCLs were derived from pleural effusion and other metastatic sites; for example, MDA-MB-361 cell line was established from brain metastatic site.

4.1.2 – Single-cell RNA-seq of CCLs

By means of the Drop-seq microfluidic platform, I performed single-cell RNA-seq of 31 CCLs and 1 basal-like normal breast epithelium cell line. I performed all experiments with 200 cell/µL and 120 bead/µL, by applying the experimental procedure that I previously optimized and described in Chapter 3; the sequencing of single-cell NGS libraries was performed with the NovaSeq 6000 Illumina sequencer. For each CCL, I prepared single-cell NGS libraries to sequence 1,000 cells, and I loaded libraries in the S1 flow cell with the 2×50 bp reagent kit, which yields as

output 1.3-1.6 Billion of single-reads CPF (cluster passing filters). The Illumina indexing strategy during NGS library generation allows to pool together more samples to parallelize the sequencing of multiple libraries (multiplexing). I pooled together up to eight single-cell NGS libraries with a unique index to parallelize the sequencing of CCLs and enable demultiplexing. I diluted each library of the pool to the library with the lowest concentration (in nM), to assign the same amount of reads to each CCLs. Then, I sequenced the pool of libraries in paired-end mode, where the read1 covered 24 bp to include 20 bp for barcode and UMI (the surplus of 4 bp was due to technical issues of the sequencer), 84 bp to the read2 for base calling of the paired transcript, and 8 bp for the read1 index.

4.2 – Sequencing reads alignment and gene expression quantification

Following sequencing of the cDNA single-cell libraries, we processed raw data using the Drop-seq tools package version 1.13 and we followed the pipeline described in the Drop-seq Core Computational Protocol (http://mccarrolllab.org/dropseq). We filtered raw sequence data to remove all read pairs with at least one base in their barcode or UMI with a quality score less than 10. We trimmed read2 at the 5' end to remove any adapter sequence, and at the 3' end to remove polyA tails. Then, we aligned reads using the STAR bioinformatics pipeline [66] on human genome (hg38 primary assembly, version 28) downloaded from the GENCODE database [67]. Following reads alignment, the UMI tool [68] was applied to perform UMI de-duplication and quantify the number of gene transcripts in each cell. In order to identify the number of sequenced cells; we used a simple (knee-like) filtering rule as implemented by the CellRanger 2.2 software. In this process, we retained only cells with: (i) at least 2,500 UMIs; (ii) morre than 1,000 captured genes, and (iii) with less than 50% of reads aligned to mitochondrial genes used as marker of dead cells (apoptosis). We discarded putative cell doublets by identifying outliers in the count depth distribution by applying the Tukey's method based on lower and upper quartiles with k equal to 3. Following pre-processing, we eventually retained a total of 35,276 cells, with an average of 1,069 cells per cell line and 3,248 genes captured per cell, as reported in Figure 4.1.

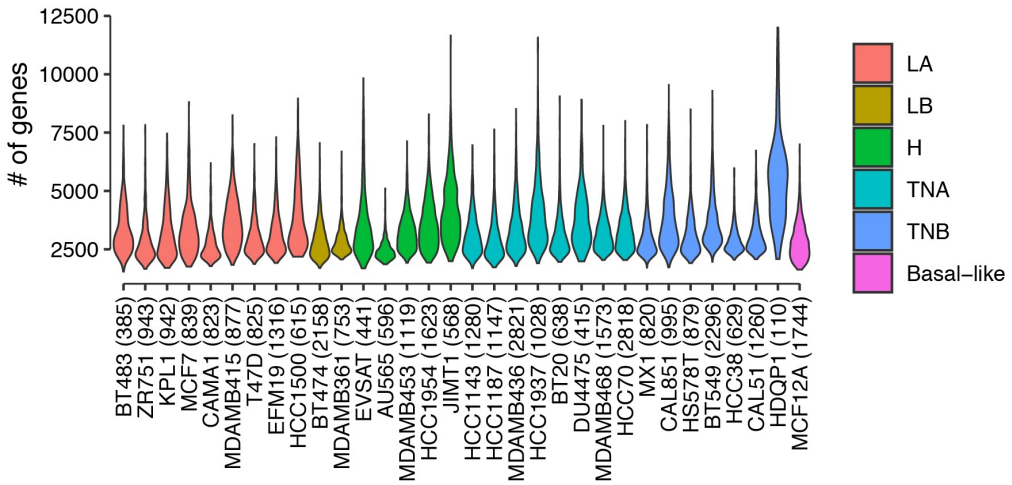


Figure 4.1 – **Violin plot of the captured genes per cell line**. Distribution of the number of captured genes per cell, across cell lines (x-axis); between parenthesis is specified the number of captured genes in the cell line.

4.3 - Breast cancer cell line single-cell atlas construction

Single-cells expression profiles were normalized by means of the GF-ICF (Gene Frequency – Inverse Cell Frequency) normalization, a method recently described in the literature and available as an R package (*gficf*) [69] (https://github.com/dibbelab/gficf). GF-ICF is based on a data transformation model called term frequency-inverse document frequency (TF-IDF) that has been extensively used in the field of text mining. Briefly, given a set of N cells, let f_{ij} be the number of transcripts of the gene i in the cell j, the gene frequency GF_{ij} of gene i in the cell j, can be defined as: $GF_{ij} = f_{ij} / \sum_{k=1}^{N} f_{kj}$ and represents its number of transcripts divided by the total number of transcripts of the cell. The Inverse Cell Frequency of gene i can be instead defined as $ICF_i = \log(N + 1/n_i + 1)$ where n_i denotes the number cells that contain gene i among the N sequenced cells. The GF-ICF score for gene i in cell j is finally defined as $GF_{ij} \times ICF_i$. GF-ICF values of each cell are then re-scaled to have Euclidean norm equal to one (L2 normalization) to account for cell depth biases. We applied GF-ICF transformation on CPM (count per million) after *EdgeR* normalization [70] and discarded genes expressed in less than 5% of the total number of sequenced cells. Finally, we summarized each cell with its first 10 Principal Components (PCs) and projected to a two-dimensional space with the UMAP package (*uwot* package, R statistical environment 3.6) [71]. We chose the number of principal components equal to 10 by selecting the elbow plot of the first 50 PCs.

The breast cancer (BC) single-cell atlas (http://bcatlas.tigem.it) encompassing 32 cell lines is shown in Figure 4.2A and was obtained by combining data across cell lines. In the atlas, cell lines derived from the same cancer subtypes tend to cluster together, while being separated from the other subtypes: luminal BC cell lines form a big "island" with multiple "peninsulas" with intermixing of cells from distinct cell lines; on the contrary, triple-negative breast cancer (TNBC) cell lines give rise to an "archipelago", where cells tend to separate into distinct islands according to the cell line of origin, thus suggesting that TNBC cell lines represent instances of distinct

diseases. Single-cell expression of clinically relevant biomarkers (Figure 4.2B,C) including oestrogen receptor 1 (ESR1), progesterone receptor (PGR), Erb-B2 Receptor Tyrosine Kinase 2 (ERBB2 a.k.a. HER2) and the epithelial growth factor receptor (EGFR) across the different cell lines are in agreement with their reported status.

4.4 – Biomarker analysis of Breast Cancer Cell lines in the atlas.

To gain further insights into each cancer cell line, we analysed the expression of 48 literature-based biomarkers of clinical relevance, as reported in Figure 4.2D. Luminal cell lines highly express luminal epithelium genes, but neither basal epithelial nor stromal markers. We detected higher expression of FOXA1 and GATA3 in luminal cell lines compared to TNBC, which have been found to be involved in ESR1-induced target genes transcriptional regulation [8], [9] as well as with luminal specific phenotype identity [72]; FOXA1 expression has been shown to control plasticity between basal and luminal breast cancer cells, not only by inducing luminal genes but also by repressing the basal phenotype [73]. In addition, both FOXA1 and GATA3 have been identified as favorable prognostic factors and associated with good survival in breast cancer patients [74], [75].

Unlike luminal cell lines, TNBC cell lines (11 out of 15) show a basal-like phenotype with the expression of at least one of cytokeratin 5, 14, or 17, with triple-negative subtype B (TNB) cell lines also expressing vimentin (VIM) and Collagen Type VI Alpha Chains (COL6A1, COL6A2, COL6A3), genes typically found in fibroblastic cells [76]. In agreement with the knwon TNBC features [2], we detected high expression of genes associated with tumor invasiveness, in particular for TNB cell lines, expressed genes include ZEB1, TWIST, SNAI2 (SLUG) transcription factors. SNAI2 is a well-known gene associated with malignant biological properties of cancer cells [77] and together with ZEB1 and TWIST plays a critical role in malignant transformation and tumor progression; ZEB1 and TWIST have been shown to downregulate CDH1 (E-cadherin), which play a migration-suppressive role, promoting the EMT process[78]. Moreover, we detected in TNBC cell lines, and in particular in the TNB subtype, expression of CDH2 (N-cadherin), a stromal marker, which, irrespective of CDH1, endows tumor cells with enhanced migratory and invasive capacity [79].

Two out of five HER2 overexpressing (HER2+) cell lines (JIMT1 and HCC1954) in the atlas are in the triple-negative “archipelago” and express cytokeratin 5 (KRT5), a basal marker in breast cancer, linked to poor prognosis and unfavourable overall survival [80]. HER2+ cell lines, including JIMT1, expressing basal markers (i.e. KRT5) were classified in the basal-HER2+ subgroup, characterized by resistance to trastuzumab [81]. JIMT1 and HCC1954 cell lines have been reported to be resistant to trastuzumab and anti-HER2 treatments, and in particular, JIMT1 has been studied as a cancer cell line model for anti-HER2 treatment resistance [82], [17]. Indeed, cells overexpressing KRT5 are more invasive, sphere-forming, and quiescent with increased resistance to endocrine and chemotherapy and trastuzumab resistance [83], [84]. Interestingly, excluding the TNBC cell lines, JIMT1 cell line is the only one that shows the expression of CDH2 stromal marker and EMT regulator genes, like SNAI2 and TWIST1, with in addition CD44, a marker associated with stemness features.

Finally, the non-tumorigenic MCF12A cell line lacks expression of ESR1, PGR, and HER2 and displays a basal-like phenotype. Indeed, I detected basal markers in MCF12A as reported in the literature, such as expression of VIM, TP63, and no expression of ACTA2 [85], as well as

KRT8/18 positivity and KRT19 negativity according to the marker analysis reported in ATCC in this cell line.

Overall, these results show that single-cell transcriptomics can be successfully used to capture the overall expression of clinically relevant markers.

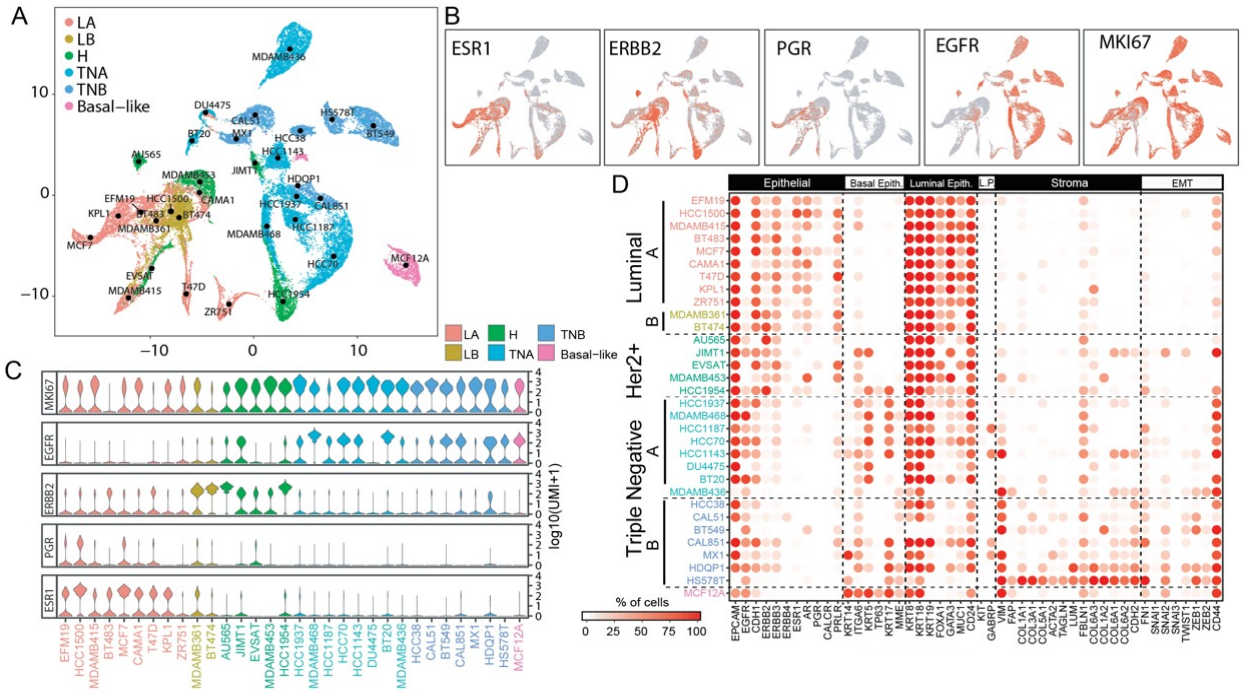


Figure 4.2 - The Breast Cancer Single Cell Atlas. (A) UMAP representation of single-cell transcriptomics of 32 cell lines for a total of 35,276 cells color-coded according to cancer subtype (LA=Luminal A, LB=Luminal B, H=Her2 positive, TNA = Triple Negative A, TNB = Triple Negative B). (B) Expression levels of the indicated biomarker genes in individual cells in the atlas, with red indicating expression, together with their (C) distribution within the cell lines, shown as a violin plot. (D) Dotplot of biomarker genes along the columns for each of the 32 sequenced cell lines along the rows. Biomarker genes are grouped by type (Basal Epith. = Basal Epithelial, Luminal Epith. = Luminal Epithelial, L.P. = Luminal Progenitor, EMT = Epithelial to Mesenchymal Transition)

4.5 – Single-cell RNA sequencing captures the expression of clinically relevant signatures across CCLs

By clustering the 35,276 single-cells in the atlas, we identified 22 clusters, as shown in Figure 4.3A. Interestingly, within the luminal island, cells did not cluster according to their cell line of origin, indeed four out of the five luminal clusters contain cells from distinct cell lines (Figure 4.3B and Appendix C – Supp. Figure C2). On the contrary, triple-negative cell lines clustered according to their cell line of origin, with each cluster containing mostly cells from the same cell line.

We identified genes specifically expressed among cells in the same cluster for a total of 22 biomarkers, one for each cluster (Figure 4.3C,D). Interestingly, neither ESR1 nor HER2 were part of this set. Literature mining confirmed the significance of some of these markers: clusters in the luminal island (Figure 4.3C) were associated to genes involved in cancer progression (BCAS3 [86] cluster 2), dissemination (SCGB2A2 [87], [88] cluster 6), proliferation (DRAIC, cluster 1), migration and invasion (CLCA2, cluster 8 and PIP, cluster 18). Interestingly, whereas DRAIC is correlated with poorer survival of luminal BC patients [89], both CLCA2 and PIP are significantly associated with a favourable prognosis.CLCA2 was shown to be downregulated in several primary breast tumors and breast CCLs, and loss of CLCA2 was associated with tumorigenicity and invasion potential ([90], [91]**)** while overexpression showed decreased proliferative, migrating and invasive features [92]. SCGB2A2 is a member of the uteroglobin protein family and has been identified to be breast specific and a candidate breast cancer associated marker [93], [94]. Several studies show that SCGB2A2 is mainly expressed in luminal ESR1 positive and HER2 positive subtypes compared to TNBC subtype [95], [96]. Correlation of SCGB2A2 with oestrogen and progesterone receptor expression, histological and nuclear grade and cell proliferation in breast cancer patient specimens indicated that SCGB2A2 expression is associated with a less aggressive tumor phenotype [88], [97].

To examine the clinical relevance of these 22 biomarkers, we analysed their expression across 937 breast cancer patients from the TGCA collection encompassing all four BC types. Out of the 22 biomarkers, two (MAGEA4 and XAGE2) could not be mapped to the TGCA dataset. As shown in Figure 4.3D, there is a marked difference in the expression of the 20 cluster-derived biomarkers across Luminal A, Luminal B, HER2 positive and TNBC patients. Moreover, it is possible to distinguish subtypes within each category, which may lead to novel diagnostic/prognostic biomarkers (Figure 4.3D). For example, one subset of triple-negative patients strongly expresses the protease kallikrein-10 (KLK10), which has been associated with poor prognosis, poor response to tamoxifen treatment and identified as potential target to reverse trastuzumab resistance [98], [99]. Whereas a second subset is characterised by actin gamma 2 expression (ACTG2), involved in different cellular processes including cell motility [100], whose overexpression has been linked in BC to cell proliferation and platinum-based chemotherapy sensitivity, including paclitaxel [101], [102], [103], [104].

Finally, we compared the performance of the 20 biomarker genes in classifying BC subtypes from bulk RNA-seq data (Appendix A - Methods) against the PAM50 gene signature (50 genes) [89] used in clinics to identify breast cancer subtypes (Figure 1I). The performances were overall comparable, with the obvious exceptions of HER2-overexpressing cancers. Indeed, when adding *ERBB2* to the list of 20 cluster-based biomarkers, classification of this subtypes markedly improved (Figure 4.3E).

Altogether, these analyses confirm that the single cell BC cell line atlas allows identifying novel clinically relevant gene signatures useful for patient stratification and tumor type classification.

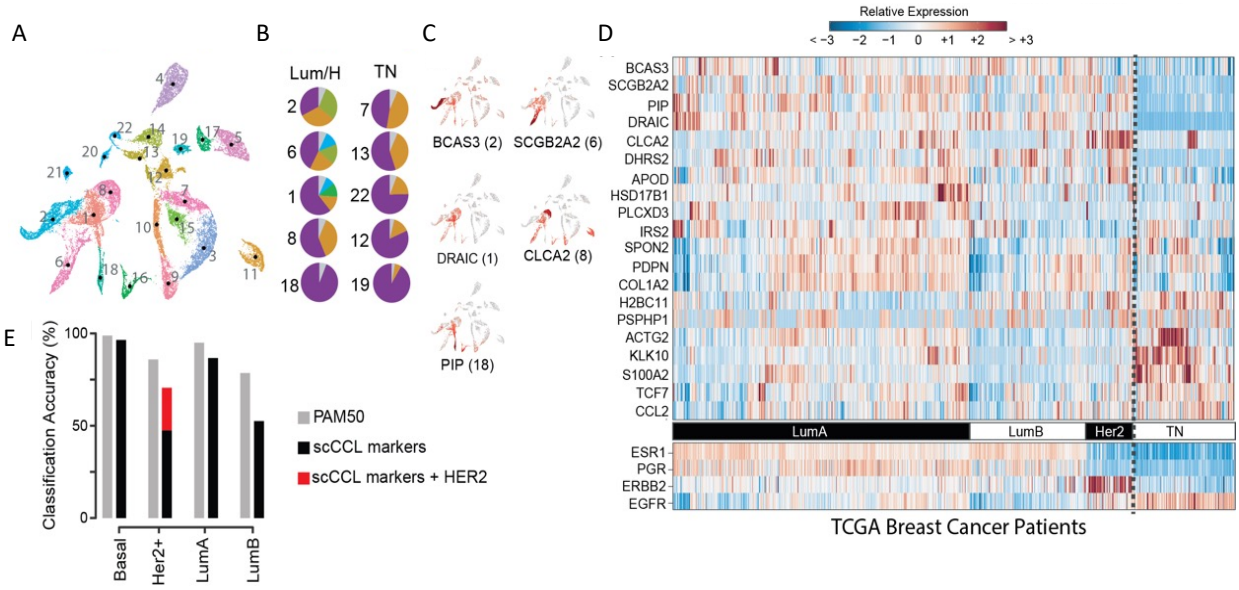


Figure 4.3 – **Clinically relevant signatures across CCLs**. (A) Graphical representation of 35,276 cells color-coded according to their cluster of origin. Clusters are numbered from 1 to 22. (B) For the indicated cluster, the corresponding pie-chart represents the cluster composition in terms of cell lines. Cell lines in the same pie-chart are distinguished by colour. Only the top 10 most heterogenous clusters are shown. Cluster 2 is the most heterogeneous while cluster 19 is the most homogeneous. Cell line with less than 1% of cells in a cluster have been merged and are represented with the grey slice. (C) Expression levels in the atlas of the five luminal biomarkers identified as the most differentially expressed in each of the five luminal clusters (1, 2, 6, 8 and 18). (D) Expression of 20 out of 22 atlas-derived biomarkers in the biopsies of 937 breast cancer patient from TCGA. (E) Accuracy in classifying tumour subtype for 937 patients from TCGA by using either PAM50 or the 22 atlas derived biomarker genes (scCCL) alone or augmented with HER2 gene (scCCL + HER2).

4.6 – The breast cancer single-cell atlas for automated cancer diagnosis

The BC atlas can be used as a reference against which to compare single cell transcriptomics data from a patient's tissue biopsy and to perform cancer subtype classification and assessment of tumour heterogeneity. To this end, I applied an algorithm developed by Gennaro Gambardella, PhD, able to map single-cell transcriptional profiles from a patient onto the BC atlas and to assign a specific cell line to each of the patient's cells. We first tested the ability of the algorithm in correctly classifying the very cells in the atlas starting from their single-cell transcriptional profiles and correctly classified 92% of the cells (Appendix C – Supp. Figure C3). We then turned to single-cell transcriptional profiles obtained from five triple-negative breast cancer patients [105]. As shown in Figure 4.4A, most, but not all the patients' cells mapped to the triple-negative "archipelago", except for the TNBC5 sample, for which most cells mapped to the luminal island. As the algorithm assigns a specific cell line to each tumour cell, it is also possible to look at the cell line composition of each patient, as reported in Figure 4.4B. For the samples TNBC1, TNBC2, TNBC4 and TNBC5 most single cell profiles (79%, 91%, 79% and 75% respectively) were

assigned to two cell lines (MX1 and HCC1187 for TNBC1 and TNBC2; MX1 and DU4475 for TNBC4, and ZR751 and T47D for TNBC5). These results demonstrates that heterogeneity is present in all the samples, as no patient's biopsy mapped to a single cell line. Moreover, information on the drug sensitivity of the individual cell lines composing the tumour may prove useful in guiding therapeutic choices.

We next tested the algorithm on spatial transcriptomics dataset obtained from the tissue biopsy of two patients, one diagnosed with $ESR1^{+}/ERBB2^{+}$ lobular estrogen positive carcinoma and the other with $ESR1^{+}/ERBB2^{+}$ ductal carcinoma (Figure 4.4C, and Appendix C – Supp. Figure C4) [106]. The dataset consists of 3,808 transcriptional profiles for patient 1 and 3,615 profiles for patient 2 (Appendix C – Supp. Figure C4), each obtained from a different tissue "tile" of size 100 μm x 100 μm x 100μm. The algorithm projected each of the spatial tiles onto the BC atlas and assigned a cell line to each tile. The algorithm projected each of the spatial tiles onto the BC atlas and assigned a cell line to each tile. We coloured the tiles according to the cell line and the BC subtype of the cell line to yield an automatic cancer subtype classification of tiles. Most of the tiles for both patients were assigned to just two cell lines and correctly classified as luminal (A or B); the remaining 13% of the tiles for patient 1 and 20% for patient 2 were instead classified either as HER2-overexpressing or Triple Negative, which could be an important information to guide therapeutic choice and to predict the occurrence of drug resistance.

As bulk gene expression profiles are more clinically relevant than single-cell gene expression profiles, we next trained a deconvolution algorithm [107] (Appendix C – Supp. Figure C5) by leveraging our single-cell atlas to predict the cell line composition from the bulk gene expression profile of a tumour sample. To test the effectiveness of this approach, we collected from the TGCA database, 937 gene expression profiles from breast cancer patients whose BC subtypes were annotated. The deconvolution algorithm assigned to each of the 937 patients the predicted cell line composition, which we then used to cluster patients, as shown in Figure 4.4D,E. Reassuringly, patients diagnosed with a specific breast cancer subtype tend to have a tumour cell line composition consisting of cell lines of the same subtype. We quantified this observation in Figure 4.4F and observed some interesting exceptions: JIMT-1 is an HER2+ cell line with an amplified ERBB2 locus, but no HER2+ patient was mapped to this cell line. Interestingly, JIMT-1 are resistant to anti-HER2 treatments [108]; another example is the cancer cell line HS578T which is reported to be triple-negative, however the majority of patients who map to this cell line are luminal; surprisingly, this cell line has been reported to be sensitive to fulvestrant [28], [29], an anti-ESR1 drug used for luminal patients.

These results show that this single cell atlas of cancer cell can be used to automatically assign cell line composition and cancer subtypes both from single-cell expression profiles and bulk gene expression profile.

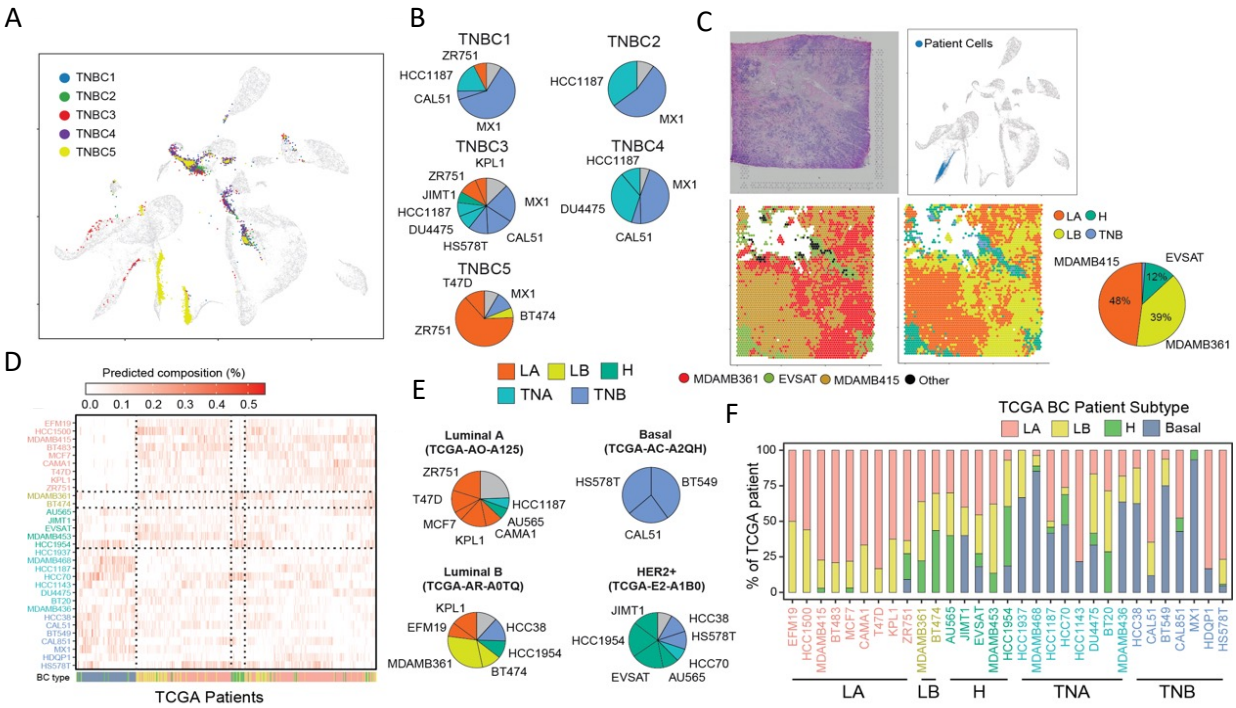


Figure 4.4 – **Clinically relevant signatures across CCLs**. (A) Cancer cells from triple negative breast cancer (TNBC) biopsies of 5 patients sequenced with 10X genomics technology are embedded in the BC atlas to predict their tumour type using K-nn algorithm. (B) Pie chart that show how cells of each TNBC patient are classified after their repositioning into the atlas. Cell lines represented with a percentage less than 5% are merged and represented with a grey slice of the pie-chart. (C) Top-left: Tissue-slide of an estrogen positive breast tumour biopsy sequenced using 10x visium spatial transcriptomics. Top-right: Cancer cells sequenced with 10X visium technology are embedded in the BC atlas to predict which cell-line they are similar. Bottom-left: Classification of each pseudo cell to show predicted cell-line in the spatial context. Bottom-right: Classification of each pseudo cell to show predicted tumour type in the spatial context. (D) *Bulk* RNA-seq of 937 TCGA patients are deconvolved into cell lines and hierarchically clustered. (E) Example of predicted cell-line composition of 4 out of 937 TCGA patients. (F) The BC subtype covered by each cell line TCGA patient deconvolution.

CHAPTER 5 - Intrapopulation gene expression heterogeneity of cancer cell lines

Single-cell transcriptomics has the potential to investigate cell-to-cell transcriptional heterogeneity. In this Chapter, I show the experiments I performed to assess the gene expression intrapopulation heterogeneity, using three representative breast cancer cell lines (MDA-MB-361, AU565, and HCC38) to investigate the heterogeneity in HER2 expression. By means of flow cytometry, I checked that the HER2 state is not only heterogeneous at the mRNA level, as assessed by scRNA-seq, but also at the protein level. I focused on the MDA-MB-361 cell line, with approximately 70% of cells expressing HER2, to demonstrate that sorted HER2+ and HER2- homogenous subpopulations spontaneously give rise to heterogeneous populations, thus confirming that heterogeneity that I observed is not caused by genomic heterogeneity. Finally, in collaboration with Gennaro Gambardella, PhD I investigated the differences in the expression programs between these two subpopulations and we found upregulated pathways indicative of epithelial-to-mesenchymal transition (EMT) in the HER2+ subpopulation, while cell cycle related pathways were upregulated in the HER2- subpopulation, suggesting the cell-cycle status could have a role in causing the observed heterogeneity.

5.1 – scRNA-seq shows intrapopulation heterogeneity within CCLs

Single-cell RNA-seq is a powerful method to unravel intrinsic heterogeneity in gene expression profiles within cells in a population. Clinically relevant receptors are heterogeneously expressed across cells belonging to the same cell line, as assessed by computing the percentage of cells in a cell line expressing the receptor in Figure 5.1A. Overall, within cell lines, I found variability in the percentage of cells expressing clinically relevant biomarkers. Consider the seven Luminal B and HER2+ cell lines present in the BC atlas, which by definition overexpress HER2: whereas more than 90% of cells in AU565, BT574, and HCC1954 cell lines express *ERBB2*, in the remaining four cell lines *ERBB2* expression ranged from 31% of EVSAT cells to 46% of JIMT1 cells and up to 64% of MDA-MB-361 cells. This happens despite both JIMT1 and MDA-MB-361 harbor a copy number gain of the locus containing the *ERBB2* gene [109]. The HER2 protein, encoded by *ERBB2* gene, is a receptor that can be pharmacologically targeted by small molecule inhibitors, like Afatinib, or by antibodies such as trastuzumab.

Therefore, I decided to assess whether the observed gene expression heterogeneity reflects the same heterogeneity at the protein level. To this purpose, I employed flow cytometry analysis to measure the HER2 protein in three representative cell lines: AU565 (high HER2 expression, 95%), MDA-MB-361 (heterogeneous HER2 expression, 64%), and HCC38 cell lines (low HER2 expression, 3%). Flow cytometry provides single-cell analysis of a statistically significant number of cells (events) employing optical filters to measure the physical features of cells and fluorescence emission. To measure the rate of HER2 (HER2+) cells in the selected cell lines, I stained cells with a dye-conjugated antibody that specifically binds HER2 protein (mouse anti-human HER2 BB700 conjugated antibody from BD biosciences). For each stained sample, I used an unstained sample to define the boundary between HER2+ and HER2- cells. The fluorescence of both stained and unstained samples of each cell line was measured with the Accuri C6 instrument (640LP BD standard filter in FL3), and then I analyzed data with the BD Accuri C6 software. The results in

Figure 5.1B show that single-cell transcriptional data agree with the cytometric analysis: the AU565 cell line resulted to be homogeneous for HER2 expression with 93% of cells stained with the for HER2 antibody, while the MDA-MB-361 population showed 69% of HER2+ cells, and a very low rate of HER2+ cells was present in the HCC38 population (6%).

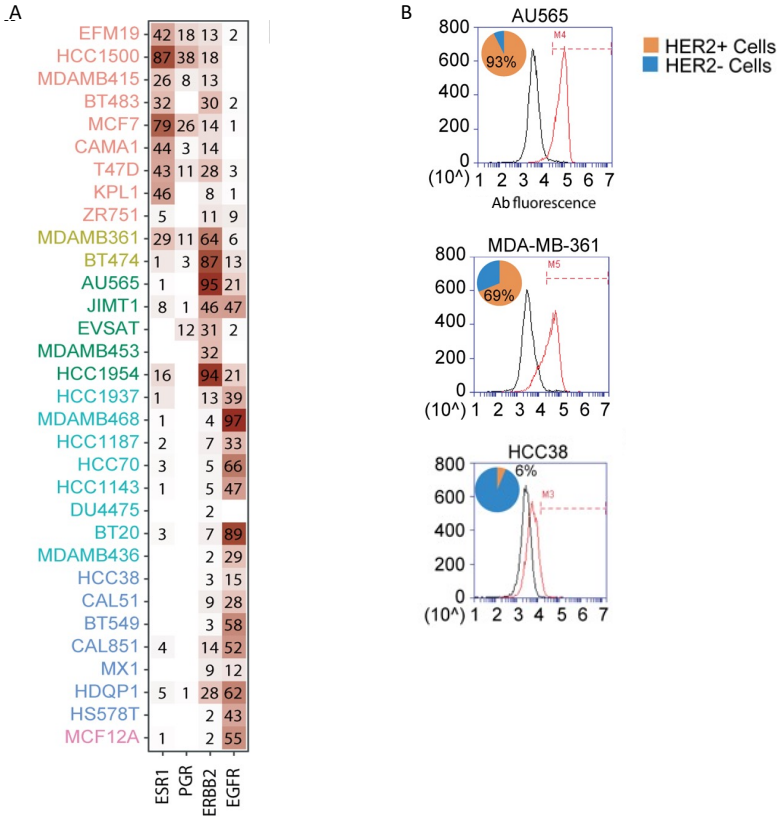


Figure 5.1 - (A) Percentage of cells expressing the indicated genes in each of the sequenced cell lines. (B) FACS analysis of HCC38, MDA-MB-361 and AU565 cell lines stained with an antibody against HER2 (BD BB700 mouse anti-human HER2).

5.2 – HER2 expression state is dynamically regulated in MDA-MB-361 cells

To exclude hereditable genetic differences as a source of heterogeneity, I sorted MDA-MB-361 cells into HER2+ and HER2- subpopulations and checked whether these homogenous subpopulations were stable over time, or rather spontaneously gave rise to heterogeneous populations. To fluorescently label HER2+ cells, I stained cells with the BB700 mouse anti-human HER2 antibody as described above, and I performed fluorescence-based sorting with the FACS

aria III sorter instrument, with the standard PerCP-Cy set (data analysis with the FACS DIVA software 8.1). As reported in Figure 5.2A, the stained MDA-MB-361 sample showed a HER2 antibody positivity of 78% when gated on the unstained sample, a value comparable with what I observed in the previous analysis, considering variability in the staining procedure and the semi-quantitative nature of flow cytometry. Then, I sorted HER2+ and HER2- cells into two subpopulations. I cultured separately 4.0×10^5 cells of both subpopulations, with the same culture medium (Appendix A. - Methods) and in the same incubation environment, to avoid any bias in the outcome due to different culture conditions. Both subpopulations re-established the original heterogeneity, demonstrating that HER2 expression in these cells is dynamic and driven by a yet undiscovered mechanism (Figure 5.2A). Interestingly, after literature mining, I found that Jordan et al. (Nature, 2016) published data that corroborate our observation; they show that HER2+ circulating tumor cells, from an ER+/HER2− breast cancer patient, spontaneously interconvert from HER2− and HER2+, with cells harboring a phenotype producing daughters of the opposite one.

In order to identify the biological processes differing between the two subpopulations, we computed the differentially expressed genes (DEGs) from the single-cell transcriptional profiles of $HER2^+$ cells against $HER2^-$ cells. Gene Set Enrichment Analyses (GSEA) against the ranked list of DEGs can be used to identify biological pathways specifically enriched for DEGs. The results of this analysis are reported in Figure 5.2B and revealed seven significantly enriched pathways (FDR<10%): four of which were upregulated in $HER2^+$ cells, but downregulated in $HER2^-$ cells, and included adipogenesis, myogenesis and OXPHOS, all indicative of Epithelial to Mesenchymal Transition (EMT) engagement, which has been reported in $HER2^+$ cells [110], [111]; EMT is a complex remodelling process that causes epithelial cell to change their nature and become able to spread to other tissues. The remaining three pathways were upregulated in $HER2^-$ cells and related to cell-cycle and specifically to G2/M phase, in agreement with our previous analysis, suggesting that cell cycle may play a role in HER2 expression in this cell line.

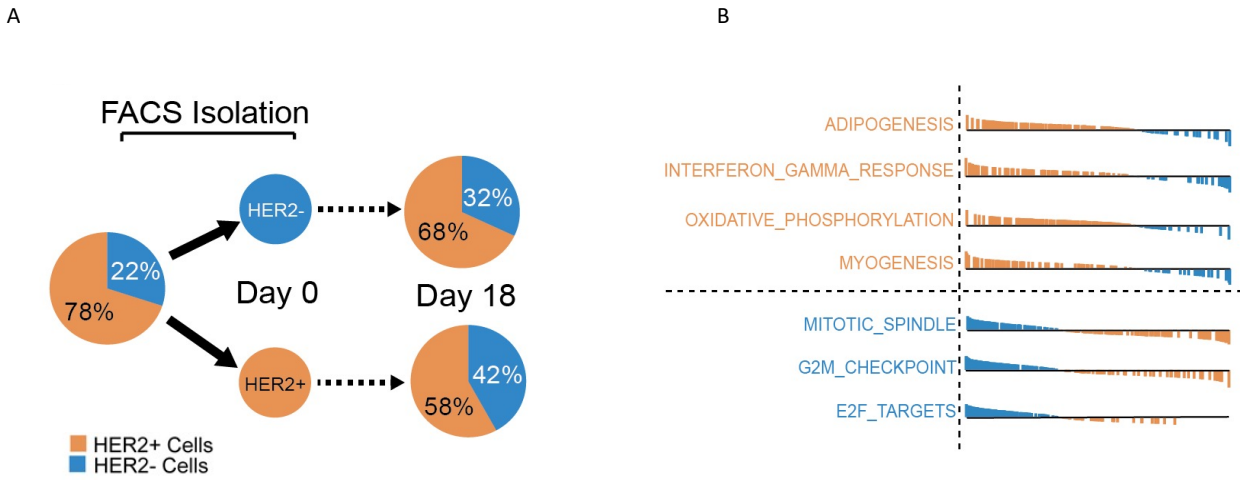


Figure 5.2 - (A) Expression of HER2 protein in MDA-MB-361 cells is dynamic and re-established in about 3 weeks. (B) Gene set enrichment analysis performed against the ranked list of differentially expressed genes obtained by comparing the single-cell transcriptional profiles of the two subpopulations of MDA-MB-361 cells: $HER2^+$ versus $HER2^-$

Pathways related to cell cycle, DNA damage repair and mitotic checkpoint regulation were found to be upregulated in the $HER2^{-}$ subpopulation of MDA-MB-361 cells. This motivated me to further check if the cell cycle phase could explain the observed heterogeneity in the MDA-MB-361 cell line.

The cell cycle is the sequence of events in which a cell grows and duplicates. The eukaryotic cell cycle is divided into four sequential phases: G1, S, G2, and M phase (Figure 5.3), and the progression through cell cycle phases are tightly regulated. The *interphase* includes G1, S, and G2 phases, in which the cell grows and DNA replication occurs, while chromosome segregation and the division process occur in the M phase or *mitosis*. The G1 phase (gap 1) is the beginning of the *interphase*. During G1 the cells double the amount of organelles and proteins (growth of the non-chromosomal components). From this phase, the cell may enter S or G0 phase, depending on several factors, including nutrient and mitogens availability, cell density, but also, as have been described for drug-tolerant cells, in response to cytotoxic or genotoxic agents [112]. G0 is a quiescent phase, in which cells are not actively dividing; cells can reversibly withdraw from the cell division cycle and enter G0 for long time or irreversibly withdraw from cell cycle into terminally differentiated or senescent state, while other cell types never enter G0 by continuously dividing [113]. DNA duplication occurs during the S phase (synthesis), and then, in the G2 phase, the cell prepares for entry in *mitosis*.

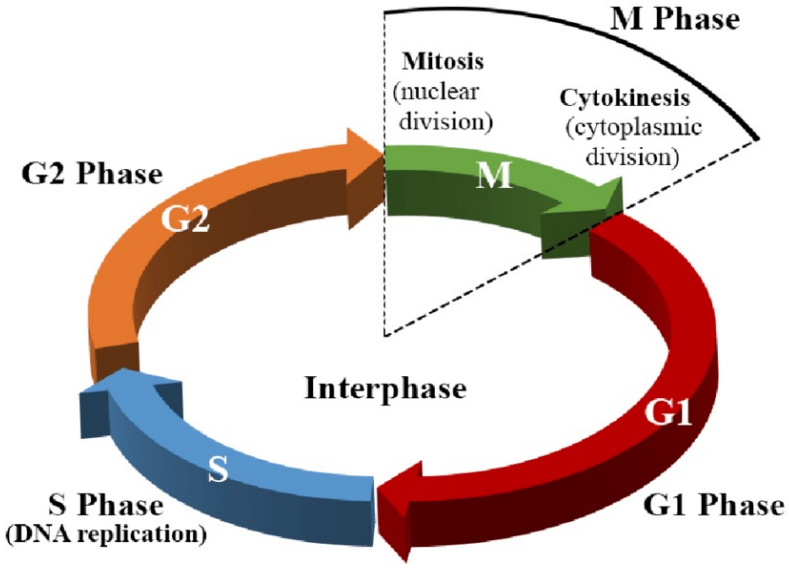


Figure 5.3 – Cell cycle representation of eucaryotic cells. Adapted from Alberts B et al., 2008.

5.3.1 – Cell cycle *in silico* prediction of the MDA-MB-361 cell line

We computationally predicted (Appendix A - Methods) the cell cycle phase of each cell in both the HER2^{-} and HER2^{+} subpopulations from single cell transcriptomics data [114]. To predict the cell cycle phase of each sequenced cell, we used the function *CellCycleScoring* of the *Seurat* tool with default parameter. A list of cell cycle marker genes was provided to *Seurat* (Tirosh et al., 2016), comprising both markers of G_2/M and markers of S phase, while cells expressing neither are considered to be in G_1 phase. A cell cycle score is then assigned to each cell based on the expression level of these cell cycle phase marker genes. As shown in Figure 5.4, a higher proportion of HER2^{-} cells was predicted to be in S/G_2/M phases when compared to HER2^{+} cells. This result is consistent with previous observations that report cell cycle arrest in G_2/M phase following HER2 inhibition [115].

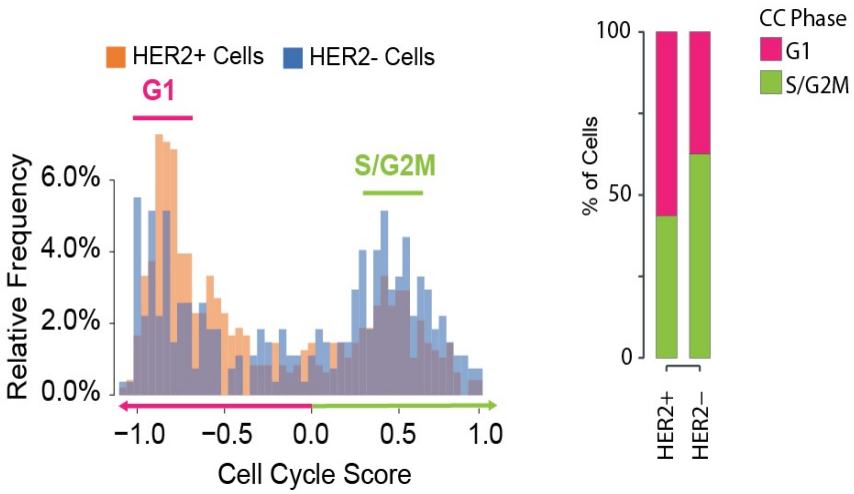


Figure 5.4 – Percentage of HER2^{+} or HER2^{-} MDA-MB-361 cells predicted to be in either G_1 or S/G_2/M phase

5.3.2 – DNA staining protocol optimization for cell cycle analysis

I next set to experimentally validate the cell cycle phase prediction based on the single-cell transcriptomics data, in the MDA-MB-361 cell line. Since cell cycle prediction showed that the HER2− subpopulation is enriched for the G_2/M phase biomarkers, while HER2+ subpopulation for the G_0/G_1 biomarkers, I planned to perform cell cycle arrest of MDA-MB-361 cells in either G_0/G_1 or G_2/M phase and check whether the percentage of HER2+ cells change or not, to assess a possible dependency of the HER2 state to the cell cycle phase.

A method to obtain the cell cycle phase of a cell sample through DNA content analysis is to stoichiometrically stain the DNA with a dye and then detect the emission intensity by flow cytometry, which is proportional to the DNA amount in the cell. Usually, cells in G_2/M phase display a 2x DNA amount than G_0/G_1 cells, since the DNA content has been doubled; cells in S phase, with DNA being synthesized, display an amount of DNA in the range between G_0/G_1 and G_2/M intensity peak.

To stain the DNA of MDA-MB-361 cells I used the propidium iodide (PI) staining protocol. PI is a stoichiometric DNA-binding dye, used from the earliest application of flow cytometry for quantitation of DNA content. I checked the literature to set-up a protocol for PI staining and cell cycle analysis. The procedure consists of two main steps: first fix cells and then stain cells with a staining buffer that includes the PI or other DNA dye. Alcohol fixation is very suitable for DNA staining; I thus added ice-cold ethanol 70% (EtOH 70%) to pelleted cells while gently vortexing them, to prevent cell aggregating during fixation. Then, I stored fixed cells overnight at -20°. During washing steps to remove EtOH, I noticed that centrifugation causes consistent loss of cells. The best practice that I found from literature mining to limit cell loss was to increase the relative centrifugal force (g) to 800g for 8-10 min, rather than using common settings (i.e. 300g for 3 min).

To stain cells with PI staining buffer, 0.025µg/mL of PI (per ~10^5 cells) was suitable for successful DNA staining. However, PI binds RNA in addition to DNA (contributing to unspecific signal), therefore, to achieve specific DNA staining, I added RNase A 50 µg/mL to the staining buffer. After 15 min of room temperature incubation, I analyzed the stained sample with the FACS Accuri C6 (with the 640LP BD standard filter in FL3) and analyzed data with the BD Accuri C6 software. Overall, this procedure yielded a good quality staining (Figure 5.5). Most of the cells were in G_0/G_1 (65.2%), while the G_2/M phase displayed 8.9% of cells and 13.9% in S phase. Other fluorescence signals come from sub-G_0 cells and polynucleated cells with intensity above the G_2/M distribution. The next necessary step to assess the cell cycle phase dependency of the HER2 state was to include in the experimental procedure the conditions for cell cycle arrest to enrich cells in G_0/G_1 and G_2/M phases.

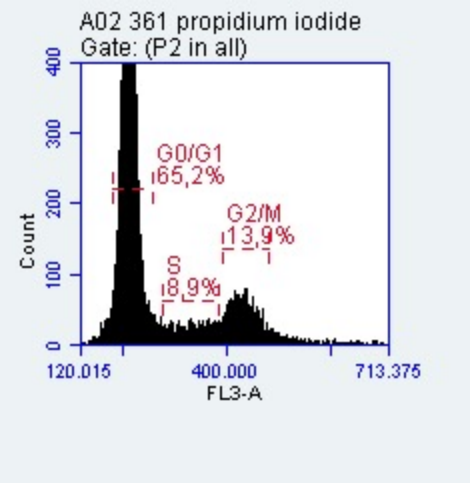


Figure 5.5 – Cell cycle distribution of MDA-MB-361 cells with propidium iodide staining.

5.3.3 – Nocodazole and HBSS protocol optimization for cell cycle arrest and analysis

After literature mining, I selected nocodazole to perform cell cycle arrest in G_2/M phase, and nutrient deprivation with HBSS (starvation) to arrest cells in G_0/G_1. Nocodazole is a compound commonly used to cause cell cycle arrest and subsequent release for cell cycle synchronization; nocodazole inhibits microtubule function by binding β-tubulin and suppressing microtubule and mitotic spindle dynamics, or inducing microtubule depolymerization [116]–[118] arresting cells in M phase (with G_2/M DNA content in flow cytometry analysis).

Starvation with HBSS medium causes arrest of the cell cycle in G_0/G_1 phase; indeed, under HBSS starvation conditions [119]–[121]. Before performing the cell cycle arrest experiment, I optimized the nocodazole and HBSS starvation conditions for MDA-MB-361 cells, in order to minimize toxicity while efficiently arresting cells. To optimize nocodazole concentration and treatment time, I performed two dose-response curves, at 24hr and 48hr of treatment, in 5 nocodazole concentration points, with the max concentration of 800µg/mL and 2-fold dilution series for the other concentration points (Figure 5.6).

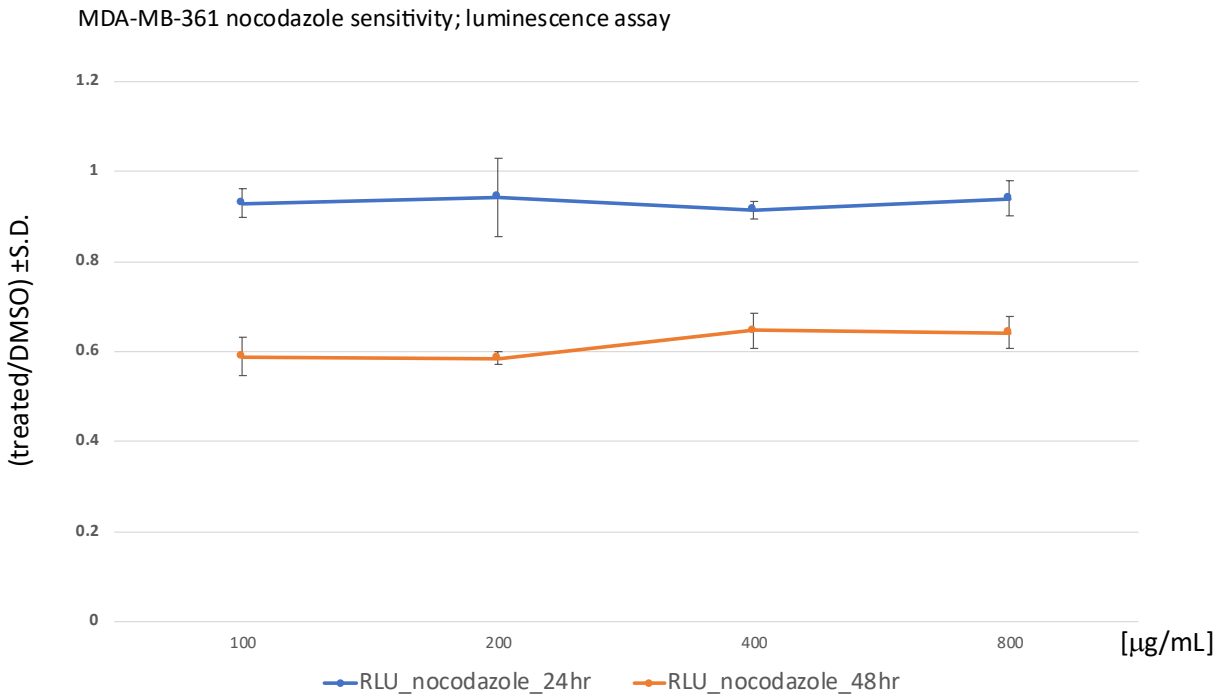


Figure 5.6 – Nocodazole dose-response curves following either 24 hr (blue line) or 48 hr (orange line). Each point is the average of three replicates (± S.D.)

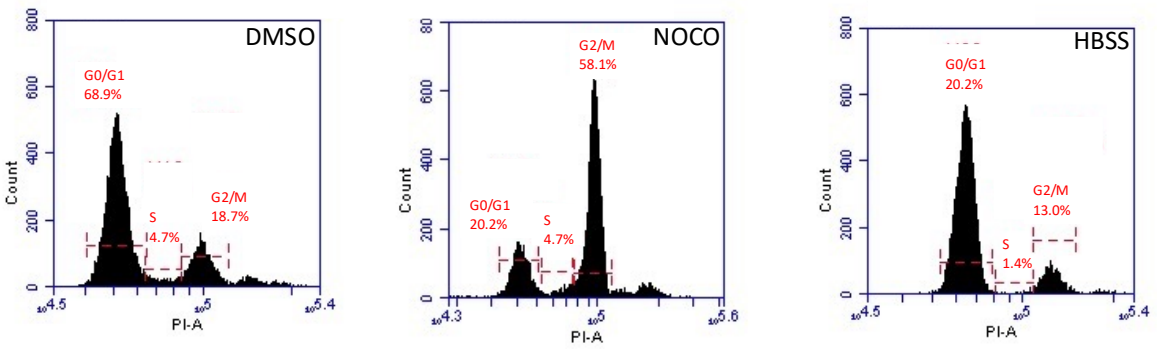
I reported cell viability measurement of nocodazole response normalized to the DMSO negative control. 48 hr of treatment resulted toxic at all concentration points, with cell viability of ~60% ((treated/control)×100); this effect was independent of the nocodazole concentration. By contrast, 24 hr of treatment showed no effect on cell viability at all concentrations (more than 90%). Therefore, I set 800 μg/mL for 24hr condition to perform nocodazole treatment. Then I optimized the recovery of nocodazole M phase arrested cells from the culture vessel, with mitotic shake-off. Indeed, M-phase cells are weakly attached to the culture vessel surface because of their more rounded-up shape, compared to cells in other cell cycle phases. This feature allows enriching for nocodazole M-phase arrested cells with mitotic shake-off, by gently hitting several times the culture vessel to detach M-phase weakly attached cells. However, I made some modifications to the protocol for the MDA-MB-361 cell line to carry out mitotic shake-off. In my experience, MDA-MB-361 cells attach very strongly to both one another and to the culture vessel surface. Hence, I improved the mitotic shake-off for this cell line by prior washing with PBS 1x and then adding a few mL of trypsin; although non-M-phase cells do not detach upon mitotic shake-off since the strong surface binding, this procedure should be fast and gentle as much as possible, to avoid retrieval of cells from other cell cycle phases.

To test HBSS starvation, I cultured cells in either HBSS for 24hr and 72hr or with growth medium as negative control. 24hr of treatment yielded very low cell death when, with a cell viability fold change of -0.20 ± 0.062, calculated as log2(*HBSS*/*control*) ± S.D (average of three replicates), where *HBSS* and *control* are respectively the cell viability that I measured in HBSS and in growth medium. I selected 24 hr of HBSS starvation as a condition to perform cell cycle arrest in G_0/G_1 since 72 hr of HBSS starvation strongly decreased the cell viability (-0.20 ± 0.062; average of three replicates)

5.3.4 – Assessment of the cell cycle contribution to the HER2 of the MDA-MB-361 cell line

In order to check the influence of the cell cycle on the percentage of HER2+ cells in the MDA-MB-361 cell line, I perfomed cell cycle arrest experiments: I incubated MDA-MB-361 cells for 24 hr with either nocodazole 800μg/mL or HBSS starvation, while a DMSO condition served as negative control, to show normal cell cycle distribution. I recovered nocodazole treated cells with mitotic shake-off (described above), while with common cell culture protocol for DMSO and HBSS incubated cells. I then divided each sample into two aliquots: the first one for BB700 mouse anti-human HER2 antibody staining to detect the percentage of HER2+ cells, while the other one for DNA content analysis by PI staining. In addition, I stained HCC38 and AU565 as respectively negative and positive control respectively of the antibody staining. As shown in Figure 5.7A,B, nocodazole treatment successfully enriched cells in G_2/M phase, while cells under HBSS nutrient deprivation displayed arrest in G_0/G_1. Compared to DMSO control (80% HER2+ cells), cells enriched in G_0/G_1 condition showed an increase in the HER2+ cells to ~88%; G_2/M enriched sample, on the contrary, showed a reduction in $HER2^+$ cells (~63%). We observed a contribution of the cell cycle in the HER2 cellular. However, the HER2 state variation, when cells are enriched in either G_0/G_1 or G_2/M phase, is too faint to address to cell cycle as the solely driver of the interconversion between HER2+ and HER2- cell state, highlighting the existence of undiscovered mechanisms in the HER2 state transition.

A



B

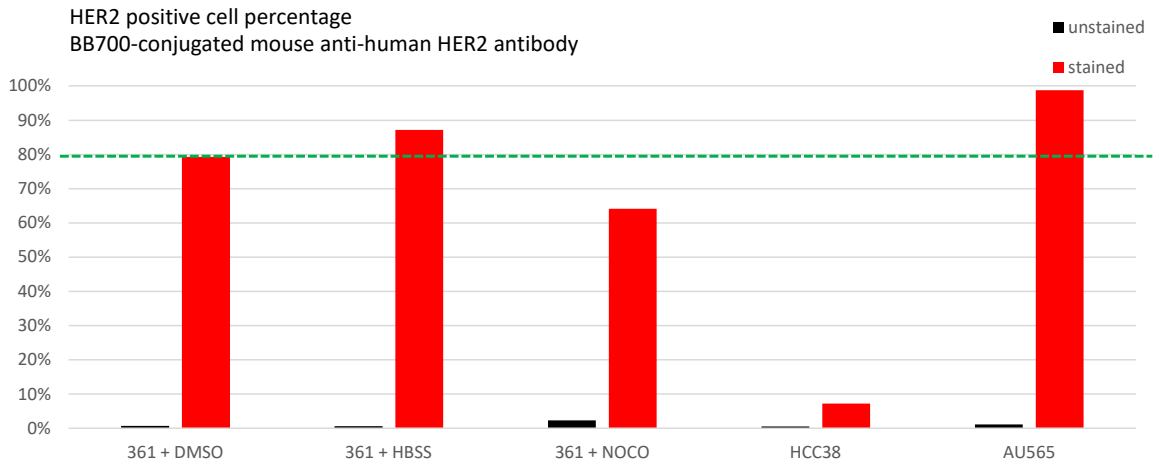


Figure 5.7 – (A) Cell cycle analysis of MDA-MB-361, with propidium iodide staining, in two different conditions (NOCO = nocodazole 800µg/mL; HBSS = HBSS starvation), plus the negative control in DMSO. (B) FACS analysis of MDA-MB-361, stained with the BD BB700 mouse anti-human HER2 in the same condition reported in panel A. I used HCC38 and AU565 as respectively negative and positive control of the antibody staining. Each bar represent the percentage of $HER2^{+}$ cells in each condition.

CHAPTER 6 - Impact of gene expression heterogeneity on drug response

Intrapopulation transcriptional heterogeneity of clinically relevant biomarkers was found to be present in most breast cancer cell lines, raising the possibility that it could be linked to a cell line sensitivity to anti-cancer drugs. In this chapter, I illustrate the experiments I performed to study the impact of intrapopulation heterogeneity on drug response. Then, I describe DREEP, a computational method we developed to predict in single cells the effect of a panel of 450 drugs. With DREEP, we found that etoposide, a common chemotherapeutic agent, is more effective on the HER2- subpopulation as compared to the HER2+ subpopulation of the MDA-MB-361 cell line. I performed the experimental validation of the effect of etoposide on both HER2+ and HER2- sorted subpopulations, which confirmed the computational predictions. Surprisingly, afatinib, an HER2 inhibitor, was equally effective on both subpopulations. To explain this observation, we formalized the interconversion between the HER2+ and HER2- cell state with a mathematical model and eventually I experimentally validated its predictions.

6.1 – Drug sensitivity correlates with intrapopulation drug target heterogeneity

To investigate the role of heterogeneity in gene expression within a cell line on the efficacy of targeted anticancer treatment, we collected large-scale in vitro drug screening data [28], [29] reporting the effect of 450 drugs on 658 cancer cell lines from solid tumours. We correlated the percentage of cells expressing the HER2 receptor in the cancer cell line using our single-cell breast cancer cell line atlas, with the toxicity caused in the cell line by treatment with specific HER2 inhibitors. The toxicity is expressed as the relative Area Under the Curve (AUC) of dose-response curves measuring the effect of the drug on cell viability. The AUC is a robust metric to measure the effect of a drug across cell lines [122], [123]. When the response to a drug is reported as cell viability of the treated sample normalized to the untreated negative control, the lower the AUC the higher the effect of the drug, over the concentration range. However, the AUC metric depends on the range of tested drug concentrations, which often varies between studies.

As show in Figure 6.1A, the sensitivity of the breast cancer cell lines to HER2 inhibitors was significantly correlated with the percentage of cells in the cell line expressing *HER2*. For example, afatinib has a stronger effect (i.e. lower relative AUC) in AU565, which is homogeneous for HER2 expression, as compared to the more heterogenous MDA-MB-361 cell lines, while lower effect on MDA-MB-361 cell lines, which is more heterogeneous, while no effect on cells not expressing *HER2* (e.g. HCC38). Interestingly, receptor expression level is substantially the same across cells expressing it, irrespective of the cell line they belong to (Figure 6.1B), except for cell lines harbouring CNVs of the *ERBB2* locus.

Furthermore, I found that the correlation between drug target expression and drug sensitivity holds true also for several other targets (Figure 1C), thus suggesting that variability in gene expression within cells of the same tumour may cause some cells to respond poorly to the drug treatment.

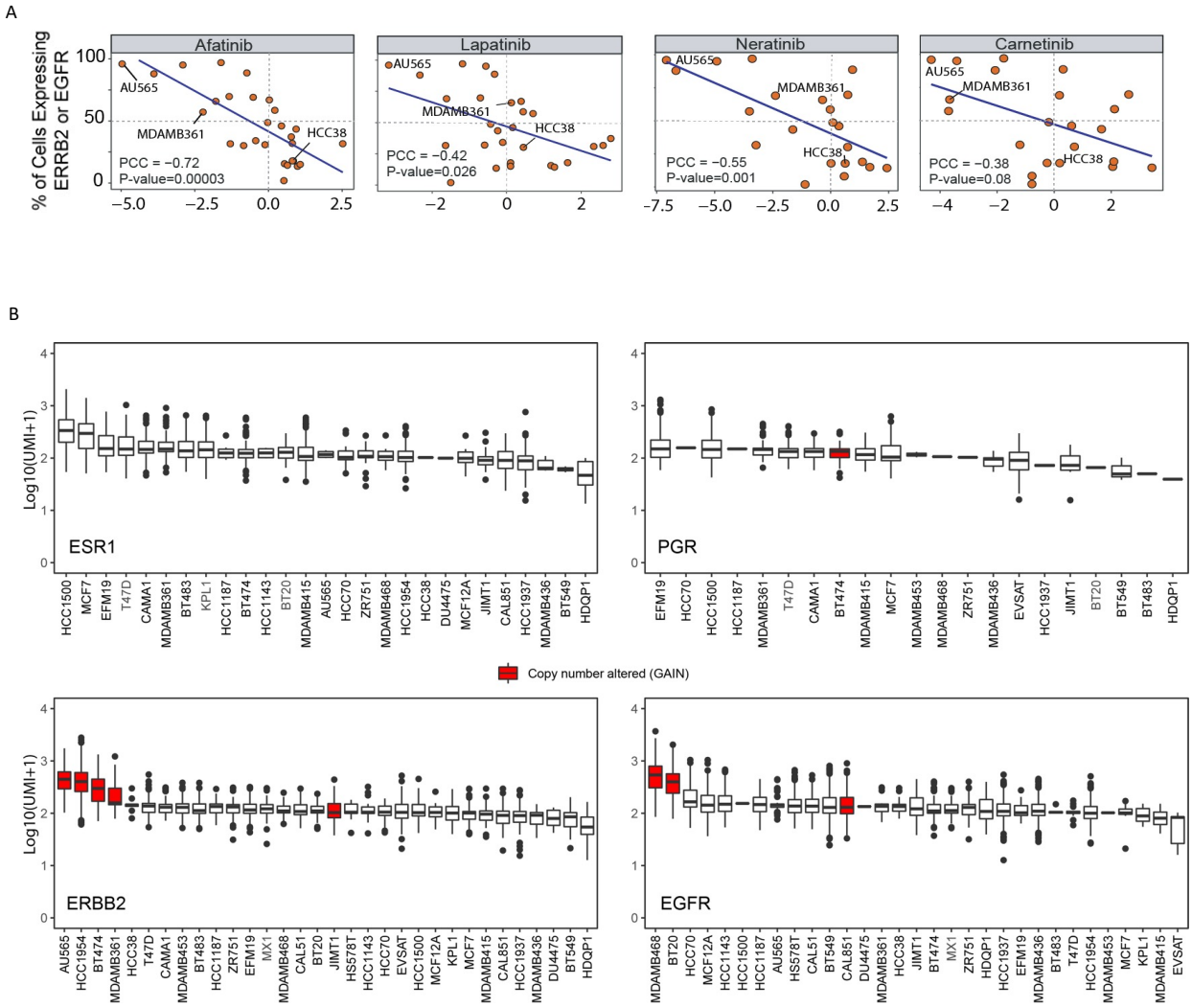


Figure 6.1 – **Drug target heterogeneity affects drug response**. (A) Relationship between percentage of cells expressing ERBB2 in the cell line [y-axis] and the drug potency in the same cell line [x-axis] for the indicated molecules. The more negative the value on the x-axis, the more potent the drug. Drug potency data were retrieved from the CTRPv2 or GDSC databases. Each dot represents a cell line we have sequenced with single cell transcriptomics. (B)). Overall expression receptor across cell lines. The expression level is the same, except for CNV, such as ERBB2 or EGFR copy number gain.

To confirm the results reported in the literature, I decided to evaluate the impact of HER2 intra-population heterogeneity by testing the effect of afatinib on AU565, MDA-MB-361 and HCC38 as representative CCL models. For each CCL, I performed the drug response assay to afatinib at 24 and 72 hr of treatment, spanning five concentrations, with the maximum concentration at 4μM (Figure 6.2A,B). To read out the effect of afatinib at 24 and 72 hr, I performed a luminescence-based assay to measure the luminescence intensity of metabolically active cells (viable cells) of the treated samples, and I normalized this value to the DMSO negative control luminescence intensity (Methods). In this way, I obtained a measurement of the cell viability, that I expressed as percentage ((treated/DMSO) × 100). Overall, the effect of afatinib reflected the intra-population HER2 expression heterogeneity. Following 24 hr of treatment, HCC38 cell line showed no response over the whole concentration range. For both AU565 and MDA-MB-361 cell lines, afatinib showed mild effect at 24 hr, but still at the higher concentrations the AU565 cell viability dropped down (53%) compared to MDA-MB-361 that never decreased below 72%. Following 72 hr of treatment, AU565 resulted highly sensitive to afatinib, even at a low concentration, with full response at 4μM (10%), while HCC38 responded only at the highest concentration of 4μM (56%), with a weak toxicity at 1μM (85%). Interestingly, the MDA-MB-361 cell line resulted partially sensitive to afatinib at 1μM (53%) and highly responsive at 4μM of afatinib (20%). Indeed, the dose-response curve of MDA-MB-361 cell line to afatinib lies in the middle between AU565 (high sensitivity) and HCC38 (low or no sensitivity). Overall, in this experiment I assessed that a homogenous cell line for a drug target like AU565 tend to be more sensitive to the drug target inhibitor than a more heterogenous cell line (MDA-MB-361), and in this condition the concentration of drug needed to observe a toxic effect is lower for the homogenous population.

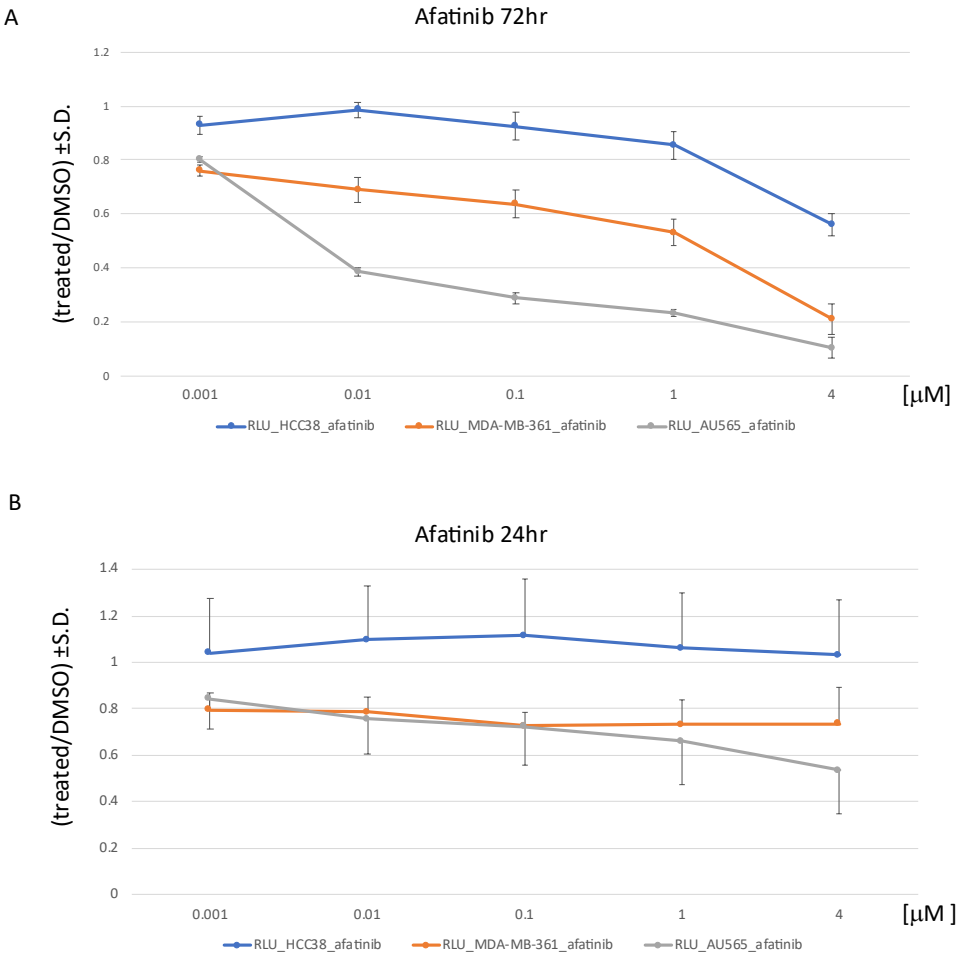


Figure 6.2 – **The percentage of cells expressing HER2 affects the effect of afatinib**. Afatinib dose response curves following 72hr (A) and 24hr (B) against HCC38 (blue line), AU565 (grey line) and MDA-MB-361 (orange line); S.D. = standard deviation; each point is the average of three replicates.

6.2 – DREEP: a bioinformatics tool to predict drug response at the single cell level

The observations described in Section 6.1 led us to develop DREEP (DRug Estimation from single-cell Expression Profiles), a novel bioinformatics tool, that, starting from single-cell transcriptional profiles, allows to predict drug response at the single cell level. To this end, we first detected expression-based biomarkers of drug sensitivity for 450 drugs [29], as schematised in Figure 6.3A (Appendix A - Methods). Briefly, we crossed data from the Cancer Cell Line Encyclopaedia (CCLE) on the response to 450 drugs across 658 cancer cell lines from solid tumours with their gene expression profiles from bulk RNA-seq. In the CCLE, drug potency is evaluated as the inverse of the Area Under the Curve (AUC) of the dose-response graph, with low values of the AUC indicating drug sensitivity, while high values implying drug resistance. For each gene and for each drug, we computed the correlation between the expression of the gene across the 658 cell lines with the drug potency in the same cell lines. Hence, genes positively

correlated with the AUC are potential markers of resistance, vice-versa, negatively correlated genes are markers of sensitivity. In this way, we generated a ranked list of expression-based biomarkers of drug sensitivity and resistance for each of the 450 drugs. We then used these biomarkers to predict drug sensitivity at the single-cell level. To this end, for each cell in the single-cell BC atlas, we selected the 250 genes most expressed in that cell and compared them against the ranked list of biomarkers for each one of 450 drugs by means of Gene Set Enrichment Analysis (GSEA) [124]. A negative enrichment score implies that highly expressed genes in that cell are enriched for biomarkers associated to sensitivity, whereas a positive enrichment score implies enrichment for resistance-associated biomarkers. At the end of this process, each cell in the atlas is associated to the drug it is most sensitive to, or to no drug, if no significant enrichment score from GSEA is found (Figure 6.3B).

To assess the algorithm's performance, we applied it to the single-cell BC atlas and estimated its performance by checking how well we could predict sensitivity of the 32 BC cell lines to 86 drugs for which this information was publicly available from GDSC database (Genomics of Drug Sensitivity in Cancer) [125]. To convert single-cell predictions to predictions at the cell line level, we simply used the percentage of cells in the cell line deemed to be sensitive to the drug by the algorithm. The algorithm precision is shown in Figure 6.3C.

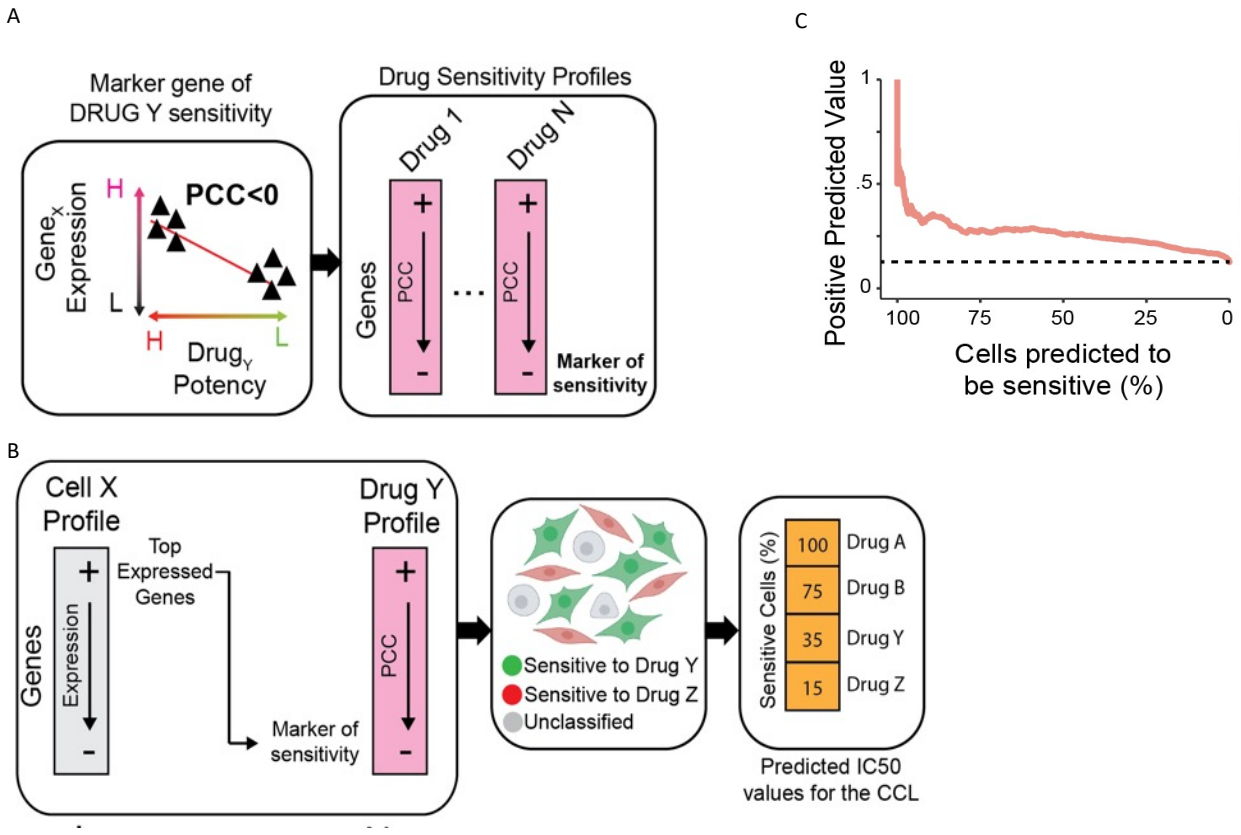


Figure 6.3 – **DREEP drug sensitivity prediction and PPV analysis**. (A) Construction of the ranked list of drug sensitivity biomarkers for 450 drugs. For each gene and for each drug, the expression of the gene is correlated with the potency of a drug expressed as a function of Area Under the Curve (AUC) across 658 cell lines. (B) The top 250 most expressed genes of a cell are used as input for a Gene Set Enrichment Analysis (GSEA) against the ranked list of biomarkers for each one of the 450 drugs, to predict single cell drug sensitivity. At the end of the process, each cell in the sample is associated to the drug it is most sensitive to, or to no drug, if no significant enrichment score from GSEA is found. Finally, for each of the 450 drugs, the number of cells predicted to be either sensitive, resistant, or not classified in the considered sample are estimated. (C) Validation of the computational method using the Breast Cancer Single Cell atlas data to predict drug sensitivity to 86 drugs for which the corresponding half maximal inhibitory concentration (IC50) was available from the GDSC database. The PPV (Positive Predicted Value) is shown as a function of the percentage of cells predicted to be sensitive for each drug-BC CCL interaction. Dashed line represents the performance of a random algorithm.

6.3 – Experimental validation of DREEP drug sensitivity prediction

To experimentally validate DREEP, I turned to the MDA-MB-361 cell line for which we found coexistence of two distinct and dynamic cell subpopulations ($HER2^+$ and $HER2^-$). We applied DREEP to each subpopulation to identify drugs able to selectively inhibit growth of either the $HER2^-$ subpopulation or the $HER2^+$ subpopulation: 42 drugs (FDR < 1%, Appendix D – Supp. Table D2) were predicted to preferentially inhibit growth of $HER2^-$ cells; the most overrepresented class among these drugs was that of inhibitors of DNA topoisomerases (TOP1/TOP2A) (Figure 6.4A) such as etoposide, that is a TOP2A poison (TOP2A resulted more expressed in the $HER2^-$ subpopulation). Surprisingly, no drug was found to specifically inhibit growth of $HER2^+$ cells, whereas 44 drugs (FDR <1%) were predicted to be equally effective on both subpopulations and unexpectedly included HER2 inhibitors, such as afatinib (Figure 6.4B,C).

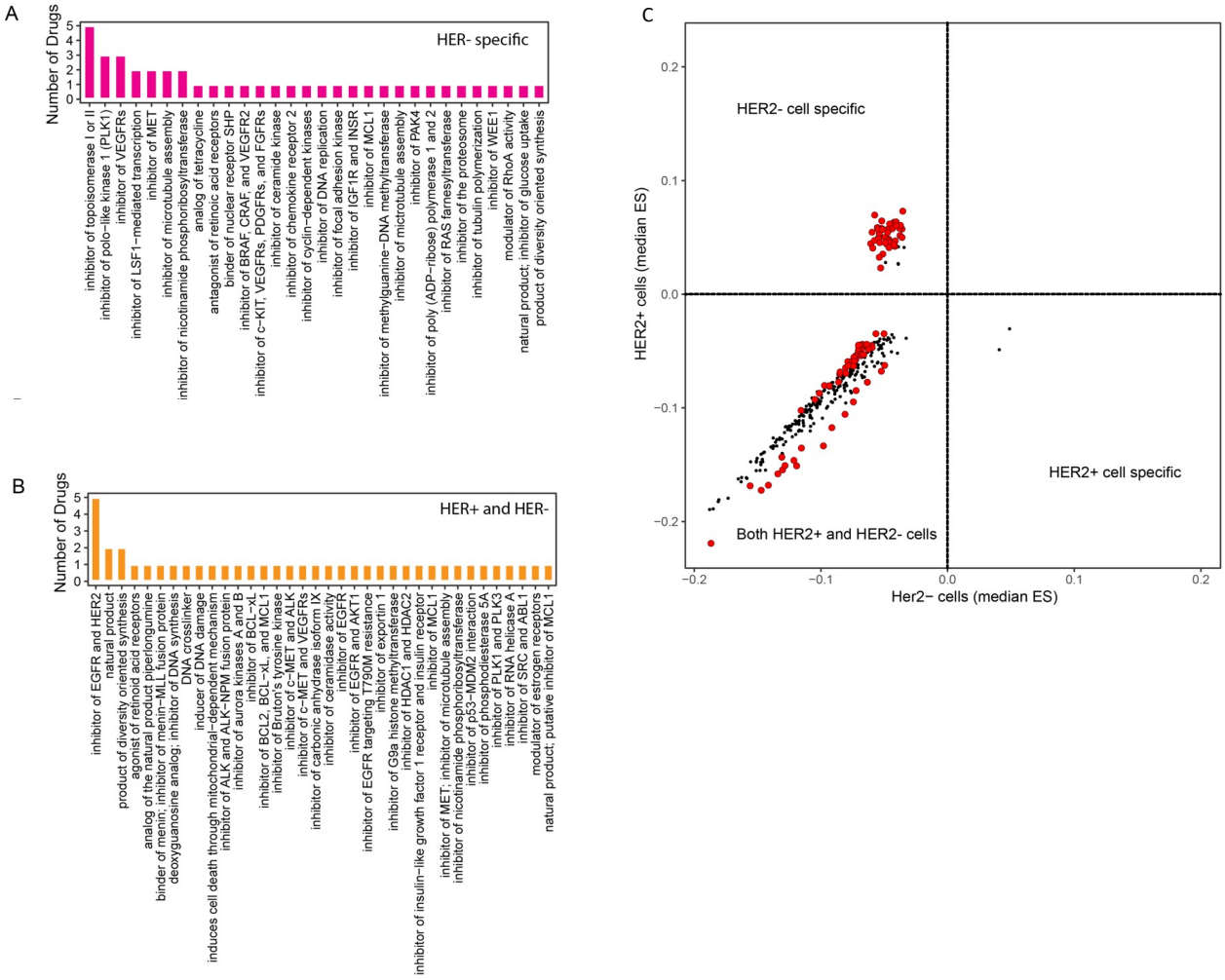


Figure 6.4 – **Drug sensitivity prediction for HER2+ and HER2- cells of MDAMB361 cell-line**. Classification of drugs predicted to specifically inhibit growth of the $HER2^-$ subpopulation, or both $HER2^+$ and $HER2^-$ subpopulations.

I selected etoposide and afatinib for the experimental validation of DREEP. The experimental plan I designed consisted of separating by FACS sorting the two subpopulations ($HER2^+$ and $HER2^-$) in the MDA-MB-361, and then check the differential drug sensitivity between the $HER2^+$ and $HER2^-$ cells against etoposide and afatinib, as schematized in Figure 6.5A. Here, I decided to read out the cell viability by nuclei staining with Hoechst 33342 and nuclei count with the Operetta microscope (Appendix A - Methods). To sort the MDA-MB-361 into $HER2^+$ and $HER2^-$ subpopulations, I stained MDA-MB-361 cells with the mouse anti-human HER2 BB700-conjugated antibody. The staining allowed to selectively label the $HER2^+$ cells, and then cells were sorted in $HER2^+$ and $HER2^-$ subpopulations with the FACS Aria III flow cytometer (PerCP-Cy set; data analysis with the FACS DIVA software 8.1). I collected both subpopulations to perform the drug sensitivity assays against etoposide and afatinib. To this purpose, I seeded 15,000 cells per well both for the $HER2^+$ and $HER2^-$ subpopulation in 96-well plate and after overnight incubation, I exposed cells to either etoposide or afatinib for 72 hr, spanning five different concentrations. In addition, I used DMSO for the negative control, in order to normalize viability

data. In agreement with DREEP predictions, HER2^{-} cells were much more sensitive to etoposide than HER2^{+} cells, which responded only at the higher concentrations, while afatinib was equally effective on both subpopulations (Figure 6.5B). This counterintuitive result was similar to that observed by Jordan et al. [126] using a BC patient's circulating tumor cells sorted into HER2^{-} and HER2^{+} subpopulations, which were found to be equally sensitive to Lapatinib (another HER2 inhibitor), but no mechanism of action was put forward.

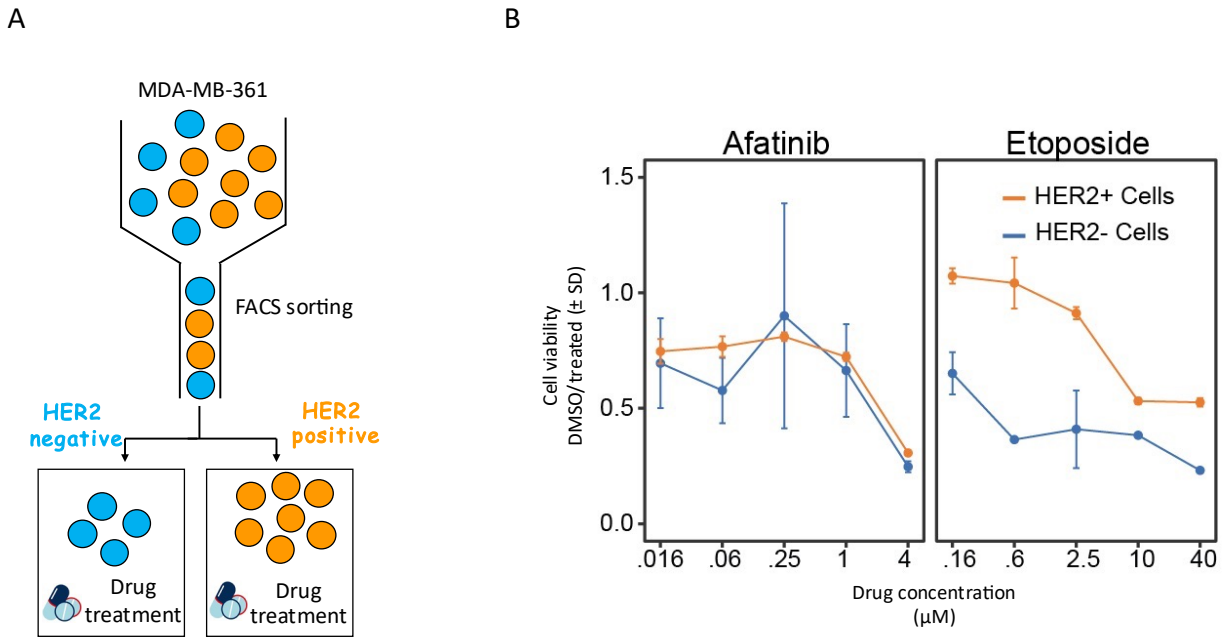


Figure 6.5 – Afatinib and etoposide drug sensitivity assay on sorted HER2^{+} and HER2^{-} subpopulations. (A) Depiction of the MDA-MB-361 HER2^{+} and HER2^{-} subpopulation sorting o. (B) Dose-response curve for afatinib and etoposide on sorted MDA-MB-361 cell populations (triplicate experiment); S.D = standard deviation.

I then performed a dose response curve to etoposide on the MDA-MB-361 cell line (without sorting) to validate what we observed in the sorting experiment on the two subpopulations, as shown in Figure 6.6. I obtained viability date by means of a luminescence assay, as described above. In Figure 6.6, I overlapped in the same plot the etoposide dose response curve result with the etoposide curves of the sorted HER2^{-} and HER2^{+} subpopulations in the previous experiment. The effect on the mixed population was very similar to the effect on the HER2^{+} subpopulation; indeed, the percentage of HER2^{+} cells in MDA-MB-361 cell line is approximately 70-80%, therefore the etoposide response of the HER2^{+} cells contributes much more to the overall response than the HER2^{-} subpopulation.

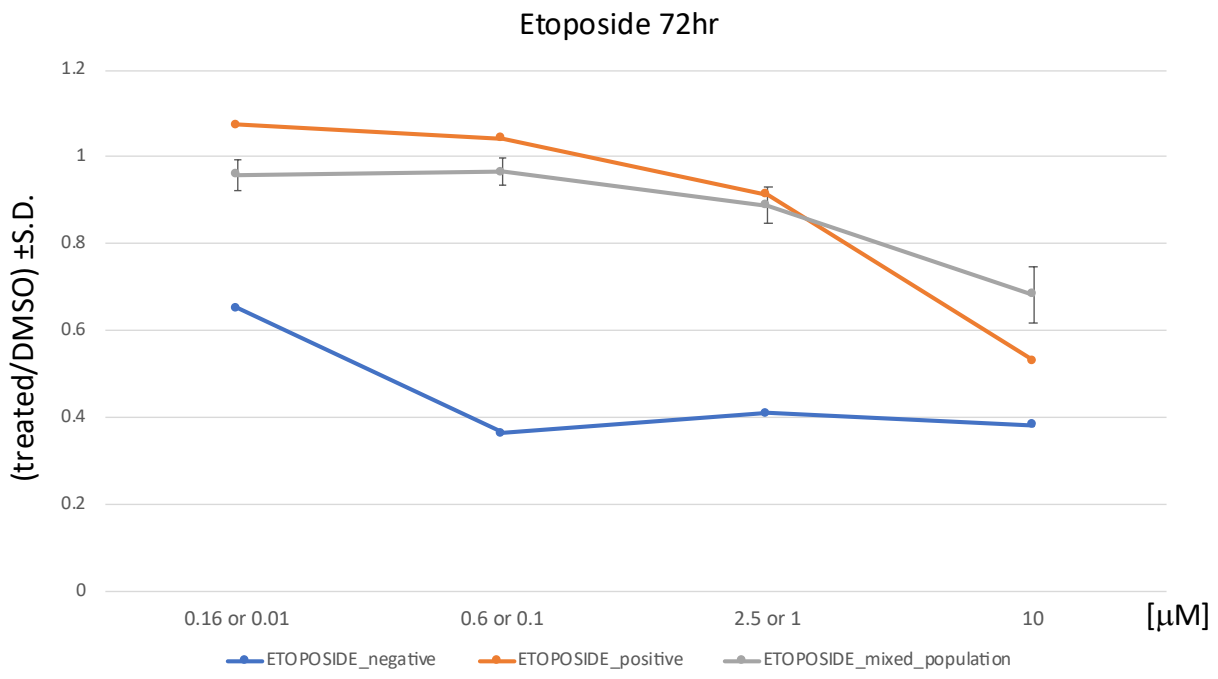
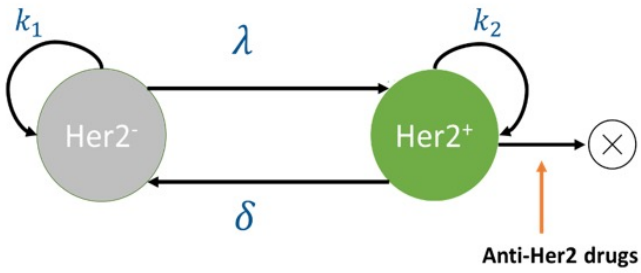


Figure 6.6 – **Etoposide drug sensitivity assay**. Orange and blue line are respectively the dose response curves obtained for the experiment in Figure 6.5; here, the blue line represents the dose response curve of the HER2$^-$ subpopulation, while the orange line represents the dose response curve of the HER2$^+$ subpopulation. I removed the higher drug concentration point for both curves. The grey line represents the etoposide dose response curve of the MDA-MD-361 cell line without sorting. For simplicity, I show only error bars for the 72hr treatment with etoposide of the unsorted MDA-MB-361 population (grey curve); refer to Figure 6.2 for all information about the blue and the orange curve.

6.4 – Mathematical model to explain the HER2 state dynamic interconversion

We hypothesised that the dynamic interconversion of MDA-MB-361 cells between the HER2$^-$ and the HER2$^+$ state may explain the counterintuitive effectiveness of Afatinib on the HER2- subpopulation of MDA-MB-361 cells. Indeed, cells not expressing HER2 should not respond to HER2 inhbition. We hypothesised that when the starting population consists of HER2$^-$ cells only, as following FACS sorting described in section 6.2, some of these cells will nevertheless interconvert to HER2$^+$ cells during afatinib treatment, and they will thus become sensitive to HER2 inhibition, explaining the observed results. We mathematically formalised this hypothesis with a simple mathematical model depicted in Figure 6.7 and Appendix B. In the model, two species (HER2$^+$ and HER2$^-$ cells) can replicate and interconvert, but only one (HER2$^+$) is affected by afatinib treatment. The model shows that if the interconversion time between the two cell states is comparable to the cell cycle duration, then afatinib treatment will have the same effect independently of whether the initial population consists of HER2$^+$ cells only, or HER2$^-$ cells only. If instead the interconversion time is much longer than the cell cycle, then afatinib will have little

effect on HER2$^-$ sorted cells, but maximal effects on HER2$^+$ sorted cells, and vice-versa, if the interconversion time is much shorter than the cell cycle, then afatinib's effect would be minimal on both HER2$^-$ and HER2$^+$ sorted cells.



λ	0.6 days^{-1}
δ	0.4 days^{-1}
k_1	0.2 days^{-1}
k_2	0.2 days^{-1}

Figure 6.7 – **A two state model of interconversion of MDAMB361 cells.** A two-state model of interconversion of MDAMB361 cells with arrows indicating reactions occurring at the rates reported on the arrow with values in the table.

Comparison of the modelling results with the experimental results thus suggests that the interconversion rate should be of the same order of the cell cycle (about 72h for MDA-MB-361 cells). The model further predicts that treating the unsorted population of MDA-MB-361 cells with afatinib will reduce the percentage of HER2$^+$ cells, since only HER2$^+$ will be affected, but that this percentage would quickly recover once afatinib treatment is halted (Figure 6.8A,B).

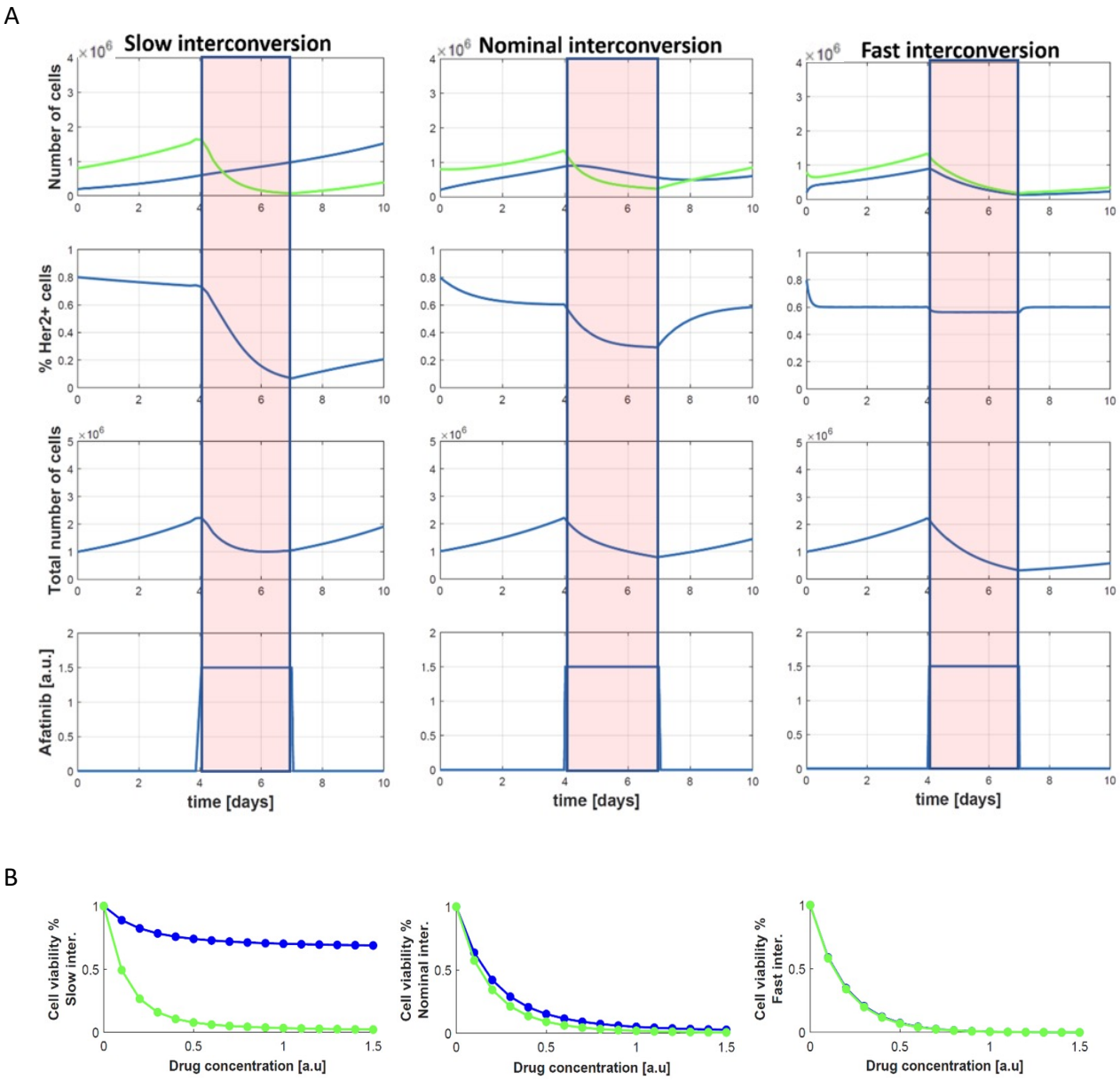


Figure 6.8 – **Simulation of the effect of afatinib.** (A) Numerical simulations of the effect of Afatinib on MDAMB361 cell line. Three different sets of parameters' values were used to investigate the effect of changing the interconversion rate on the response of MDMA261 cells to afatinib treatment. Simulations start with a total of 1 million cells, of which 0.8 million $HER2^+$ and 0.2 million $HER2^-$. In the first row, the green line stands for $HER2^+$ cells and the blue line for $HER2^-$ cell. (B) Simulated dose response curve to afatinib for MDAMB361 cells. The simulated cell viability was obtained by setting both the HER2- and HER2+ populations at $5x10^5$ cells at simulation time zero and then running the simulation with or without Afatinib at the indicated concentrations for 10 days, and then dividing the resulting number of $HER2^-$ cells (resp. $HER2^+$) treated with Afatinib by the number of $HER2^-$ cells (resp. $HER2^+$) grown in the absence of Afatinib. Cell viability was simulated for the three different set of parameters' values. Green line refers to $HER2^+$ cells while the blue line to $HER2^-$ cells.

6.5 – Experimental validation of the modelling prediction

In order to validate the modelling predictions, I exploited the MDA-MB-361 cell line without sorting to check for a possible variation in the percentage of $HER2^+$ cells upon afatinib and etoposide treatment, as assessed by flow cytometry. I incubated MDA-MB-361 for 72 hr with either 10μM etoposide or 1μM afatinib, and DMSO for the negative control. Following treatment, I stained cells with the mouse anti-human HER2 BB700-conjugated antibody and processed cell by flow cytometry with the FACS Accuri C6 (detection with the 640LP standard filter) and then analyzed data with the Accuri C6 software. As negative and positive control of the antibody staining, I respectively stained, HCC38 and AU565 cell lines. As shown in Figure 6.9B, etoposide and afatinib had the same effect on the MDA-MB-361 cell viability by luminescence assay (normalized to the DMSO negative control), however, as reported in Figure 6.9A, etoposide increased the percentage of $HER2^+$ cells, in agreement with the increased sensitivity of $HER2^-$ cells to this treatment, whereas afatinib treatment strongly decreased the percentage of $HER2^+$ cells, confirming that its effect is specific for $HER2^+$ cells only.

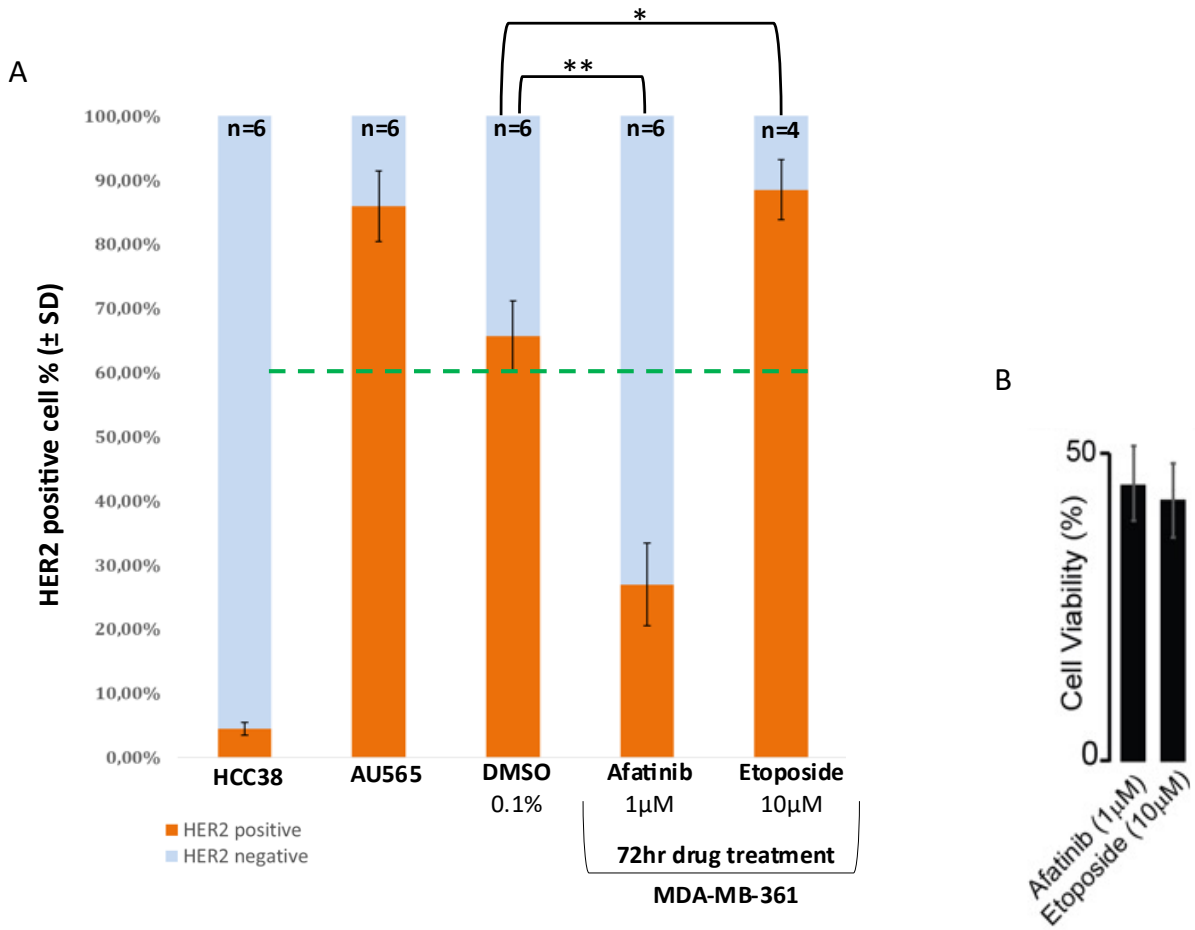


Figure 6.9 – FACS analysis and cell viability of the effect of afatinib and etoposide on the MDA-MB-361 cell line. (A) Percentage of HER2+ cells in MDA-MB-361 after 72h treatment with either afatinib (statistic: two-sided t-test, *P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001) or etoposide, with the number of replicates specified for each measurement, and (B) measured cell viability after the treatment (average of three replicates ± S.D.).

To further confirm the modelling results, I then decided to test the interconversion dynamics of HER2 in MDA-MB-361 following afatinib perturbation (Figure 6.10). I incubated MDA-MB-361 with afatinib 0.1 μM for 48 hr, and also HCC38 and AU565 cell lines in the same conditions; HCC38 here is a negative control for the toxicity of afatinib (since nearly no cells express HER2), while AU565 was meant as positive control of the afatinib toxicity and also to compare the HER2 state variation with the MDA-MB-361. I measured the percentage of $HER2^+$ cells and cell viability following afatinib removal from the medium at three time points: t_0, 72 hr and 144 hr, where t_0 corresponded exactly to afatinib removal (Figure 6.10). Results of the experiment are reported in Figure 6.11; as expected, for the HCC38 cell line, no $HER2^+$ cell percentage variation occurred and HCC38 cells were not sensitive to afatinib by checking the cell number over time, with the exception of the 144 hr time point, where cells died for overgrowth in the culture vessel. In the case of MDA-MB-361 cells, afatinib showed very little toxicity effect following 48 hr at 0.1 μM, while the percentage of $HER2^+$ cells in the population decreased as expected (41% of $HER2^+$ cells at t_0). Within 72 hr following afatinib removal, the percentage of $HER2^+$ cells quickly increased from 41% to 74% (compared to the DMSO) and was fully recovered at 144 hr (85%). This observation successfully confirmed our modelling prediction of the HER2 state interconversion dynamic. By checking the cell number over time, the growth of afatinib treated cells resulted slower than DMSO, suggesting that afatinib treated cells where not growing until drug removal; a possible explanation is that afatinib exerted a cytostatic effect at 0.1 μM. Interestingly, AU565 lost the HER2 homogenous state at 72 hr (66% at 72hr compared to 91% of the DMSO), showing a slower drop down of the $HER2^+$ cell percentage than MDA-MB-361, while after 144 hr the HER2 state returned to the original proportion.

All together our results show that dynamic heterogeneity in gene expression does play a significant role in how the cell population will respond to the drug treatment.

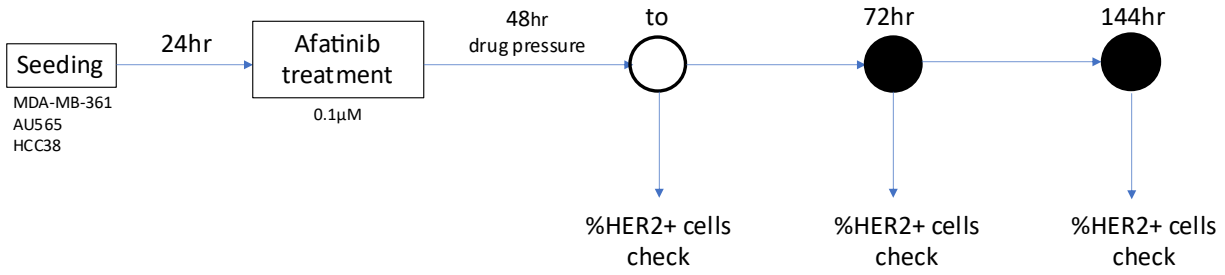


Figure 6.10 – Schematic representation of the experimental protocol to measure interconversion dynamics of MDA-MB-361 cells following afatinib treatment and the HER2 state measurement by flow cytometry. I incubated cells with afatinib for 48hr, after incubation of 24hr. At t_0 I removed the medium with afatinib and replaced with growth medium without afatinib. At t_0, 72 and 144hr I checked by flow cytometry the percentage of $HER2^+$ cells.

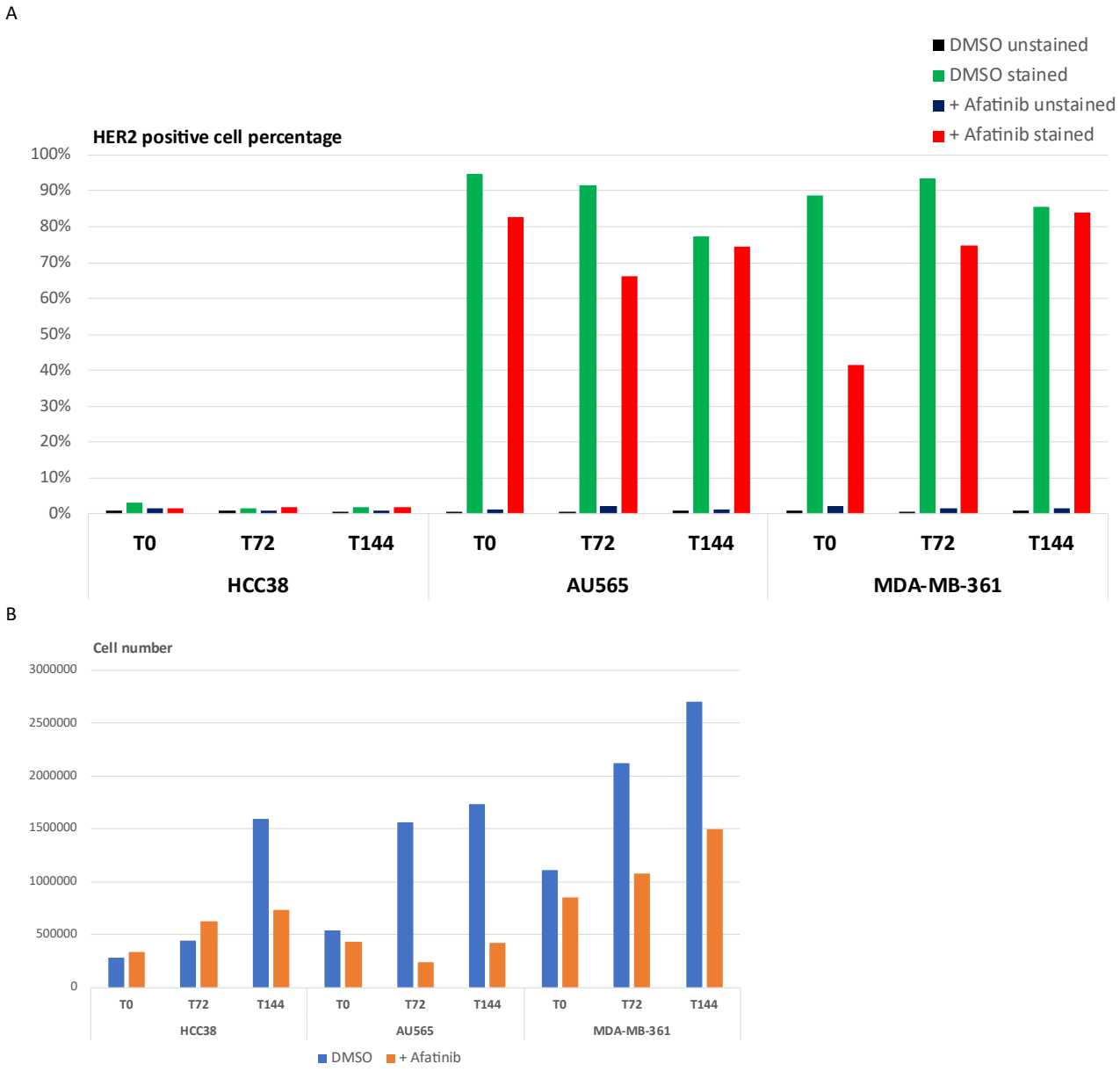


Figure 6.11 – HER2 interconversion dynamics following afatinib treatment. (A) Percentage of HER2 positive cells in MDA-MB-361 cell-line either after 48h of afatinib pre-treatment (red bars) or without any afatinib pre-treatment (green bars) and (B) number of counted cells at each timepoint.

6.6 – Normalized growth rate inhibition to assess the afatinib effect on growth rate

In the experiments described on Section 6.5, I noticed a possible cytostatic effect of afatinib on the MDA-MB-361 cell line at a concentration 0.1 μM. In order to confirm this effect, and to find the concentration at which afatinib switches from cytostatic to cytotoxic, I decided to perform follow-up experiments. In drug sensitivity experiments metrics of drug potency and efficacy (i.e. IC_{50} and E_{max}) are used to estimate the drug effect, which takes into account the relative cell viability of the treated sample normalized to the negative control. However, differences in the proliferation rates can confound the drug effect read out and a cytostatic effect, that blocks treated cell growth, could be confounded as cytotoxic when compared to the negative control. To verify the cytostatic effect of afatinib on MDA-MB-361, I used the drug-induced growth rate inhibition with the GR metric [127]. The GR metric is based on comparing growth rates between the treated samples and the negative control (in the presence and absence of drug), which allows compensating for the confounding effects of division rate on drug response measurements. A GR value that depends only on the drug concentration is calculated as:

$$GR(c) = 2^{\frac{k(c)}{k(0)}} - 1$$

where k(c) is the growth rate of drug-treated cells and k(0) is the growth rate of untreated control cells, and calculated as:

$$k(c) = log_2(\frac{x(c)}{x_0})$$

$$k(0) = log_2(\frac{x(ctrl)}{x_0})$$

Where $x(c)$ is the count in the presence of drug and x_{ctrl} is the cell count for control cells (i.e. DMSO), while x_0 is the cell count from a sample grown in parallel and measured just prior to drug exposure. Therefore, it is possible to calculate a GR value for each drug concentration; As reported in the paper [127], the sign of the GR value relates directly to response phenotype: it lies between 0 and 1 in case of partial growth inhibition, equals 0 in the case of complete cytostasis and lies between 0 and −1 in case of cell death. Values of 1 or higher means that the drug is probably enhancing the growth, however no description of this specific case has been reported in the paper.

I seeded MDA-MB-361 cells and after overnight incubation, I treated cells for 72 hr with five drug concentrations (c_1,…,c_5) in order to obtain each $x(c)$ value (five GR values). I added DMSO in an additional sample as negative control and to obtain x_{ctrl} , and, in addition, a further sample to be measured exactly at the time of the afatinib and DMSO exposure to obtain x_0, that correspond to the measurement at time zero. I performed all measurements with a luminescence assay, as described above. All the results are reported in Table 6.1.

µM	Luminescence_value
0.001	1.96E+05
0.01	1.70E+05
0.1	1.36E+05
1	1.21E+05
4	7.82E+04
x0	1.41E+05
DMSO	2.02E+05

Parameter calculation		
µM	**k(c)**	**GR_values**
0.001	0.4719813	0.88515964
0.01	0.26728638	0.43196241
0.1	-0.0545002	-0.0705945
1	-0.217235	-0.2530909
4	-0.8522205	-0.6817085
	k(0)	
	0.51600324	

Table 6.1 – Normalized growth rate inhibition values (GR) for afatinib treatment of the MDA-MB-361 cell line.

At 0.1 µM of afatinib the GR value is very close to 0. According to my hypothesis, this means that 0.1 µM is the concentration at which afatinib exerts a cytostatic effect, and the switch to cytotoxic approximately happens at 1 µM (at least one order of magnitude more), where the negative GR is indicative of cell death.

CHAPTER 7 - Conclusions

In this thesis, I describe how I implemented the Drop-seq single-cell sequencing technology from scratch in the laboratory, by employing microfluidic techniques in order to fabricate the microfluidic devices to generate droplets able to capture and isolate single cells. In addition, I optimized an experimental procedure to produce single-cell cDNA libraries and I improved the Drop-seq microfluidic device to increase the single-cell capture efficiency.

The Drop-seq technology enabled me to perform single-cell transcriptional profiling of 35,276 cells from 31 breast cancer cell lines, and 1 non-tumorigenic breast cell line, including all relevant breast cancer subtypes (i.e. luminal A and B, HER2 positive and TNBC). With single-cell transcriptomic, I was able to demonstrate that it is possible to successfully measure the expression of clinically relevant biomarkers, and that these are heterogeneously expressed across cells within the same cell line.

Indeed, an important achievement of my thesis work, is the observation that gene expression intrapopulation heterogeneity not only is present and it is dynamic. To further investigate the intrapopulation heterogeneity, I focused on the MDA-MB-361 cell line, for which approximately the 70% of cells express the HER2 receptor. I carried out the separation of the $HER2^{+}$ subpopulation from the $HER2^{-}$ by means of the fluorescence-activated cell sorting, and I separately cultured these homogenous subpopulations. I assessed by flow cytometry, that both subpopulations re-established the HER2 expression heterogeneity of the initial population after a period of approximately three weeks in culture.

This surprising result, highlights that there exists a dynamic plasticity in the regulation of HER2 expression in the MDA-MB-361 cell line, a phenomenon recently observed also in circulating tumour cells (CTC) of a breast cancer patient (Jordan et al., 2016). Moreover, this observation excluded the possibility that this mechanism is driven by genetic mechanisms. I performed the cell cycle analysis of the MDA-MB-361 cell line to evaluate a possible implication of cell cycle in the HER2 state. However, I found that cell cycle status only partially explains the observed results.

Another key result of my work is the demonstration that drug target expression heterogeneity in a cell line is correlated with drug response. By combining publicly available drug sensitivity data from large-scale in vitro drug screening with breast single-cell dataset, I was able to observe negative correlation between the percentage of cells in the same cell line that express the drug target (e.g. HER2) and the sensitivity of the cell line to the specific drug inhibitor.

This observation led us to develop DREEP, a computation algorithm that predicts the effect of 450 anticancer drugs in single cells. Thanks to DREEP, I was able to show that a common chemotherapeutic agent, etoposide, is able to specifically target HER2- cells in the MDA-MB-61 cell lines, and it may be thus used in conjunction to HER2 inhibitors to improve the effectiveness of these drugs and to prevent a resistant subpopulation to arise during the treatment. Future work will be needed to establish whether this observation is true also in more relevant clinical model of HER2+ breast cancer, such as breast cancer organoids and mouse PDX models.

To conclude, my thesis work shows the importance of performing single-cell RNA sequencing on the available cancer models, including cell lines and organoids to build a set of known cell cancer states with known phenotypes and drug response to which patients' tumour can be mapped to.

References

[1] N. Harbeck *et al.*, “Breast cancer.,” *Nat. Rev. Dis. Prim.*, vol. 5, no. 1, p. 66, 2019, doi: 10.1038/s41572-019-0111-2.

[2] X. Dai, H. Cheng, Z. Bai, and J. Li, “Breast cancer cell line classification and Its relevance with breast tumor subtyping,” *J. Cancer*, vol. 8, no. 16, pp. 3131–3141, 2017, doi: 10.7150/jca.18457.

[3] M. C. U. Cheang *et al.*, “Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer.,” *J. Natl. Cancer Inst.*, vol. 101, no. 10, pp. 736–50, May 2009, doi: 10.1093/jnci/djp082.

[4] X. Dai, L. Xiang, T. Li, and Z. Bai, “Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes.,” *J. Cancer*, vol. 7, no. 10, pp. 1281–94, 2016, doi: 10.7150/jca.13141.

[5] X. Dai, A. Chen, and Z. Bai, “Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling.,” *Sci. Rep.*, vol. 4, p. 6566, Oct. 2014, doi: 10.1038/srep06566.

[6] M. Lacroix *et al.*, “Gene regulation by phorbol 12-myristate 13-acetate in MCF-7 and MDA-MB-231, two breast cancer cell lines exhibiting highly different phenotypes.,” *Oncol. Rep.*, vol. 12, no. 4, pp. 701–7, Oct. 2004, doi: 10.3892/or.12.4.701.

[7] C. M. Perou *et al.*, “Molecular portraits of human breast tumours.,” *Nature*, vol. 406, no. 6797, pp. 747–52, Aug. 2000, doi: 10.1038/35021093.

[8] G. M. Bernardo *et al.*, “FOXA1 represses the molecular phenotype of basal breast cancer cells.,” *Oncogene*, vol. 32, no. 5, pp. 554–63, Jan. 2013, doi: 10.1038/onc.2012.62.

[9] V. Theodorou, R. Stark, S. Menon, and J. S. Carroll, “GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility.,” *Genome Res.*, vol. 23, no. 1, pp. 12–22, Jan. 2013, doi: 10.1101/gr.139469.112.

[10] T. Sorlie *et al.*, “Repeated observation of breast tumor subtypes in independent gene expression data sets.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 14, pp. 8418–23, Jul. 2003, doi: 10.1073/pnas.0932692100.

[11] T. Sørlie *et al.*, “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 19, pp. 10869–74, Sep. 2001, doi: 10.1073/pnas.191367098.

[12] T.-A. Moo, R. Sanford, C. Dang, and M. Morrow, “Overview of Breast Cancer Therapy.,” *PET Clin.*, vol. 13, no. 3, pp. 339–354, Jul. 2018, doi: 10.1016/j.cpet.2018.02.006.

[13] C. S. Vallejos *et al.*, “Breast cancer classification according to immunohistochemistry markers: subtypes and association with clinicopathologic variables in a peruvian hospital database.,” *Clin. Breast Cancer*, vol. 10, no. 4, pp. 294–300, Aug. 2010, doi: 10.3816/CBC.2010.n.038.

[14] J. D. Brenton, L. A. Carey, A. A. Ahmed, and C. Caldas, “Molecular classification and molecular forecasting of breast cancer: ready for clinical application?,” *J. Clin. Oncol.*, vol. 23, no. 29, pp. 7350–60, Oct. 2005, doi: 10.1200/JCO.2005.03.3845.

[15] C. Sotiriou *et al.*, “Breast cancer classification and prognosis based on gene expression profiles from a population-based study.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 18, pp. 10393–8, Sep. 2003, doi: 10.1073/pnas.1732912100.

[16] D. Cameron *et al.*, “11 years’ follow-up of trastuzumab after adjuvant chemotherapy in

HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial.," *Lancet (London, England)*, vol. 389, no. 10075, pp. 1195–1205, 2017, doi: 10.1016/S0140-6736(16)32616-2.
[17] K. Köninki *et al.*, "Multiple molecular mechanisms underlying trastuzumab and lapatinib resistance in JIMT-1 breast cancer cells.," *Cancer Lett.*, vol. 294, no. 2, pp. 211–9, Aug. 2010, doi: 10.1016/j.canlet.2010.02.002.
[18] Y. Nagata *et al.*, "PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients.," *Cancer Cell*, vol. 6, no. 2, pp. 117–27, Aug. 2004, doi: 10.1016/j.ccr.2004.06.022.
[19] X. Dai *et al.*, "Breast cancer intrinsic subtype classification, clinical use and future trends.," *Am. J. Cancer Res.*, vol. 5, no. 10, pp. 2929–43, 2015, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26693050.
[20] K. Aysola *et al.*, "Triple Negative Breast Cancer - An Overview.," *Hered. Genet. Curr. Res.*, vol. 2013, no. Suppl 2, 2013, doi: 10.4172/2161-1041.S2-001.
[21] R. M. Neve *et al.*, "A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.," *Cancer Cell*, vol. 10, no. 6, pp. 515–27, Dec. 2006, doi: 10.1016/j.ccr.2006.10.008.
[22] E. Charafe-Jauffret *et al.*, "Gene expression profiling of breast cell lines identifies potential new basal markers.," *Oncogene*, vol. 25, no. 15, pp. 2273–84, Apr. 2006, doi: 10.1038/sj.onc.1209254.
[23] B. Čunderlíková, "Extracellular matrix in gene expression profiling of cancer," *Transl. Cancer Res.*, vol. 5, no. 1, 2015, [Online]. Available: https://tcr.amegroups.com/article/view/5649.
[24] A. R. Tan, Ed., *Triple-Negative Breast Cancer*. Cham: Springer International Publishing, 2018.
[25] R. Costa *et al.*, "Targeting Epidermal Growth Factor Receptor in triple negative breast cancer: New discoveries and practical insights for drug development.," *Cancer Treat. Rev.*, vol. 53, pp. 111–119, Feb. 2017, doi: 10.1016/j.ctrv.2016.12.010.
[26] D. Furrer, F. Sanschagrin, S. Jacob, and C. Diorio, "Advantages and disadvantages of technologies for HER2 testing in breast cancer specimens.," *Am. J. Clin. Pathol.*, vol. 144, no. 5, pp. 686–703, Nov. 2015, doi: 10.1309/AJCPT41TCBUEVDQC.
[27] A.-E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, "Single-cell RNA-seq: advances and future challenges.," *Nucleic Acids Res.*, vol. 42, no. 14, pp. 8845–60, Aug. 2014, doi: 10.1093/nar/gku555.
[28] F. Iorio *et al.*, "A Landscape of Pharmacogenomic Interactions in Cancer.," *Cell*, vol. 166, no. 3, pp. 740–754, Jul. 2016, doi: 10.1016/j.cell.2016.06.017.
[29] M. G. Rees *et al.*, "Correlating chemical sensitivity and basal gene expression reveals mechanism of action.," *Nat. Chem. Biol.*, vol. 12, no. 2, pp. 109–16, Feb. 2016, doi: 10.1038/nchembio.1986.
[30] J. C. Costello *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms.," *Nat. Biotechnol.*, vol. 32, no. 12, pp. 1202–12, Dec. 2014, doi: 10.1038/nbt.2877.
[31] R. Fisher, L. Pusztai, and C. Swanton, "Cancer heterogeneity: implications for targeted therapeutics.," *Br. J. Cancer*, vol. 108, no. 3, pp. 479–85, Feb. 2013, doi: 10.1038/bjc.2012.581.
[32] F. Andre *et al.*, "Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy

for Women With Early-Stage Invasive Breast Cancer: ASCO Clinical Practice Guideline Update-Integration of Results From TAILORx.," *J. Clin. Oncol.*, vol. 37, no. 22, pp. 1956–1964, 2019, doi: 10.1200/JCO.19.00945.

[33] M. J. Garnett *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells.," *Nature*, vol. 483, no. 7391, pp. 570–5, Mar. 2012, doi: 10.1038/nature11005.

[34] W. D. Foulkes, I. E. Smith, and J. S. Reis-Filho, "Triple-negative breast cancer.," *N. Engl. J. Med.*, vol. 363, no. 20, pp. 1938–48, Nov. 2010, doi: 10.1056/NEJMra1001389.

[35] S. V Sharma *et al.*, "A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations.," *Cell*, vol. 141, no. 1, pp. 69–80, Apr. 2010, doi: 10.1016/j.cell.2010.02.027.

[36] S. M. Shaffer *et al.*, "Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance.," *Nature*, vol. 546, no. 7658, pp. 431–435, 2017, doi: 10.1038/nature22794.

[37] S. Ebinger *et al.*, "Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia.," *Cancer Cell*, vol. 30, no. 6, pp. 849–862, Dec. 2016, doi: 10.1016/j.ccell.2016.11.002.

[38] A. S. Meyer and L. M. Heiser, "Systems biology approaches to measure and model phenotypic heterogeneity in cancer.," *Curr. Opin. Syst. Biol.*, vol. 17, pp. 35–40, Oct. 2019, doi: 10.1016/j.coisb.2019.09.002.

[39] J. E. Wiedmeier, P. Noel, W. Lin, D. D. Von Hoff, and H. Han, "Single-Cell Sequencing in Precision Medicine.," *Cancer Treat. Res.*, vol. 178, pp. 237–252, 2019, doi: 10.1007/978-3-030-16391-4_9.

[40] V. Svensson *et al.*, "Power analysis of single-cell RNA-sequencing experiments.," *Nat. Methods*, vol. 14, no. 4, pp. 381–387, Apr. 2017, doi: 10.1038/nmeth.4220.

[41] C. Ziegenhain, B. Vieth, S. Parekh, I. Hellmann, and W. Enard, "Quantitative single-cell transcriptomics.," *Brief. Funct. Genomics*, vol. 17, no. 4, pp. 220–232, 2018, doi: 10.1093/bfgp/ely009.

[42] S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg, "Smart-seq2 for sensitive full-length transcriptome profiling in single cells.," *Nat. Methods*, vol. 10, no. 11, pp. 1096–8, Nov. 2013, doi: 10.1038/nmeth.2639.

[43] T. M. Gierahn *et al.*, "Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput.," *Nat. Methods*, vol. 14, no. 4, pp. 395–398, Apr. 2017, doi: 10.1038/nmeth.4179.

[44] H. Gong, D. Do, and R. Ramakrishnan, "Single-Cell mRNA-Seq Using the Fluidigm C1 System and Integrated Fluidics Circuits.," *Methods Mol. Biol.*, vol. 1783, pp. 193–207, 2018, doi: 10.1007/978-1-4939-7834-2_10.

[45] P. See, J. Lum, J. Chen, and F. Ginhoux, "A Single-Cell Sequencing Guide for Immunologists.," *Front. Immunol.*, vol. 9, p. 2425, 2018, doi: 10.3389/fimmu.2018.02425.

[46] E. Z. Macosko *et al.*, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.," *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015, doi: 10.1016/j.cell.2015.05.002.

[47] A. M. Klein *et al.*, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.," *Cell*, vol. 161, no. 5, pp. 1187–1201, May 2015, doi: 10.1016/j.cell.2015.04.044.

[48] Illumina, "https://www.illumina.com/science/technology/next-generation-

sequencing.html," [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing.html.
[49] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, "Overview of Next-Generation Sequencing Technologies.," *Curr. Protoc. Mol. Biol.*, vol. 122, no. 1, p. e59, 2018, doi: 10.1002/cpmb.59.
[50] D. R. Link, S. L. Anna, D. A. Weitz, and H. A. Stone, "Geometrically mediated breakup of drops in microfluidic devices.," *Phys. Rev. Lett.*, vol. 92, no. 5, p. 054503, Feb. 2004, doi: 10.1103/PhysRevLett.92.054503.
[51] N. Shembekar, C. Chaipan, R. Utharala, and C. A. Merten, "Droplet-based microfluidics in drug discovery, transcriptomics and high-throughput molecular genetics.," *Lab Chip*, vol. 16, no. 8, pp. 1314–31, Apr. 2016, doi: 10.1039/c6lc00249h.
[52] M. DE MENECH, P. GARSTECKI, F. JOUSSE, and H. A. STONE, "Transition from squeezing to dripping in a microfluidic T-shaped junction," *J. Fluid Mech.*, vol. 595, pp. 141–161, Jan. 2008, doi: 10.1017/S002211200700910X.
[53] D. T. Papageorgiou, "On the breakup of viscous liquid threads," *Phys. Fluids*, vol. 7, no. 7, pp. 1529–1544, Jul. 1995, doi: 10.1063/1.868540.
[54] H.-S. Moon *et al.*, "Inertial-ordering-assisted droplet microfluidics for high-throughput single-cell RNA-sequencing.," *Lab Chip*, vol. 18, no. 5, pp. 775–784, 2018, doi: 10.1039/c7lc01284e.
[55] M. Nooranidoost, M. Haghshenas, M. Muradoglu, and R. Kumar, "Cell encapsulation modes in a flow-focusing microchannel: effects of shell fluid viscosity," *Microfluid. Nanofluidics*, vol. 23, no. 3, p. 31, Mar. 2019, doi: 10.1007/s10404-019-2196-z.
[56] D. C. Duffy, J. C. McDonald, O. J. A. Schueller, and G. M. Whitesides, "Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane)," *Anal. Chem.*, vol. 70, no. 23, pp. 4974–4984, Dec. 1998, doi: 10.1021/ac980656z.
[57] N. R. Beer *et al.*, "On-chip single-copy real-time reverse-transcription PCR in isolated picoliter droplets.," *Anal. Chem.*, vol. 80, no. 6, pp. 1854–8, Mar. 2008, doi: 10.1021/ac800048k.
[58] T. Kawakatsu, G. Trägårdh, C. Trägårdh, M. Nakajima, N. Oda, and T. Yonemoto, "The effect of the hydrophobicity of microchannels and components in water and oil phases on droplet formation in microchannel water-in-oil emulsification," *Colloids Surfaces A Physicochem. Eng. Asp.*, vol. 179, no. 1, pp. 29–37, Apr. 2001, doi: 10.1016/S0927-7757(00)00498-2.
[59] J. Clausell-Tormos *et al.*, "Droplet-based microfluidic platforms for the encapsulation and screening of Mammalian cells and multicellular organisms.," *Chem. Biol.*, vol. 15, no. 5, pp. 427–37, May 2008, doi: 10.1016/j.chembiol.2008.04.004.
[60] D. J. Collins, A. Neild, A. DeMello, A.-Q. Liu, and Y. Ai, "The Poisson distribution and beyond: methods for microfluidic droplet production and single cell encapsulation," *Lab Chip*, vol. 15, no. 17, pp. 3439–3459, 2015, doi: 10.1039/C5LC00614G.
[61] "Dropletex." https://dropletex.com/tools-and-resources/.
[62] X. Zhang *et al.*, "Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems.," *Mol. Cell*, vol. 73, no. 1, pp. 130-142.e5, 2019, doi: 10.1016/j.molcel.2018.10.020.
[63] R. Salomon *et al.*, "Droplet-based single cell RNAseq tools: a practical guide," *Lab Chip*, vol. 19, no. 10, pp. 1706–1727, 2019, doi: 10.1039/C8LC01239C.
[64] S. Slovin *et al.*, "Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview.,"

Methods Mol. Biol., vol. 2284, pp. 343–365, 2021, doi: 10.1007/978-1-0716-1307-8_19.
[65] S. Yang *et al.*, "Decontamination of ambient RNA in single-cell RNA-seq with DecontX," *Genome Biol.*, vol. 21, no. 1, p. 57, Dec. 2020, doi: 10.1186/s13059-020-1950-6.
[66] T. Smith, A. Heger, and I. Sudbery, "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy," *Genome Res.*, vol. 27, no. 3, pp. 491–499, Mar. 2017, doi: 10.1101/gr.209601.116.
[67] G. Gambardella and D. di Bernardo, "A Tool for Visualization and Analysis of Single-Cell RNA-Seq Data Based on Text Mining," *Front. Genet.*, vol. 10, Aug. 2019, doi: 10.3389/fgene.2019.00734.
[68] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.
[69] N. A. Soliman and S. M. Yussif, "Ki-67 as a prognostic marker according to breast cancer molecular subtype.," *Cancer Biol. Med.*, vol. 13, no. 4, pp. 496–504, Dec. 2016, doi: 10.20892/j.issn.2095-3941.2016.0066.
[70] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Feb. 2018.
[71] M. Karaayvaz *et al.*, "Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq," *Nat. Commun.*, vol. 9, no. 1, p. 3588, Dec. 2018, doi: 10.1038/s41467-018-06052-0.
[72] S. L. Kong, G. Li, S. L. Loh, W.-K. Sung, and E. T. Liu, "Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state.," *Mol. Syst. Biol.*, vol. 7, p. 526, Aug. 2011, doi: 10.1038/msb.2011.59.
[73] J. S. Carroll *et al.*, "Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1.," *Cell*, vol. 122, no. 1, pp. 33–43, Jul. 2005, doi: 10.1016/j.cell.2005.05.008.
[74] I. Wolf, S. Bose, E. A. Williamson, C. W. Miller, B. Y. Karlan, and H. P. Koeffler, "FOXA1: Growth inhibitor and a favorable prognostic factor in human breast cancer.," *Int. J. cancer*, vol. 120, no. 5, pp. 1013–22, Mar. 2007, doi: 10.1002/ijc.22389.
[75] H. O. Habashy *et al.*, "Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance.," *Eur. J. Cancer*, vol. 44, no. 11, pp. 1541–51, Jul. 2008, doi: 10.1016/j.ejca.2008.04.020.
[76] O. De Wever *et al.*, "Molecular and pathological signatures of epithelial-mesenchymal transitions at the cancer invasion front.," *Histochem. Cell Biol.*, vol. 130, no. 3, pp. 481–94, Sep. 2008, doi: 10.1007/s00418-008-0464-1.
[77] D. Li *et al.*, "Prognostic values of SNAI family members in breast cancer patients.," *Ann. Transl. Med.*, vol. 8, no. 15, p. 922, Aug. 2020, doi: 10.21037/atm-20-681.
[78] Y. Soini *et al.*, "Transcription factors zeb1, twist and snai1 in breast carcinoma.," *BMC Cancer*, vol. 11, p. 73, Feb. 2011, doi: 10.1186/1471-2407-11-73.
[79] K. M. Mrozik, O. W. Blaschuk, C. M. Cheong, A. C. W. Zannettino, and K. Vandyke, "N-cadherin in cancer metastasis, its emerging role in haematological malignancies and potential as a therapeutic target in cancer.," *BMC Cancer*, vol. 18, no. 1, p. 939, Oct. 2018, doi: 10.1186/s12885-018-4845-0.
[80] C. Ricciardelli *et al.*, "Keratin 5 overexpression is associated with serous ovarian cancer recurrence and chemotherapy resistance.," *Oncotarget*, vol. 8, no. 11, pp. 17819–17832, Mar. 2017, doi: 10.18632/oncotarget.14867.

[81] B. Martin-Castillo *et al.*, “Basal/HER2 breast carcinomas: integrating molecular taxonomy with cancer stem cell dynamics to predict primary resistance to trastuzumab (Herceptin).,” *Cell Cycle*, vol. 12, no. 2, pp. 225–45, Jan. 2013, doi: 10.4161/cc.23274.
[82] N. A. O’Brien *et al.*, “Activated phosphoinositide 3-kinase/AKT signaling confers resistance to trastuzumab but not lapatinib.,” *Mol. Cancer Ther.*, vol. 9, no. 6, pp. 1489–502, Jun. 2010, doi: 10.1158/1535-7163.MCT-09-1171.
[83] B. Martin-Castillo *et al.*, “Cytokeratin 5/6 fingerprinting in HER2-positive tumors identifies a poor prognosis and trastuzumab-resistant basal-HER2 subtype of breast cancer.,” *Oncotarget*, vol. 6, no. 9, pp. 7104–22, Mar. 2015, doi: 10.18632/oncotarget.3106.
[84] S. D. Axlund *et al.*, “Progesterone-inducible cytokeratin 5-positive cells in luminal breast cancer exhibit progenitor properties.,” *Horm. Cancer*, vol. 4, no. 1, pp. 36–49, Feb. 2013, doi: 10.1007/s12672-012-0127-5.
[85] M. F. Sweeney, C. Sonnenschein, and A. M. Soto, “Characterization of MCF-12A cell phenotype, response to estrogens, and growth in 3D.,” *Cancer Cell Int.*, vol. 18, p. 43, 2018, doi: 10.1186/s12935-018-0534-y.
[86] M. Bärlund *et al.*, “Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer,” *Genes Chromosom. Cancer*, vol. 35, no. 4, pp. 311–317, 2002, doi: 10.1002/gcc.10121.
[87] F. S. Al Joudi, “Human mammaglobin in breast cancer: a brief review of its clinical utility,” *Indian J Med Res*, 2014.
[88] B. K. Zehentner and D. Carter, “Mammaglobin: a candidate diagnostic marker for breast cancer.,” *Clin. Biochem.*, vol. 37, no. 4, pp. 249–57, Apr. 2004, doi: 10.1016/j.clinbiochem.2003.11.005.
[89] M. C. U. Cheang *et al.*, “Defining Breast Cancer Intrinsic Subtypes by Quantitative Receptor Expression,” *Oncologist*, vol. 20, no. 5, pp. 474–482, May 2015, doi: 10.1634/theoncologist.2014-0372.
[90] Y.-Y. Qiang *et al.*, “Along with its favorable prognostic role, CLCA2 inhibits growth and metastasis of nasopharyngeal carcinoma cells via inhibition of FAK/ERK signaling,” *J. Exp. Clin. Cancer Res.*, vol. 37, no. 1, p. 34, Dec. 2018, doi: 10.1186/s13046-018-0692-8.
[91] A. D. Gruber and B. U. Pauli, “Tumorigenicity of human breast cancer is associated with loss of the Ca2+-activated chloride channel CLCA2.,” *Cancer Res.*, vol. 59, no. 21, pp. 5488–91, Nov. 1999, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10554024.
[92] D. Zhao and J. T. Dong, “Upregulation of long non-coding RNA DRAIC correlates with adverse features of breast cancer,” *Non-coding RNA*, vol. 4, no. 4, pp. 1–9, 2018, doi: 10.3390/ncrna4040039.
[93] M. Zafrakas *et al.*, “Expression analysis of mammaglobin A (SCGB2A2) and lipophilin B (SCGB1D2) in more than 300 human tumors and matching normal tissues reveals their co-expression in gynecologic malignancies.,” *BMC Cancer*, vol. 6, p. 88, Apr. 2006, doi: 10.1186/1471-2407-6-88.
[94] M. A. Watson and T. P. Fleming, “Mammaglobin, a mammary-specific member of the uteroglobin gene family, is overexpressed in human breast cancer.,” *Cancer Res.*, vol. 56, no. 4, pp. 860–5, Feb. 1996, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8631025.
[95] E. Sasaki, N. Tsunoda, Y. Hatanaka, N. Mori, H. Iwata, and Y. Yatabe, “Breast-specific

expression of MGB1/mammaglobin: an examination of 480 tumors from various organs and clinicopathological analysis of MGB1-positive breast cancers.," *Mod. Pathol.*, vol. 20, no. 2, pp. 208–14, Feb. 2007, doi: 10.1038/modpathol.3800731.

[96] G. H. Lewis *et al.*, "Relationship between molecular subtype of invasive breast carcinoma and expression of gross cystic disease fluid protein 15 and mammaglobin.," *Am. J. Clin. Pathol.*, vol. 135, no. 4, pp. 587–91, Apr. 2011, doi: 10.1309/AJCPMFR6OA8ICHNH.

[97] M. J. Núñez-Villar *et al.*, "Elevated mammaglobin (h-MAM) expression in breast cancer is associated with clinical and biological features defining a less aggressive tumour phenotype.," *Breast Cancer Res.*, vol. 5, no. 3, pp. R65-70, 2003, doi: 10.1186/bcr587.

[98] Z. Wang *et al.*, "Identification of KLK10 as a therapeutic target to reverse trastuzumab resistance in breast cancer.," *Oncotarget*, vol. 7, no. 48, pp. 79494–79502, Nov. 2016, doi: 10.18632/oncotarget.13104.

[99] L.-Y. Luo, E. P. Diamandis, M. P. Look, A. P. Soosaipillai, and J. A. Foekens, "Higher expression of human kallikrein 10 in breast cancer tissue predicts tamoxifen resistance," *Br. J. Cancer*, vol. 86, no. 11, pp. 1790–1796, 2002, doi: 10.1038/sj.bjc.6600323.

[100] V. Dugina, G. Shagieva, N. Khromova, and P. Kopnin, "Divergent impact of actin isoforms on cell cycle regulation," *Cell Cycle*, vol. 17, no. 23, pp. 2610–2621, Dec. 2018, doi: 10.1080/15384101.2018.1553337.

[101] X. Lu *et al.*, "Establishment of a Predictive Genetic Model for Estimating Chemotherapy Sensitivity of Colorectal Cancer with Synchronous Liver Metastasis," *Cancer Biother. Radiopharm.*, vol. 28, no. 7, pp. 552–558, Sep. 2013, doi: 10.1089/cbr.2012.1431.

[102] K. Edfeldt, P. Hellman, G. Westin, and P. Stalberg, "A plausible role for actin gamma smooth muscle 2 (ACTG2) in small intestinal neuroendocrine tumorigenesis," *BMC Endocr. Disord.*, vol. 16, no. 1, p. 19, Dec. 2016, doi: 10.1186/s12902-016-0100-3.

[103] C.-Z. Xu *et al.*, "Gene and microRNA expression reveals sensitivity to paclitaxel in laryngeal cancer cell line.," *Int. J. Clin. Exp. Pathol.*, vol. 6, no. 7, pp. 1351–61, 2013, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23826416.

[104] N. M. Verrills *et al.*, "Alterations in γ-Actin and Tubulin-Targeted Drug Resistance in Childhood Leukemia," *JNCI J. Natl. Cancer Inst.*, vol. 98, no. 19, pp. 1363–1374, Oct. 2006, doi: 10.1093/jnci/djj372.

[105] R. Gao *et al.*, "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes," *Nat. Biotechnol.*, vol. 39, no. 5, pp. 599–608, May 2021, doi: 10.1038/s41587-020-00795-2.

[106] "10x Genomics." .

[107] B. Jew *et al.*, "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information," *Nat. Commun.*, vol. 11, no. 1, p. 1971, 2020, doi: 10.1038/s41467-020-15816-6.

[108] M. Tanner *et al.*, "Characterization of a novel cell line established from a patient with Herceptin-resistant breast cancer," *Mol. Cancer Ther.*, vol. 3, no. 12, pp. 1585 LP – 1592, Dec. 2004.

[109] M. Ghandi *et al.*, "Next-generation characterization of the Cancer Cell Line Encyclopedia," *Nature*, vol. 569, no. 7757, pp. 503–508, May 2019, doi: 10.1038/s41586-019-1186-3.

[110] D. Ishay-Ronen *et al.*, "Gain Fat—Lose Metastasis: Converting Invasive Breast Cancer Cells into Adipocytes Inhibits Cancer Metastasis," *Cancer Cell*, vol. 35, no. 1, pp. 17-32.e6, Jan. 2019, doi: 10.1016/j.ccell.2018.12.002.

[111] S. Ingthorsson, K. Andersen, B. Hilmarsdottir, G. M. Maelandsmo, M. K. Magnusson, and T. Gudjonsson, "HER2 induced EMT and tumorigenicity in breast epithelial progenitor cells is inhibited by coexpression of EGFR," *Oncogene*, vol. 35, no. 32, pp. 4244–4255, Aug. 2016, doi: 10.1038/onc.2015.489.
[112] K. Dökümcü and R. M. Farahani, "Evolution of Resistance in Cancer: A Cell Cycle Perspective.," *Front. Oncol.*, vol. 9, p. 376, 2019, doi: 10.3389/fonc.2019.00376.
[113] G. H. Williams and K. Stoeber, "The cell cycle and cancer.," *J. Pathol.*, vol. 226, no. 2, pp. 352–64, Jan. 2012, doi: 10.1002/path.3022.
[114] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nat. Biotechnol.*, vol. 36, no. 5, pp. 411–420, May 2018, doi: 10.1038/nbt.4096.
[115] Y. Yan *et al.*, "A novel function of HER2/Neu in the activation of G2/M checkpoint in response to γ-irradiation," *Oncogene*, vol. 34, no. 17, pp. 2215–2226, Apr. 2015, doi: 10.1038/onc.2014.167.
[116] A. L. Blajeski, V. A. Phan, T. J. Kottke, and S. H. Kaufmann, "G(1) and G(2) cell-cycle arrest following microtubule depolymerization in human breast cancer cells.," *J. Clin. Invest.*, vol. 110, no. 1, pp. 91–9, Jul. 2002, doi: 10.1172/JCI13275.
[117] M. A. Jordan, D. Thrower, and L. Wilson, "Effects of vinblastine, podophyllotoxin and nocodazole on mitotic spindles. Implications for the role of microtubule dynamics in mitosis.," *J. Cell Sci.*, vol. 102 (Pt 3, pp. 401–16, Jul. 1992, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1506423.
[118] M. Kuhn, "The microtubule depolymerizing drugs nocodazole and colchicine inhibit the uptake of Listeria monocytogenes by P388D1 macrophages.," *FEMS Microbiol. Lett.*, vol. 160, no. 1, pp. 87–90, Mar. 1998, doi: 10.1111/j.1574-6968.1998.tb12895.x.
[119] A. B. Pardee, "A restriction point for control of normal animal cell proliferation.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 71, no. 4, pp. 1286–90, Apr. 1974, doi: 10.1073/pnas.71.4.1286.
[120] P. H. O'Farrell, "Quiescence: early evolutionary origins and universality do not imply uniformity.," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 366, no. 1584, pp. 3498–507, Dec. 2011, doi: 10.1098/rstb.2011.0079.
[121] T. Ly, A. Endo, and A. I. Lamond, "Proteomic analysis of the response to cell cycle arrests in human myeloid leukemia cells.," *Elife*, vol. 4, Jan. 2015, doi: 10.7554/eLife.04534.
[122] N. Pozdeyev, M. Yoo, R. Mackie, R. E. Schweppe, A. C. Tan, and B. R. Haugen, "Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies.," *Oncotarget*, vol. 7, no. 32, pp. 51619–51625, Aug. 2016, doi: 10.18632/oncotarget.10010.
[123] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data.," *Pac. Symp. Biocomput.*, pp. 63–74, 2014, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24297534.
[124] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.
[125] W. Yang *et al.*, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D955–D961, Nov. 2012, doi: 10.1093/nar/gks1111.

[126] N. V. Jordan *et al.*, “HER2 expression identifies dynamic functional states within circulating breast cancer cells,” *Nature*, vol. 537, no. 7618, pp. 102–106, Sep. 2016, doi: 10.1038/nature19328.
[127] M. Hafner, M. Niepel, M. Chung, and P. K. Sorger, “Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs.,” *Nat. Methods*, vol. 13, no. 6, pp. 521–7, 2016, doi: 10.1038/nmeth.3853.

APPENDIX A - Materials and Methods

Cell culture

We obtained the 32 cell lines used in this study from commercial providers. I cultured cell lines in ATCC recommended complete media at 37°C and 5% CO2.

Drop-seq platform set-up

I performed the single-cell transcriptomic of the 32 cell lines with Drop-seq platform. I fabricated the microfluidics device for the generation of droplet using a bio-compatible, silicon-based polymer, polydimethylsiloxane (PDMS) that was rendered hydrophobic with Aquapel® treatment. In each sequencing experiment, I loaded cell suspension, barcoded bead suspension and carrier oil (QX200 droplet generation oil, Bio-Rad) in syringes and then I placed them in syringe pumps (Leafluid). Flow rates of syringe pumps were set at 4,000 µL/hr for both cell and barcoded bead suspensions while carrier oil syringe pump was set at 15,000 µL/hr. In the human-mouse mixture experiment, I resuspended cells at the concentration of 500 cell/µL (250 cell/µL in the droplet final volume) and barcoded beads at the concentration of 120 bead/µL. For CCL sequencing experiments, I diluted cells at the concentration of 200 cell/µL in PBS with BSA 0.01% (Merck) and 120 bead/µL in lysis buffer. I used the self-built magnetic stirrer system to keep in suspension barcoded beads. I performed tests to count the occurrence of a single cell together with a barcoded bead without lysis buffer in the barcoded bead suspension. To break the water-in-oil emulsion, I collected droplets in a 50 mL falcon and I added 1 mL of Perfluoro-1-octanol. Captured RNA was reverse transcribed in a single reaction following the original protocol described in and then digested with exonuclease 1 to degrade unbound primers.

Cell clustering and identification of marker genes

We found transcriptionally similar subpopulations of cells using a Phenograph like approach as implemented in the *clustcells* function of *gficf* package. Briefly, we initially built a graph of cells by using the K-Nearest Neighbours (KNN) algorithm applied on the PC-reduced space where each cell was connected to its 50 most similar cells using the manhattan distance. Then, to build the final graph of cells edge weight between any two cells was refined with Jaccard similarity by computing the proportion of neighbours they share. We used the Louvain algorithm with resolution parameter equal to 0.25 to find communities of cells in this graph of cells. We identified differentially expressed genes in each cluster by using *findClusterMarkers* function of *gficf* package that compare the expression of a gene in each cluster versus all the other by using the Wilcoxon rank-sum test.

TCGA BC bulk expression dataset and deconvolution into BC cell-lines

We collected raw BC bulk expression data and relative patient clinical information from the Genomic Data Commons (GDC) portal by using the *TCGAbiolinks* package. Then, we normalized raw counts using the *EdgeR* package into R statistical environment 3.6. We used Bisque tool (available at https://github.com/cozygene/bisque) to estimate the cell-line proportion in BC TCGA bulk expression. Specifically, we used the *ReferenceBasedDecomposition* function with parameter *bulk.eset* equal to the BC TCGA expression dataset in log2 scale, the parameter sc.eset equal to our BC atlas were normalized raw counts were rescaled in log2, *use.overlap* parameter equal to FALSE and *markers* parameter equal to the marker genes across the 32 BC cell-lines estimated by using the function *findClusterMarkers* function of *gficf* package. As in the original manuscript of Bisque tool [107], only marker genes with an FDR<0.5 and Log2 fold change greaten then 0.25 were used for deconvolution purpose. Before to be used both

Spatial sequencing data

We downloaded spatial transcriptomic data of BC patient from 10x Genomic website (https://www.10xgenomics.com/resources/datasets). We used only tiles reported to be "in tissue" according to the related metadata of each patient slide.

Embed new cells into the BC atlas and prediction of the cancer type

We embedded new points into the UMAP space via *embedNewCells* function of *gficf* package. Briefly, we normalized tiles from 10x spatial transcriptomic with GF-ICF method as described above but using the ICF weight estimated on the BC atlas. Then, tiles are projected in the existing PC space using gene loadings estimated on the BC atlas. After this transformation, we embedded tiles in the BC atlas via *umap_transform* function of *uwot* package. Finally, we predicted the cancer type of each new embedded point using the function *classify.cells* of the package *gficf* with the k parameter equal to 7. This function performs a k-nearest neighbour classifier to classify the new embedded points using the coordinates of the UMAP space.

Single-cell drug sensitivity prediction

We obtained the basal expression profile of about 1,000 cancer cell line from RNA-sequencing data downloaded from the Cancer Cell Line Encyclopaedia (CCLE) portal. We discarded cell line belonging to liquid tumour and we retained only 658 cell lines belonging to solid tumours for further analysis. We normalized the raw counts of each gene with edgeR package and transformed in log10(CPM+1). Lowly expressed genes and genes whose entropy was in the fifth percentile were excluded from the analysis. We crossed the expression profiles of the 658 CCLs with drug sensitivity data from work of Rees and colleges [29]. This dataset was originally composed by 481 small molecules, but after removing drugs for which the in vitro response was available for more than 25 CCLs only 450 small molecules were retained for further analysis. Then, we computed for each gene and for each of the 450 drugs the Pearson correlation coefficient (PCC) between the expression of the gene and the effect of the drug expressed in terms of Area Under the Curve (AUC) across the 658 cell lines. Since the AUC reflects the in vitro response of a cell line to different concertation of a drug in a timeframe of 72 hours, lower values of AUC are associated with sensitivity whereas higher values with resistance to the drug by the tested cell line. Hence, genes positively correlated with the AUC are potential markers of resistance (the more expressed the gene, the higher the concentration needed to inhibit growth), vice-versa, negatively correlated genes are markers of sensitivity. With this approach, we generated a ranked list of expression-based biomarkers of drug sensitivity and resistance for each of the 450 drugs where genes positively correlated with the AUC are at the top, and those negatively correlated at the bottom. Finally, to predict drug sensitivity at the single-cell level, we used the top 250 expressed genes of each cell as input of Gene Set Enrichment Analysis (GSEA) against the ranked list of biomarkers for each one of 450 drugs built as described above. Hence, while a negative enrichment score implies that genes associated to drug sensitivity are highly expressed by the cell, a positive one indicates the cell express genes conferring drug resistance. GSEA and associated p-values were estimating using the *fgsea* package in the R statistical environment version 3.6.

Differential drug sensitivity prediction between HER2+ and HER2- cells in the MDA-MB-361 cell line

For each sequenced cell of the MDA-MB-361, we predicted the effect to 450 anticancer drugs as described above. Then, for each of the 450 drugs, we used the Mann-Whitney test to assess if there was a difference between enrichment scores of HER2+ (UMI>0) and HER2- cells. Obtained pvalues were corrected for false discovery rate using

Benjamini-Hochberg correction. We considered a drug specific for HER2- cell population if and only if its FDR was less than 0.05 and the median enrichment score across HER2- cells less than zero while its median enrichment score across HER2+ cells greater than zero. Conversely, we considered a drug specific for HER2+ cell population if and only if FDR was less than 0.05 and the median enrichment score across HER2+ cells less than zero while its median enrichment score across HER2- cells greater than zero.

Validation of drug sensitivity prediction

Precision of the proposed method in predicting drug sensitivity was evaluated using an independent drug screening dataset produced from Iorio and colleagues [9] composed by 1,001 CCLs and their maximal inhibitory concentration (IC50) values at 265 small molecules. Hence, we used the described method on our 32 BC cell lines to predict from single cell transcriptional data the percentage of sensitive cells at 86 common drugs between the two datasets. The "golden standard" was built by assigning to each of 32 x 86 (=2,752) cell line/drug pair the value 1 if the cell line was sensitive to the drug and 0 otherwise. To determine if a cell line was sensitive or not to a specific drug, we converted for each drug its IC50 distribution in Z-scores using all the 1,001 available cell lines and then defined a cell line sensitive to the drug if and only if its Z-score was in the 5% percentile. Finally, Positive Predicted Values (PPV) were defined as TP/(TP+FP) where TP represents the number of true positives and FP the number of false positives predicted cell lines/drug pairs.

Prediction of cell cycle phase from scRNA-seq

We predicted the cell cycle phase of each sequenced cell using the function *CellCycleScoring* of the *Seurat* tool with default parameter and we followed what was suggested in the corresponding vignette (https://satijalab.org/seurat).

HER2 antibody staining procedure for flow cytometry analysis

Before staining, I first washed **c**ells with phosphate-buffered saline (PBS) 1x, detached with 0.05% trypsin-EDTA, resuspended and harvested with the appropriate medium in single-cell suspension. Then, I counted cells, washed with PBS-FBS 1%, and finally incubated for 15 min at 4° in the dark at the concentration of 1.0×10^6 cell/mL with staining buffer. I prepared the staining buffer by diluting the mouse anti-human HER2 antibody (BD BB700) at the final concentration of 0.00114 ng/mL. Then, to remove unbound antibody, I washed cells three times with PBS-FBS 1%. I performed flow cytometry measurements on either BD Accuri C6 or BD FACSAria III instruments. To define antibody positive and negative cells, I used the unstained samples to set the threshold.

HER2 expression dynamics experiment

I performed the sorting of MDA-MB-361 HER2-positive and HER2-negative cells following the antibody staining procedure that I described above with the only exception that before sorting, I resuspended each sample in sorting buffer (PBS 1x, FBS 1%, trypsin 0.1%, EDTA 2mM). Then, I collected 4.0×10^5 cells for each cell subpopulation (*i.e.* HER2-positive and HER2-negative); I seeded cells in their appropriate medium, and incubated at 37°. After 18 days, I checked the percentage of cells expressing HER2 protein by performing the antibody staining procedure described above.

Drug sensitivity assay

I seeded **c**ells in the appropriate format (96-well microplates (PerkinElmer)); I specifically optimized the seeding cell confluency for each cancer cell line to have cells in growth phase at the end of the assay. After overnight incubation at 37°, I treated cells with DMSO (Merck) for the negative control and with selected drugs in triplicate, depending on the assay, as well as for the incubation time at 37°. I assessed cell viability by measuring either luminescence with GloMax® Discover instrument from Promega or by nuclei count using the Operetta instrument from PerkinElmer. I normalized luminescence measurements using background wells as manufacturer protocol. For luminescence measurement, I treated cells with Promega CellTiter-Glo® Luminescent Cell Viability Assay according to the manufacturer protocol. For nuclei count, I washed attached cells with PBS 1x, fixed with paraformaldehyde (PFA) 4% for 10 min at room temperature, washed again with PBS 1x, incubated at room temperature in the dark with HOECHST 33342 (Thermo Fisher Scientific) diluted 1:1000 in PBS 1x for 10 min and finally I washed with PBS 1x. I performed nuclei count by using the Columbus image analysis software (PerkinElmer). All drugs I used in this study were purchased from Selleckchem.

APPENDIX B – Mathematical model of the HER2 interconversion dynamics

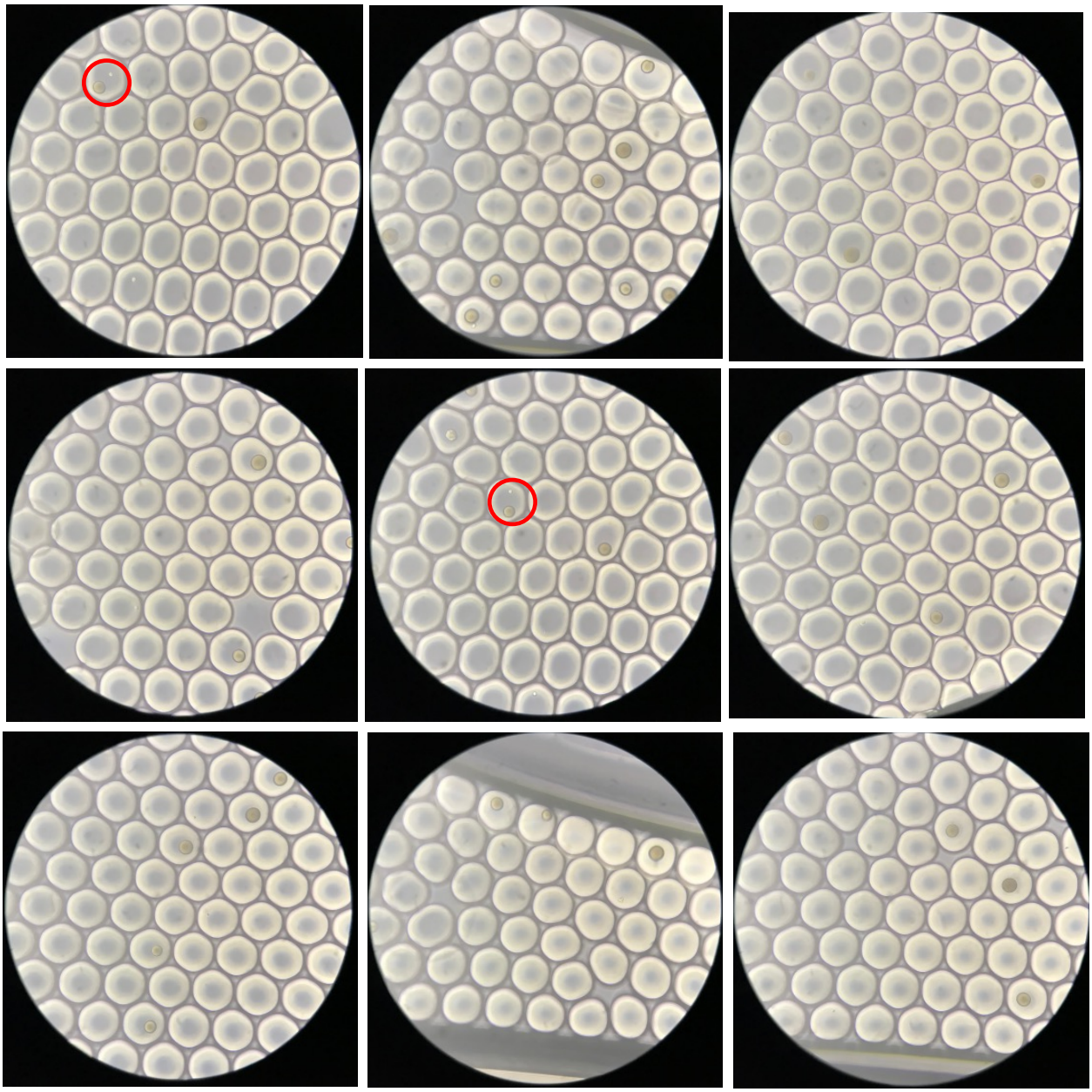
The model assumes that each cell can be in either one of two states (HER2$^-$ and HER2$^+$) and can switch dynamically between the two with rates λ, δ. Moreover, independently of the state, the cell can replicate with rates k_1 and k_2. Finally, the effect of anti-HER2 drugs is present as an additional degradation term on HER2$^+$ cells.
Using standard mass action kinetics, the following equations describing the model in Figure 6.7 can be derived:

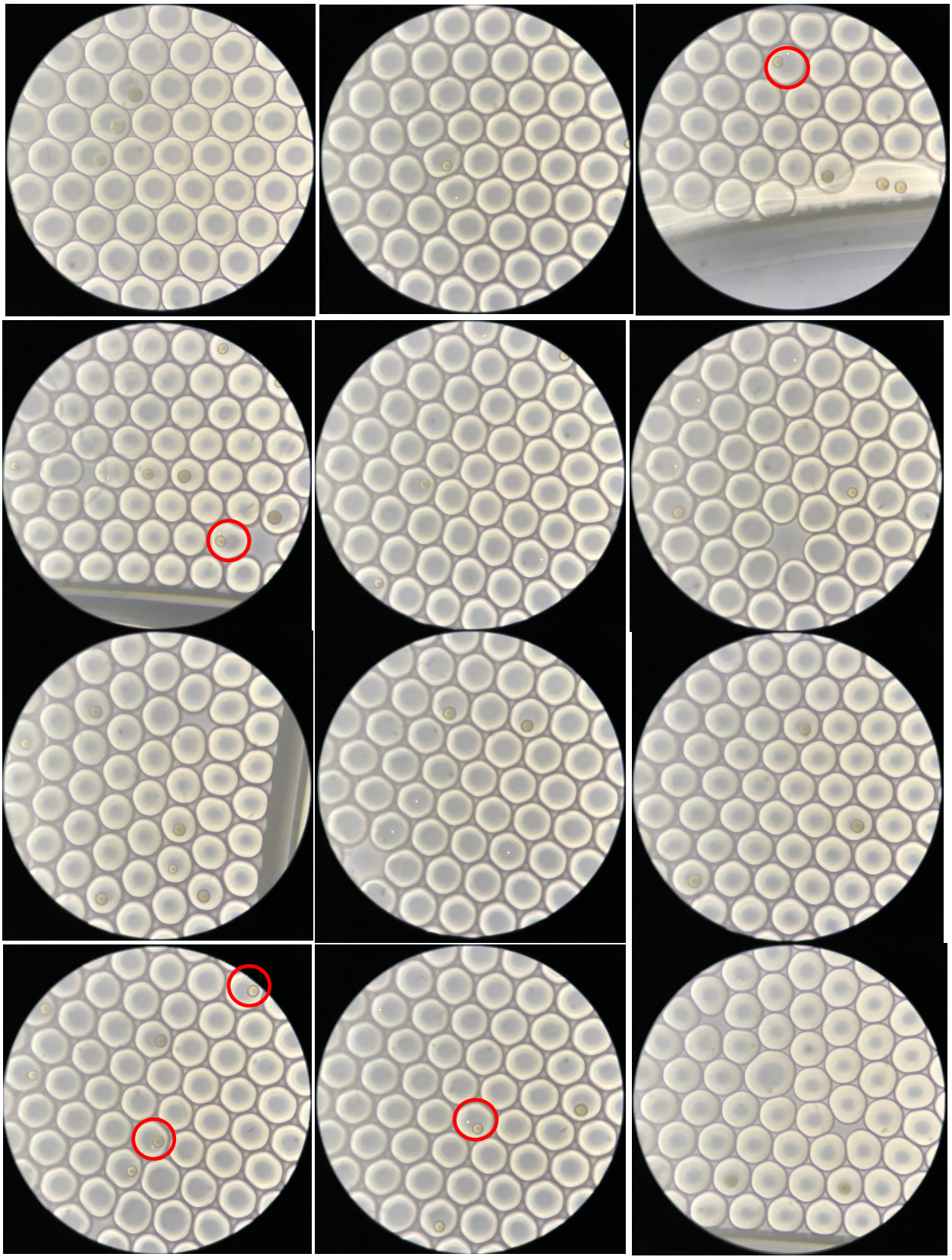
$$\begin{cases} \dot{h^-} = (k_1 - \lambda)h^- + \delta h^+ \\ \dot{h^+} = (k_2 - \delta)h^+ + \lambda h^- - uh^+ \end{cases}$$

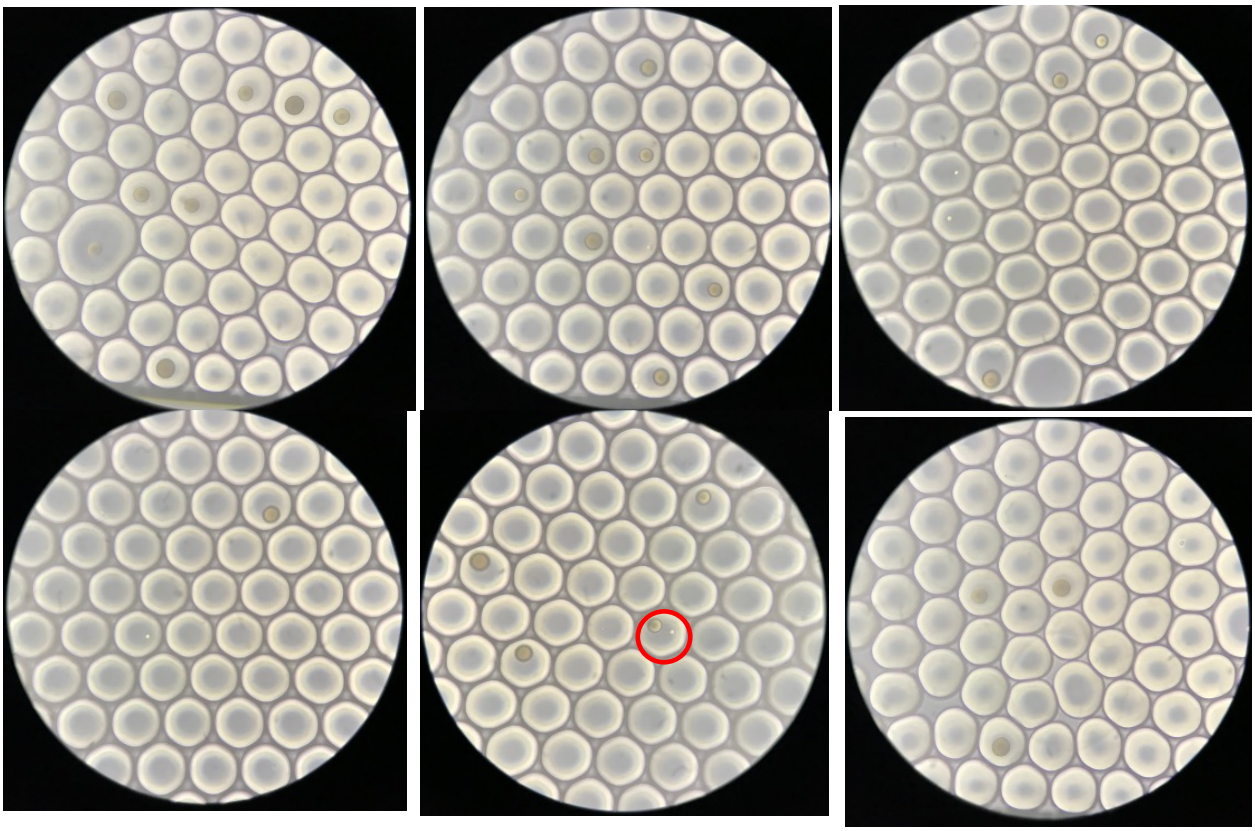
Where h^- stands for HER2$^-$ cells, and h^+ for HER2$^+$ cells, whereas u quantifies the effect of anti-HER2 drugs (e.g., Afatinib). To simplify the model, the replication rates k1 and k2 are assumed to be the same. The parameters values for λ, δ determine the percentage of HER2$^+$ cells in the cell population, which can be shown to be equal to $\frac{\lambda}{\lambda+\delta}$ after a transient, when no drug is present (u=0). The parameters' values, reported in Supplementary Figure 13, were set to yield a doubling rate of the total cell population ($h^- + h^+$) of approx. 3.5 days, like the observed cell cycle rate of the MDAMB361 cell line, and a percentage of HER2$^+$ cells of 60%, close to the value tha we measured (Figure 5.1). With these nominal values, the replication rates and the interconversion rates are of the same order of magnitude.
Figure 6.A shows the numerical simulations of the model behaviour following treatment with afatinib, with a starting population of 1x10^6 cells, of which 0.9x10^6 are HER2$^+$ cells and 0.1x10^6 HER2$^-$ cells (i.e. 90% HER2$^+$ cells): for nominal values of the parameters, in the absence of Afatinib, both HER2$^-$ and HER2$^+$ cells grow exponentially, while the percentage of Her2+ cells quickly stabilises at 60%; upon Afatinib treatment for 3 days, both the number of HER2$^+$ and HER2$^-$ cells decrease, while the percentage of HER2$^+$ cells settles at 30%. Finally, upon removal of Afatinib, the number of cells increases while the percentage of HER2$^+$ cells recover to 60%. When the interconversion rates (λ, δ) are much slower than the growth rate (k), then in the absence of Afatinib the percentage of HER2$^+$ cells take longer to stabilise at 60%, whereas the effect of a 3 days Afatinib is much more pronounced, causing the percentage of HER2 cells to quickly drop to approx. 10%. This can be explained by the fact that HER2$^-$ cells keep increasing in number during Afatinib treatment as their growth rate is much faster than their interconversion rate, while HER2$^+$ are removed by Afatinib treatment and cannot escape its effect as they convert to HER2$^-$ cells too slowly. Upon Afatinib removal, both the number of cells and the percentage of HER2$^+$ cells start increasing. For fast interconversion rates, the situation is reversed, that is cells increase in number in the absence of Afatinib with the percentage of HER2$^+$ cells quickly reaching 60%. Interestingly, while the effect of Afatinib is almost absent in terms of changes in the percentage of HER2$^+$ cells, the total number of cells drops substantially, as HER2$^-$ are much more affected by Afatinib treatment because of their fast interconversion to HER2$^+$ cells.
We also simulated dose response curves at increasing concentrations of drugs (i.e., the value of u in the model) for the model for each set of parameters' values (slow, nominal and fast), as reported in Figure 6.8B. As expected, only in the case of slow interconversion, it is possible to appreciate a difference in the response of HER2$^-$ cells versus HER2$^+$ cells following treatment with Afatinib.

APPENDIX C – Supplementary Figures

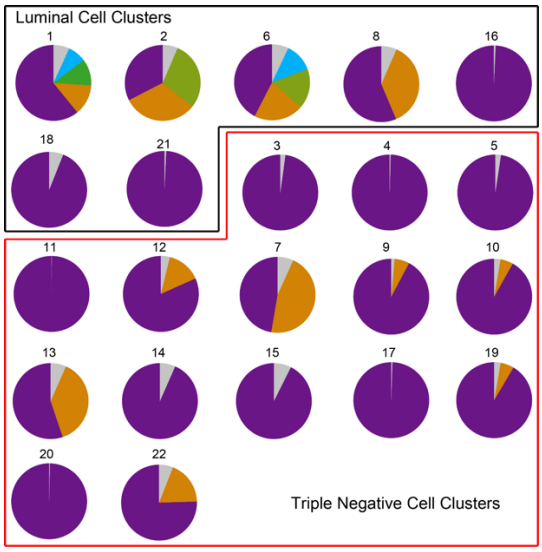
Supp. Figure C1 - Fields I captured to count the cell and barcoded bead occupancy (red circles) for the experiment described in Chapter 3, Section 3.4.2.



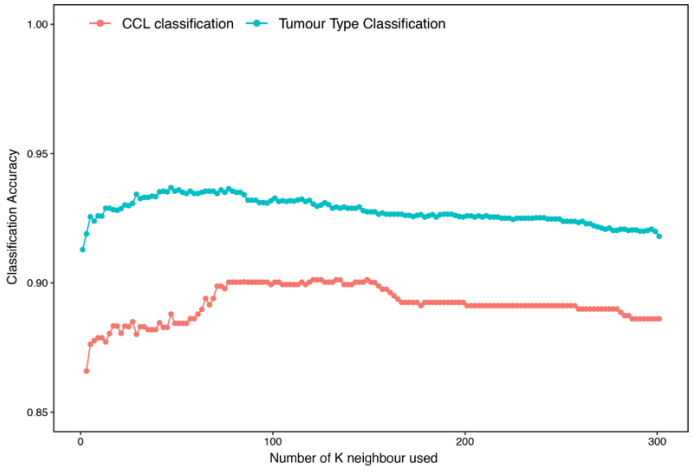




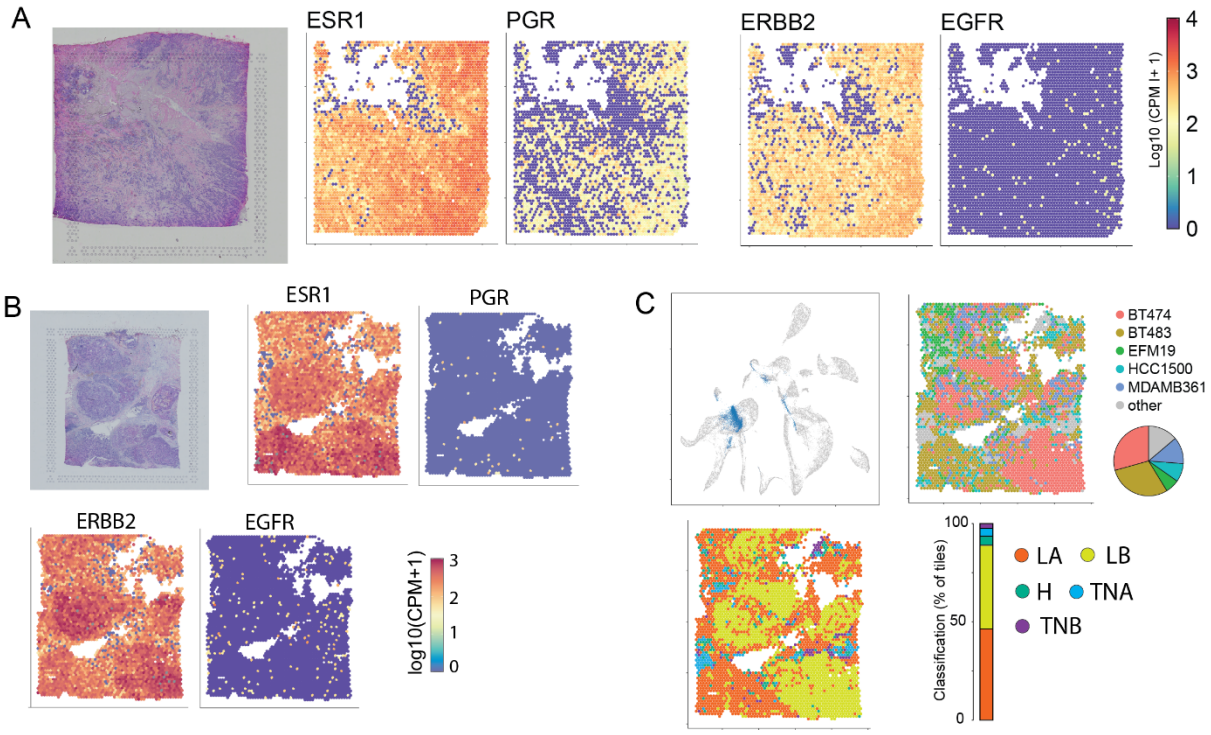
Supp. Figure C2 - Composition of the clusters in the Atlas. For the indicated cluster, the corresponding pie-chart represents the cluster composition in terms of cell lines. Cell-lines in the same pie-chart are distinguished by colour.



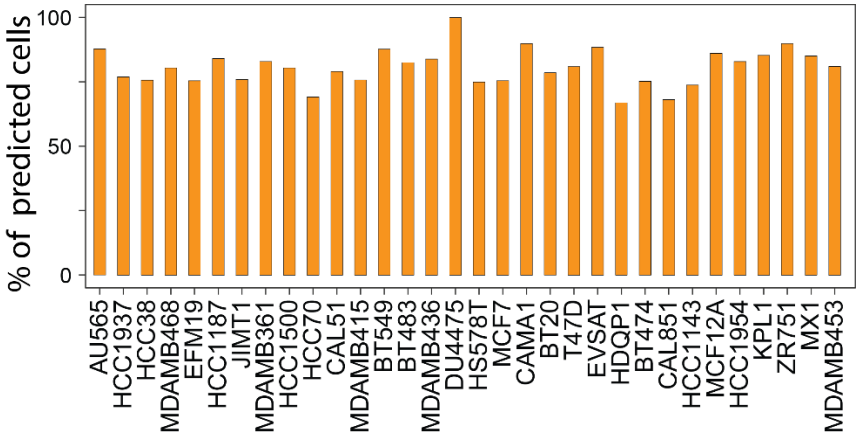
Supp. Figure C3 - Single cell cancer type classification performances. Seventy five percent of cells of each cell-line were collected and used as the training set while the remaining 25% was used as test set. Cells in the training set were used to reconstruct the breast cancer atlas from scratch while the cell line type of each cell in the test set was predicted by mapping them into the atlas as "new cells". Finally average classification accuracy as a function of the number of neighbourhood cells was estimated by using the function classify.cells of the package gficf (Methods). The analyses described in the main text were performed using a number of neighbours K=100.



Supp. Figure C4 - Automatic detection of cell line composition of spatial transcriptomics profiles. (A) Tissue-slide of the lobular BC tumour biopsy (presented in Figure 2C) sequenced by 10x spatial transcriptomics and the spatial expression of ESR1, PGR, ERBB2 and EGFR genes. (B) Tissue-slide of a ductal BC tumour biopsy sequenced by 10x spatial transcriptomics and the spatial expression of ESR1, PGR, ERBB2 and EGFR genes. (C) Top-left: Cancer cells sequenced with spatial 10X genomics technology are embedded in the BC atlas to predict which cell-line they are similar using K-nn algorithm. Top-right: Classification of each pseudo cell to show predicted cell-line in the spatial context with the pie-chart showing the percentage of cells predicted to be similar a specific cell-line. Bottom-left: Classification of each pseudo cell to show predicted tumour type in the spatial context. Bottom-right: Quantification of bottom-left plot where the percentage of pseudo cells predicted for each tumour type is reported.



Supp. Figure C5 - Performance in predicting cell line composition from bulk RNA-seq using computational deconvolution. The Bisque deconvolution algorithm was first trained on the BC single-cell atlas and then its performances estimated by predicting cell-line composition from bulk RNA-seq obtained by averaging single cell expression profiles for each cell-line.



APPENDIX D – Supplementary Tables

Supp. Table D1 – Cancer cell lines information table. ER=ESR1= estrogen receptor; PR=PgR=progesterone receptor; H= HER2+; LA = Luminal A; LB = Luminal B; TNA = triple negative A; TNB = triple negative B. In each column is reported the (A) cell line name; (B-D) expression marker status; (E) BRCA mutational status; (F) subtype; (G) culture condition; (H) derivation site of the cell line.

CCLs	ER	PR	HER2	BRCA1	Subtype	Growth.medium	derivation_site
AU565	-	-	+	WT	H	RPMI 2mM L-Glutamine, 10 %FBS	metastatic site: malignant pleural effusion
BT20	-	-	-	WT	TNA	MEM 2mM L-Glutamine, 10 %FBS	mammary gland/breast
BT474	+	+	+	WT	LB	DMEM 2mM L-Glutamine, 10 %FBS	mammary gland; breast/duct
BT483	+	+/-	-	WT	LA	RPMI 2mM L-Glutamine, 0.01 mg/mL bovine insulin, 20 %FBS	mammary gland; breast
BT549	-	-	-	WT	TNB	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland; breast
CAL51	-	-	-	WT	TNB	DMEM 2mM L-Glutamine, 10 %FBS	metastatic site: pleural effusion
CAL851	-	-	-	WT	TNB	DMEM 2mM L-Glutamine, 10 %FBS, 1mM sodium pyruvate	relapsing invasive galactophoric breast adenocarcinoma
CAMA1	+	+/-	-	WT	LA	MEM 2mM L-Glutamine, 10 %FBS, 1% NEAA	metastatic site: pleural effusion
DU4475	-	-	-	WT	TNA	RPMI 2mM L-Glutamine, 20 %FBS	metastatic site: skin carcinoma
EFM19	+	+	-	ND	LA	RPMI 2mM L-Glutamine, 10 %FBS	metastatic site: pleural effusion
EVSAT	-	+/-	+	ND	H	MEM 2mM L-Glutamine, 10 %FBS, 25 mM HEPES	metastatic site: malignant ascitic effusion
HCC1143	-	-	-	ND	TNA	RPMI 2mM L-Glutamine, 20 %FBS	mammary gland; breast/duct
HCC1187	-	-	-	ND	TNA	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland; breast
HCC1500	+	+/-	-	ND	LA	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland; breast/duct
HCC1937	-	-	-	MU	TNA	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland; breast/duct
HCC1954	-	-	+	WT	H	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland; breast/duct
HCC38	-	-	-	ND	TNB	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland; breast/duct
HCC70	-	-	-	WT	TNA	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland; breast/duct
HDQP1	-	-	-	MU	TNB	DMEM 2mM L-Glutamine, 10 %FBS	primary ductal infiltrating breast carcinoma
HS578T	-	-	-	WT	TNB	RPMI 2mM L-Glutamine, 10 %FBS	mammary gland/breast
JIMT1	-	-	+	ND	H	DMEM 2mM L-Glutamine, 10 %FBS	metastatic site: pleural effusion
KPL1	+	-	-	ND	LA	DMEM 2mM L-Glutamine, 10 %FBS	metastatic site: pleural fluid
MCF12A	-	-	-	ND	Basal-like	DMEM/F12 - 1:1 mixture of DMEM and F-12, 20 ng/mL EGF, 0.03 µM ITS, 10% FBS	mammary gland; non-tumorigenic, fibrocystic breast diseas
MCF7	+	+	-	WT	LA	MEM 2mM L-Glutamine, 10 %FBS, 1% NEAA	metastatic site: pleural effusion
MDAMB175VII	+	-	-	WT	LA	RPMI 2mM L-Glutamine, 10 %FBS	metastatic site: pleural effusion
MDAMB361	+	+/-	+	WT	LB	RPMI 2mM L-Glutamine, 20 %FBS	metastatic site: brain
MDAMB415	+	+/-	-	WT	LA	RPMI 2mM L-Glutamine, 10 %FBS, 10 µg/mL insulin, 10 µg/mL glutathione, 15% FB	metastatic site: pleural effusion
MDAMB436	-	-	-	MU	TNA	RPMI 2mM L-Glutamine, 10 %FBS	metastatic site: pleural effusion
MDAMB453	-	-	+	WT	H	RPMI 2mM L-Glutamine, 10 %FBS	metastatic site: pericardial effusion
MDAMB468	-	-	-	WT	TNA	RPMI 2mM L-Glutamine, 10 %FBS	metastatic site: pleural effusion
MFM223	-	-	-	WT	TNA	MEM 2mM L-Glutamine, 15 %FBS, 0.03 µM ITS	metastatic site: pleural effusion
MX1	-	-	-	WT	TNB	DMEM/F12 - 1:1 mixture of DMEM and F-12, 2 mM L-Glutamine, 10% FBS	Primary xenotransplant from an infiltrating duct carcinoma
T47D	+	+	-	WT	LA	RPMI 2mM L-Glutamine, 10 %FBS, 1 mM sodium pyruvate, 10 mM HEPES	metastatic site: pleural effusion
ZR751	+	+/-	-	WT	LA	RPMI 2mM L-Glutamine, 10 %FBS	metastatic site: ascites

Supp. Table D2 – Drug prediction for HER2+ and HER- cell subpopulations of MDAMB361 cell-line. For each predicted drug we report: (A) its name; (B) its Enrichment Score computed on HER2+ cells; (C) its Enrichment Score computed on ES HER2- cells; (D) for which cells drug is specific; (E) The P-value; (F) Bonferroni corrected P-value; (G) the target gene of the drug; (H) the drug Mode of Action.

drug name	ES HER2+ cells	ES HER2- cells	specificity	P	Bonferroni	drug.MoA_to_plot
SN-38	0.073062055	-0.035355036	HER2- specific	7.72507E-08	3.47628E-05	inhibitor of topoisomerase I or II
3-Cl-AHPC	0.06962767	-0.057546287	HER2- specific	2.71798E-07	0.000122309	binder of nuclear receptor SHP
PHA-793887	0.064467955	-0.051961343	HER2- specific	9.97543E-07	0.000448895	inhibitor of cyclin-dependent kinases
NVP-231	0.063645982	-0.040608342	HER2- specific	9.19924E-08	4.13966E-05	inhibitor of ceramide kinase
BRD-K66453893	0.063348932	-0.042213632	HER2- specific	7.65066E-08	3.4428E-05	product of diversity oriented synthesis
paclitaxel	0.061798576	-0.045749746	HER2- specific	3.14917E-07	0.000141713	inhibitor of microtubule assembly
teniposide	0.061586979	-0.0442706	HER2- specific	1.49555E-07	6.72997E-05	inhibitor of topoisomerase I or II
daporinad	0.060561546	-0.037169888	HER2- specific	1.32572E-07	5.96575E-05	inhibitor of nicotinamide phosphoribosyltransferase
manumycin A	0.06	-0.04064293	HER2- specific	4.46806E-07	0.000201063	inhibitor of RAS farnesyltransferase
narciclasine	0.058661241	-0.055418108	HER2- specific	1.19042E-07	5.35691E-05	modulator of RhoA activity
SU11274	0.057989827	-0.052156663	HER2- specific	6.05375E-08	2.72419E-05	inhibitor of MET
FQI-1	0.057957274	-0.037995931	HER2- specific	8.85168E-08	3.98326E-05	inhibitor of LSF1-mediated transcription
rigosertib	0.057489318	-0.043774161	HER2- specific	4.0148E-07	0.000180666	inhibitor of polo-like kinase 1 (PLK1)
foretinib	0.057355036	-0.03597762	HER2- specific	3.79328E-08	1.70698E-05	inhibitor of MET
BI-2536	0.057188199	-0.054844354	HER2- specific	8.62256E-07	0.000388015	inhibitor of polo-like kinase 1 (PLK1)
parbendazole	0.056899288	-0.047291963	HER2- specific	5.44768E-07	0.000245146	inhibitor of microtubule assembly
MK-1775	0.056883011	-0.047615463	HER2- specific	5.32131E-08	2.39459E-05	inhibitor of WEE1
phloretin	0.056862665	-0.051513733	HER2- specific	1.69027E-06	0.000760623	natural product; inhibitor of glucose uptake
PF-3758309	0.054526958	-0.059586979	HER2- specific	6.87567E-07	0.000309405	inhibitor of PAK4
etoposide	0.05436826	-0.047304171	HER2- specific	6.21627E-07	0.000279732	inhibitor of topoisomerase I or II
gemcitabine	0.053367243	-0.053814852	HER2- specific	7.5843E-08	3.41294E-05	inhibitor of DNA replication
doxorubicin	0.05160529	-0.037558494	HER2- specific	1.148E-07	5.16602E-05	inhibitor of topoisomerase I or II
CAY10618	0.05002645	-0.049348932	HER2- specific	8.58282E-08	3.86227E-05	inhibitor of nicotinamide phosphoribosyltransferase
BMS-195614	0.049945066	-0.035916582	HER2- specific	4.13E-06	0.001858501	antagonist of retinoic acid receptors
CHM-1	0.049749746	-0.054520855	HER2- specific	5.7486E-08	2.58687E-05	inhibitor of tubulin polymerization
BRD-K70511574	0.048695829	-0.054382503	HER2- specific	4.49669E-07	0.000202351	inhibitor of polo-like kinase 1 (PLK1)
Ki8751	0.048419125	-0.046417091	HER2- specific	4.08498E-08	1.83824E-05	inhibitor of VEGFRs
axitinib	0.047344863	-0.046883011	HER2- specific	3.29792E-07	0.000148406	inhibitor of VEGFRs
vincristine	0.047291963	-0.057700916	HER2- specific	3.76864E-07	0.000169589	inhibitor of mictrotubule assembly
BMS-536924	0.046933876	-0.041304171	HER2- specific	1.64887E-08	7.41991E-06	inhibitor of IGF1R and INSR
SB-225002	0.04563174	-0.046024415	HER2- specific	1.21144E-06	0.000545149	inhibitor of chemokine receptor 2
FQI-2	0.045375381	-0.053513733	HER2- specific	7.24966E-07	0.000326235	inhibitor of LSF1-mediated transcription
sorafenib	0.045257375	-0.049167854	HER2- specific	1.36865E-06	0.000615891	inhibitor of BRAF, CRAF, and VEGFR2
ML311	0.044459817	-0.060640895	HER2- specific	2.77148E-07	0.000124716	inhibitor of MCL1
olaparib	0.044056968	-0.04802645	HER2- specific	3.90592E-07	0.000175767	inhibitor of poly (ADP-ribose) polymerase 1 and 2
tigecycline	0.043491353	-0.042317396	HER2- specific	9.05721E-06	0.004075744	analog of tetracycline
cediranib	0.042107833	-0.041540183	HER2- specific	1.95335E-06	0.000879006	inhibitor of VEGFRs
MG-132	0.041009156	-0.046217701	HER2- specific	1.34252E-06	0.000604134	inhibitor of the proteosome
topotecan	0.040549339	-0.059053917	HER2- specific	2.1868E-06	0.000984061	inhibitor of topoisomerase I or II
lomeguatrib	0.035316378	-0.051096643	HER2- specific	6.8368E-08	3.07656E-05	inhibitor of methylguanine-DNA methyltransferase
PF-573228	0.032520855	-0.053894201	HER2- specific	2.98217E-06	0.001341976	inhibitor of focal adhesion kinase
nintedanib	0.023088505	-0.052821974	HER2- specific	2.17368E-06	0.000978156	inhibitor of c-KIT, VEGFRs, PDGFRs, and FGFRs