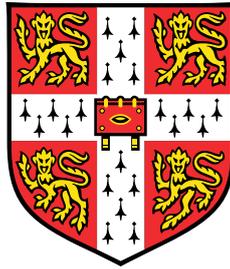


CLONAL DECONVOLUTION OF TRANSCRIPTOMIC  
SIGNATURES AND THEIR SPATIAL ORGANISATION  
IN A MOUSE MODEL OF BREAST CANCER



**Sophia Alisa Wild**

Supervisor: Prof. Gregory J. Hannon

Dr. Kirsty Sawicka

Cancer Research UK Cambridge Institute

Jesus College

University of Cambridge

This dissertation is submitted for the degree of

*Doctor of Philosophy*

November 2021



*“It is not the strongest of the species that survives, not the most intelligent that survives. It is the one that is the most adaptable to change.”*

Charles Darwin



## Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text and authors contributions at the beginning of each Chapter.

This thesis is the result of work carried out at the Cancer Research UK – Cambridge Institute between October 2017 and August 2021. It is not substantially the same as any that I have submitted or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted or is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

This thesis does not exceed the prescribed word limit of 60,000 words for the Clinical Medicine and Veterinary Medicine Degree Committee.

Sophia Alisa Wild  
November 2021



# **Clonal deconvolution of transcriptomic signatures and their spatial organisation in a mouse model of breast cancer**

Sophia Alisa Wild

Intratumour heterogeneity is a phenomenon during cancer progression in which cancer cells diverge and form clonal populations with distinct phenotypic, genetic, or epigenetic states within the same tumour. This intrinsic heterogeneity provides a fuel for cancer evolution enabling tumour cell populations to adapt to selective pressures imposed by the tumour microenvironment or therapeutic interventions. Lineage-tracing approaches shed light into the clonal dynamics of complex populations, but generally lack the ability to directly associate clonal lineage with measurements that infer phenotype such as epigenetics and transcriptomics. In contrast, single-cell sequencing techniques can provide insight into the makeup of complex biological ecosystems, revealing the presence of rare cell populations that are typically masked in bulk analyses, but lack the ability to link these cell populations to clonal lineages.

To address this challenge, we developed the WILDseq platform, a novel approach that allows clonal characterisation at the single-cell transcriptomic level while facilitating the prospective analysis of dynamic regulation of phenotypic heterogeneity under the selective pressure of therapeutic intervention. WILDseq relies on uniquely labelling individual cells with a heritable, expressed DNA barcode coupled with high-throughput single-cell RNA-sequencing. Importantly, this lentiviral-labelling approach can be deployed in any model system that is susceptible to viral transfection. Thus, this platform allows the comprehensive and systematic characterisation of clonal phenotypic differences within complex populations.

Here we demonstrate how this technology can be used to determine clonal populations which are sensitive or resistant to a particular therapeutic intervention, identify transcriptomic signatures that correlate with these phenotypes and analyse how these cells adapt their transcriptomes to escape therapy. We have applied WILDseq to the study of differential clonal responses to chemotherapy in the heterogeneous 4T1 model of breast cancer and validated transcriptomic signatures of therapeutic resistance and sensitivity in primary patient data. We additionally used WILDseq to study the clonal response to the epigenetic regulator JQ1 which revealed intrinsic signatures

that primed clones to JQ1 sensitivity. We observed JQ1-dependent depletion of CD8+ cytotoxic T-cells and suggest that this drives changes in clonal distribution. Finally, we are working on developing a high throughput FISH assay to leverage the WILDseq technology for mapping clonal and transcriptional identities spatially.

Collectively, this thesis contributes to the characterisation and understanding of breast cancer heterogeneity and the impact of clonal architecture on tumour progression and response to therapy.



## Acknowledgements

After 4 years as a PhD and one year as a master student in the Hannon lab and the UK, I want to thank all the people that have contributed to make me the person I am today.

I first would like to acknowledge Greg. I am extremely grateful to have been given the opportunity to work on this exciting, challenging and extremely expensive project. I have learned and grown as a scientist and person beyond my craziest expectations. I need to thank Greg especially for the freedom and challenges he gave me. Thanks for all the highly inspiring and intellectually stimulating conversations.

Two people, that miraculously joined our lab in my second year and played a key role during my PhD, are Kirsty and Ian. Both of you, you made my time truly special here and I would like to thank you for all the support and guidance, for always being there and cheering for me, and for teaching me hell of a lot of things in bioinformatics and cancer biology. WILDseq would not be where it is today without you - it was a great pleasure to be part of our little team. I think we can definitely consider ourselves now as a big family with regard to the numbers of self-isolation we survived together.

Thanks to my labmates, especially my lovely office mate Tati who suffered with me through these last four years and for always drowning PhD sorrows together with a glass of wine. A special shoutout also goes to Ashley for being my cheerleader and keeping me sane with our runs that always end up in sprints. Thanks to everyone in the lab for all the fun, the summer parties and Christmas dinners.

Thanks to all the people that have supported me along my journey throughout these years. Thanks to Joaquina, for introducing me to the Hannon lab and giving me the chance to grow with multiple challenging, interesting projects, and for all the lessons and tricks you taught me. Thanks to all the Core Facilities for all their help and support, in particular from the BRU, Genomics and Flow Cytometry core. Additionally, I would like to thank my thesis committee members, Jason Carroll and Sakari Vanharanta, for their support and the scientific discussions and Ann Kaminski,

our graduate admissions tutor, for always having an open door and taking care of the students.

I would also like to extend my deepest appreciation to all my amazing friends all over the world! Thanks to all my German friends from school and university days for always having my back and coming to visit me despite being on the other side of the pond. Thanks to my friends at Jesus College - Demetris, Kev, Krit, Tamara and Tori - for the countless amounts of workouts, coffees, dinners and all the wine. A special thanks goes to Lisa. You have been my family away from home for the last four years. Thanks for caring, listening, for being there during the highest and lowest moments, for all our amazing trips and the endless amount of prosecco and wine we had during lockdown. I will cherish the memories from all our experiences and cannot wait for many more to come. Finally, thank you to my family who have supported me in this endeavor and for sending lots of care packages.



# Table of contents

List of Abbreviations	xv
List of figures	xvii
List of tables	xix
<b>1 Introduction</b>	<b>1</b>
1.1 Tumour heterogeneity . . . . .	1
1.1.1 Breast cancer tumour heterogeneity . . . . .	2
1.1.2 The 4T1 mouse model . . . . .	4
1.1.3 Tumour heterogeneity as an obstacle for therapeutic response . . . . .	5
1.2 Single-cell RNA sequencing technologies . . . . .	7
1.3 Lineage tracing strategies . . . . .	9
1.3.1 Lineage tracing by barcode-sequencing . . . . .	10
1.3.2 Retrospective approaches . . . . .	11
1.4 Single-cell transcriptomics meets lineage tracing . . . . .	13
1.5 Spatial tumour heterogeneity . . . . .	15
1.5.1 Molecular profiling of tumours in space . . . . .	15
1.5.2 Spatial transcriptomic approaches . . . . .	16
<b>2 WILDseq: Development of an integrative barcoding approach</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.2 Material and Methods . . . . .	21
2.2.1 Cell culture . . . . .	21
2.2.2 WILDseq library design and cloning . . . . .	21
2.2.3 Virus production and transduction . . . . .	22
2.2.4 Library complexity analysis . . . . .	22
2.2.5 Bottlenecking and characterization of WILDseq pools. . . . .	23

2.2.6	Whitelist generation of WILDseq barcodes . . . . .	23
2.2.7	Single cell library preparation . . . . .	24
2.2.8	Enrichment library preparation . . . . .	25
2.2.9	Animals . . . . .	25
2.2.10	Bioinformatic analysis of WILDseq scRNA-seq data . . . . .	25
2.3	Results . . . . .	28
2.3.1	Design of the WILDseq platform and proof-of principle experiment	28
2.3.2	Robust detection of WILDseq barcodes <i>in vivo</i> . . . . .	31
2.3.3	Characterisation of WILDseq pool <i>in vitro</i> . . . . .	34
2.4	Discussion . . . . .	36
<b>3</b>	<b>Investigating therapeutic response using WILDseq in breast cancer</b>	<b>39</b>
3.1	Introduction . . . . .	39
3.2	Material and Methods . . . . .	41
3.2.1	Tissue preparation for scRNA-seq experiments . . . . .	41
3.2.2	Animals and <i>in vivo</i> dosing . . . . .	41
3.2.3	Differential gene expression analysis . . . . .	42
3.2.4	Analysis of baseline transcriptomic signatures . . . . .	42
3.3	Results . . . . .	43
3.3.1	Validation of WILDseq in a the triple-negative breast cancer cell line D2A1-m2 . . . . .	47
3.3.2	Characterising baseline signatures of major clones <i>in vivo</i> . . . . .	48
3.3.3	Defining resistance and sensitivity signatures to docetaxel . . . . .	50
3.3.4	Defining resistance and sensitivity signatures to JQ1 . . . . .	54
3.3.5	Discussion . . . . .	63
<b>4</b>	<b>CloneSTAR: Visualising clonal populations in space</b>	<b>67</b>
4.1	Introduction . . . . .	67
4.2	Material and Methods . . . . .	70
4.2.1	Designing and cloning 40-mer barcodes for CloneSTAR . . . . .	70
4.2.2	Creating barcoded clones for CloneSTAR validation experiments	70
4.2.3	Slide coating for cell and tissue experiments . . . . .	71
4.2.4	Sample preparation . . . . .	71
4.2.5	STARmap procedure for cells and tissue slices . . . . .	71
4.2.6	STARmap imaging and <i>in situ</i> sequencing . . . . .	72

---

4.2.7	STARmap and CloneSTAR probe design . . . . .	72
4.2.8	Single-cell RNA sequencing and marker selection . . . . .	73
4.3	Results . . . . .	75
4.3.1	Design of CloneSTAR vector and detecting of CloneSTAR barcodes <i>in vitro</i> . . . . .	75
4.3.2	Visualising differential gene expression profiles of clones with STARmap . . . . .	78
4.3.3	Establishing STARmap in breast cancer tissue . . . . .	83
4.3.4	Visualisation of clones <i>in vivo</i> . . . . .	87
4.4	Discussion . . . . .	90
<b>5</b>	<b>Conclusion</b>	<b>93</b>
	<b>References</b>	<b>97</b>



# List of Abbreviations

## Roman Symbols

<i>(sc)WES</i>	(single-cell) Whole exome sequencing
<i>(sc)WGS</i>	(single-cell) Whole genome sequencing
<i>AUC</i>	Area under the recovery curve
<i>bp</i>	Base pairs
<i>CBC</i>	Cell barcode
<i>CNV</i>	Copy number variation
<i>CRISPR</i>	Clustered regularly interspaced short palindromic repeats
<i>DCIS</i>	Ductal carcinoma <i>in situ</i>
<i>DDR</i>	DNA damage repair
<i>DMEM</i>	Dulbecco's modified eagle medium
<i>DMSO</i>	Dimethyl sulfoxide
<i>EMT</i>	Epithelial-to-mesenchymal transition
<i>FACS</i>	Fluorescence-activated cell sorting
<i>FISH</i>	Fluorescence <i>in situ</i> hybridisation
<i>GEM</i>	Gel bead in emulsion
<i>GTF</i>	Gene transfer format
<i>IDC</i>	Invasive ductal carcinoma

<i>MHC</i>	Major histocompatibility complex
<i>MMTV</i>	Mouse mammary tumour virus
<i>MOI</i>	Multiplicity of infection
<i>MRD</i>	Minimal residual disease
<i>mRNA</i>	messenger RNA
<i>NHEJ</i>	Non-homologous end joining
<i>PBS</i>	Phosphate-buffered saline
<i>PCA</i>	Principle component analysis
<i>PCAWG</i>	Pan-Cancer Analysis of Whole Genomes
<i>PCR</i>	Polymerase chain reaction
<i>PGK</i>	Phosphoglycerate kinase promoter
<i>polyA</i>	polyadenylated
<i>RT</i>	Reverse transcription
<i>scRNA – seq</i>	single-cell RNA sequencing
<i>SEDAL</i>	Sequencing with error-reduction by dynamic annealing and ligation
<i>SEM</i>	Standard error of mean
<i>smFISH</i>	Single-molecule fluorescence <i>in situ</i> hybridisation
<i>SNV</i>	Single nucleotide variation
<i>TNBC</i>	Triple negative breast cancer
<i>UMI</i>	Unique molecular identifier
<i>UTR</i>	Untranslated region
<i>WILDseq</i>	Wholistic interrogation of lineage dynamics by sequencing

# List of figures

1.1	Tumour heterogeneity . . . . .	2
1.2	Tumour evolution and emergence of therapy resistance through non-genetic mechanisms. . . . .	6
1.3	10X Genomics Platform. . . . .	8
1.4	Prospective lineage tracing approaches. . . . .	11
1.5	Integrative technologies for lineage tracing at single-cell level. . . . .	14
2.1	WILDseq pipeline and barcode design. . . . .	29
2.2	WILDseq barcode detection in 4T1 cells <i>in vitro</i> . . . . .	30
2.3	WILDseq barcode detection <i>in vivo</i> . . . . .	31
2.4	WILDseq barcode detection <i>in vivo</i> and bottlenecking strategy of clone pools . . . . .	33
2.5	Characterisation of 4T1 WILDseq pool <i>in vitro</i> . . . . .	34
3.1	WILDseq pipeline <i>in vivo</i> for docetaxel and JQ1 treatment . . . . .	43
3.2	WILDseq applied to study clonal dynamics under drug treatment <i>in vivo</i> . . . . .	44
3.3	Comparing WILDseq pool <i>in vitro</i> versus <i>in vivo</i> . . . . .	46
3.4	WILDseq characterisation of D2A1 tumours . . . . .	47
3.5	Characterisation of basal transcriptomic signatures of major clones <i>in vivo</i> . . . . .	49
3.6	Characterising clonal dynamics in 4T1 WILDseq tumours to docetaxel treatment . . . . .	51
3.7	Clonal transcriptomic signatures of docetaxel resistance and sensitivity . . . . .	52
3.8	Correlation of differentially expressed genes in clone 679 and clone 238 versus all other clones . . . . .	53
3.9	Transcriptomic signatures of docetaxel resistance and sensitivity in patient treated with taxane-based chemotherapy . . . . .	54
3.10	Characterisation of clonal dynamics in JQ1-treated WILDseq tumours . . . . .	55

---

3.11	Baseline signature of clone 473 primes it for JQ1 sensitivity . . . . .	56
3.12	Global downregulation of <i>MYC</i> targets in JQ1-treated tumour . . . . .	57
3.13	JQ1 treatment results in specific depletion of CD8+ T-cells in breast cancer tumours . . . . .	58
3.14	Global JQ1-dependent gene expression changes . . . . .	60
3.15	JQ1 resistance correlates with $\beta$ -2-microglobulin ( <i>B2m</i> ) expression . . .	61
4.1	Schematic illustration of STARmap protocol . . . . .	68
4.2	CloneSTAR barcode design and testing barcode detection <i>in vitro</i> in 4T1 clones - F, G, J, and T . . . . .	77
4.3	ScRNA-seq of CloneSTAR labelled 4T1 cells <i>in vitro</i> and selection of clonal markers . . . . .	79
4.4	Detection of 16-gene library and CloneSTAR barcodes <i>in vitro</i> . . . . .	80
4.5	CloneSTAR barcode detection aligns with clonal marker genes <i>in vitro</i>	82
4.6	Selection of tumour microenvironment markers . . . . .	84
4.7	Identification of cell types <i>in vivo</i> using STARmap . . . . .	87
4.8	CloneSTAR barcode detection <i>in vivo</i> using scRNA-seq and STARmap	88

# List of tables

2.1	Barcode oligos . . . . .	22
2.2	RT-PCR and gDNA PCR primers . . . . .	24
2.3	Enrichment PCR primers . . . . .	25
4.1	CloneSTAR oligos . . . . .	70
4.2	Tissue pre-treatment conditions tested on 4T1 breast tumour tissue . .	85



# Chapter 1

## Introduction

### 1.1 Tumour heterogeneity

Most human tumours are composed of genetically and phenotypically heterogeneous cancer cell populations that evolve dynamically in space and time following principles of Darwinian evolution. This extensive genetic and phenotypic diversity exists not only between patient tumours (intertumoural heterogeneity), but also within individual tumours (intratumoural heterogeneity) (**Figure 1.1**). Intratumour heterogeneity is a key driver of cancer progression underpinning important emergent features such as drug resistance and metastasis and thus, poses a major clinical challenge. The origins and dynamics of tumour heterogeneity are still poorly understood and different underlying mechanisms have been proposed, including cell-autonomous (e.g. genetic and epigenetic) and non-cell-autonomous (e.g. tumour microenvironment) as well as stochastic events.

Intratumour heterogeneity can manifest in a spatial manner, describing the varying distribution of a heterogeneous cell population within the primary site or across different disease sites, while temporal heterogeneity illustrates the dynamic variation of the tumour cell population over time.

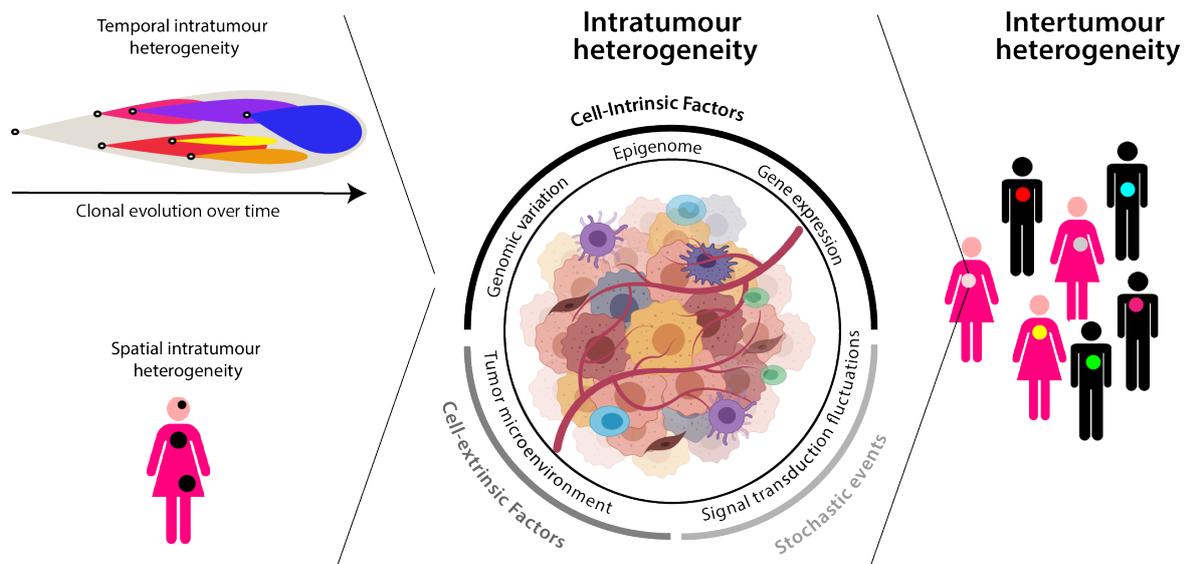


Figure 1.1 **Tumour heterogeneity.** Tumour heterogeneity can be classified into intertumoural heterogeneity, describing the variations in the molecular makeup of tumours between patients, and intratumoural heterogeneity which can be observed within individual tumours. Intratumoural heterogeneity can either occur between different geographical regions of a tumour (spatial) or as molecular evolution of tumours over time (temporal). Inter- as well as intratumoural heterogeneity can complicate diagnosis and challenge therapeutic approaches.

Unravelling tumor heterogeneity has a major clinical impact, as intratumour heterogeneity is a mechanism of therapeutic resistance. The evolution of resistant subpopulations and cellular changes in phenotypes largely affects therapeutic outcome. However, intratumour heterogeneity still remains poorly characterized across cancer types. Several studies have revealed the presence of genetic heterogeneity, the most studied aspect of intratumour heterogeneity, using whole-exome multiregion spatial sequencing or next-generation sequencing mutational profiling (Gerlinger et al., 2012; Shah et al., 2012). In an effort to characterise genetic intratumour heterogeneity in multiple cancer types, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium identified the presence of rich subclonal architectures with linear and branching evolutionary paths as well as transcriptomic alteration signatures which were associated with DNA mutational signatures (Calabrese et al., 2020; Gerstung et al., 2020).

### 1.1.1 Breast cancer tumour heterogeneity

Breast cancer is the most frequently diagnosed cancer and is the second most lethal cancer in women worldwide after lung cancer. Breast cancer comprises a heterogeneous group of distinct subtypes characterised by their genomic, phenotypic and biological

features. Based on the presence of clinical and pathological features, including estrogen and progesterone receptors (ER and PR), and HER2, tumours are classified into ER<sup>+</sup>, HER2<sup>+</sup>, and ER<sup>-</sup>PR<sup>-</sup>HER2<sup>-</sup> (triple-negative) subtypes.

Early transcriptomic profiling using microarrays classified breast cancer tumours into four major intrinsic subtypes (luminal-A, luminal-B, basal-like, and HER2-enriched) and a normal breast-like group that showed significant differences in clinical outcome, incidence and therapy response (Carey et al., 2006; Perou et al., 2000; Sørlie et al., 2001). Molecular characterisation of intrinsic subtypes using the PAM50 score based on the expression of 50 genes, further refined classification of patients into these prognostic groups and allowed prediction of risk of recurrence and treatment response (Nielsen et al., 2010; Parker et al., 2009; Perou et al., 2000). Additional gene-expression analysis further revealed another intrinsic subtype referred to as claudin-low (Prat et al., 2010).

More recently, genomic and transcriptomic analysis of over 2,000 breast tumours has been used to stratify breast tumours into 11 different integrative cluster (IntClust). Each cluster has a distinct pattern of copy number aberrations (CNAs) and gene expression changes. The integrative subtypes each show differences in prognostic values and/or predictive value (Curtis et al., 2012; Nielsen et al., 2010; Pereira et al., 2016; The Cancer Genome Atlas Network, 2012).

Previous work in triple-negative breast cancer (TNBC) patients have shown high levels of somatic mutations, frequent mutations in TP53, and complex genomic rearrangements revealing the presence of extensive intratumour heterogeneity (Gao et al., 2016; Navin et al., 2011; Shah et al., 2012; Wang et al., 2014). Sequencing of TNBC and ER<sup>+</sup> breast cancer has suggested that breast tumours have the ability to undergo major shifts in genomic aberrations upon neoadjuvant therapy (Kim et al., 2018; Yates et al., 2015). Kim *et al.* further showed that while the genotypes were pre-existing and adaptively selected in TNBC patients, transcriptional reprogramming was induced upon treatment (Kim et al., 2018). In a more recent study, Saheli *et al.* generated single-cell whole-genome sequencing from one HER2<sup>+</sup> and three TNBC patient-derived xenografts (PDX) (Salehi et al., 2021). Notably, phylogenetic analysis revealed up to eleven major clones in the TNBC PDXs, while only four clones were observed in the HER2<sup>+</sup> subtype. Drug treatment of TNBC PDXs resulted in a change of clonal dynamics with cisplatin-resistant clones emerging. Based on the findings of

Caswell-Jin *et al.* breast tumours seem to have a higher and more widely variable level of intratumour heterogeneity in untreated as well as treated tumours compared to other cancer types, such as colon, lung or esophagus cancer (Caswell-Jin *et al.*, 2019). The high level of intratumour heterogeneity could be explained through several factors, including high selective pressure during primary tumour growth, differences in the microenvironment, spatial boundaries within the tumour, or different growth modes. Thus, it is of highest priority to gain a better understanding of clonal diversity in breast tumours as this might be critical indicator to how tumour cells will encounter various selective pressures.

### 1.1.2 The 4T1 mouse model

The murine mammary cancer cell line 4T1 is a syngenic mammary carcinoma model for breast cancer and in particular TNBC. This cell line was originally derived from a single spontaneously arising mammary tumour (410.4 tumour) of a mouse mammary tumor virus (MMTV) positive BALB/c mouse foster nursed on a C3H mother (BALB/BfC3H) (Heppner *et al.*, 1978). The 4T1 line was selected from the 410.4 tumour through its 6-thioguanine-resistance without mutagen treatment (Aslakson and Miller, 1992). Injection of a 4T1 cell suspension into the mammary fat-pad results in the formation of 4T1 tumours which are highly tumourigenic, invasive with the ability to spontaneously metastasize from the primary tumour to multiple distant sites, including lymph nodes, blood, lung, brain, and bone (Lelekakis *et al.*, 1999; Pulaski and Ostrand-Rosenberg, 1998).

Based on the lack of ER, PgR, and ErbB2 expression, 4T1 is widely used as a model for TNBC and closely resembles the clinical course of the disease in patients, including the location of the primary tumour and its metastatic spread (Schroers *et al.*, 2020).

Using the 4T1 cell line, the Hannon laboratory has developed a functional model of breast cancer heterogeneity capable of identifying different critical phenotypical drivers in a mixed population. Wagenblast *et al.* created a polyclonal population of 4T1 cells by single-cell sorting 4T1 cells labelled with a high-complexity DNA barcode library and mixing 23 clonal lines at equal ratios for tumour formation. The different subclones varied in their abilities to dominate the primary tumour or spontaneously metastasize to distant organs implicating the presence of clones with diverse transcriptomic

and phenotypic traits in a relative small subpopulation of cells (Wagenblast et al., 2015).

### 1.1.3 Tumour heterogeneity as an obstacle for therapeutic response

Therapeutic resistance continues to be a major clinical obstacle. Resistance to treatment can be classified as intrinsic (primary), manifested by the lack of a clinical response following therapy, or acquired (secondary), representing local or distant recurrence following initial response to therapy (Marine et al., 2020). At a cellular level, intrinsic resistance occurs due to resistance-mediating factors pre-existing in the bulk of tumour cells making therapy ineffective, whereas in acquired resistance, tumours show a dramatic initial response in most cases, but eventually develop resistance during treatment over time (Burrell and Swanton, 2014; Holohan et al., 2013). Acquired resistance can arise either through genetic evolution by the acquisition of *de novo* mutations or non-genetic mechanisms, including epigenetic and phenotypic changes, allowing cancer cells to adapt to the selective pressure. With tumours being dynamically evolving entities, intratumour heterogeneity has been recognized as a substantial source of poor prognosis and therapy resistance (Andor et al., 2016; Bhang et al., 2015; Das Thakur et al., 2013; Jamal-Hanjani et al., 2017; Juric et al., 2015; Zhang et al., 2014). Cancer therapy constitutes a strong, directional selection pressure that shapes the tumour evolutionary landscape (**Figure 1.2**) (Almendro et al., 2014). Most patients with advanced or metastatic disease experience therapeutic resistance leaving residual cancer cells behind, traditionally called "minimal residual disease" (MRD), and thus providing a reservoir for local or distant disease relapse. While genetic evolution was conceptually the driving force of acquired resistance, there is growing evidence that it is unlikely to be the main mechanism for therapeutic evasion (Marine et al., 2020). Notably, drug resistance can arise in a minor pre-existing subpopulation either through therapy-induced selection of cells which are intrinsically highly resistant to treatment or through an adaptive response, such as rewiring of signalling pathways or phenotypic switch (Holohan et al., 2013). In both cases, tumour heterogeneity poses a major problem substantially fuelling drug resistance.

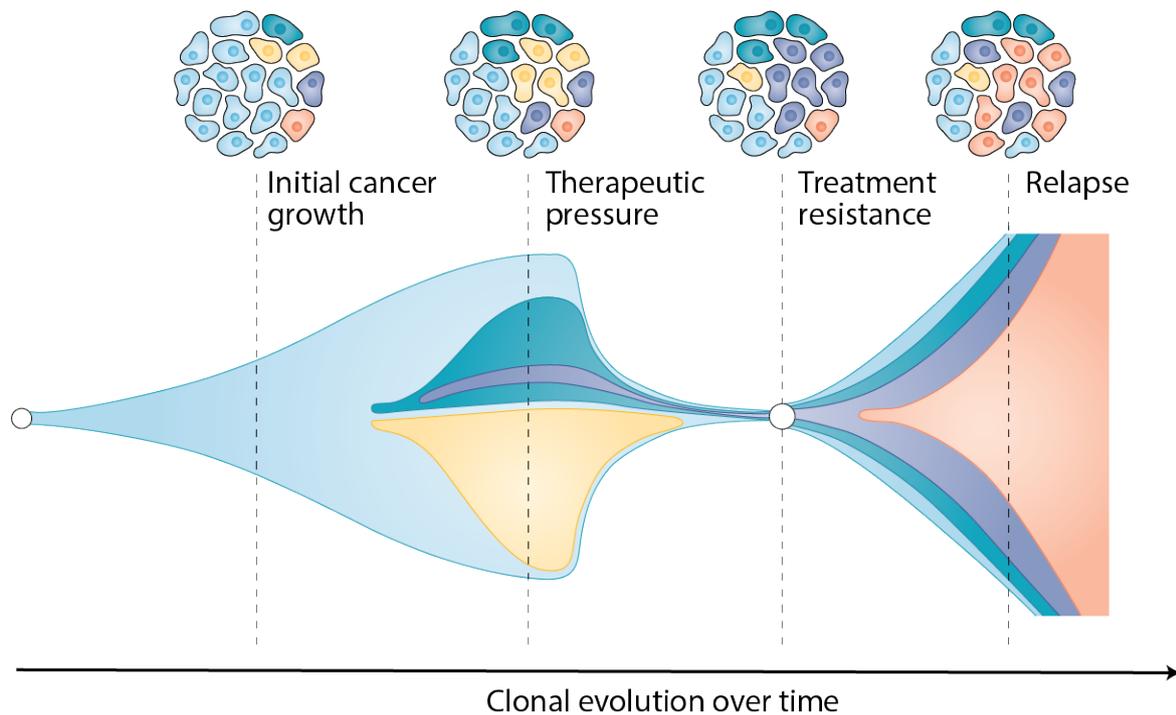


Figure 1.2 **Tumour evolution and emergence of therapy resistance through non-genetic mechanisms.** Clonal evolution is a key feature of cancer progression enabling tumour cell populations to adapt to selective pressures. When challenged with therapy, the bulk of clones are eradicated, whereas others confer a clonal advantage under therapeutic pressure. Cancer relapse is driven by a rare subpopulation of clones that are able to survive the initial therapeutic challenge. *Adapted from Marine et al. (2020).*

Despite treatment induced changes in the clonal composition of tumours, relatively few studies have been done in tumours, with the exception of hematopoietic malignancies (Ding et al., 2010; Landau et al., 2013). Using WGS/WES, studies in hematopoietic cancers compared diagnostic and relapse samples of leukemia patients. Ding *et al.* discovered either a new mutation in the original founding clone at relapse, or the expansion of a subclone of the founding clone with additional mutations (Ding et al., 2012). Moreover, Landau *et al.* found that ten out of 12 treated patients underwent clonal evolution, in contrast to one patient of the six untreated cases. Interestingly, they observed an enrichment of subclonal driver mutations with treatment and an increase in subclonal complexity suggesting a selection process for more aggressive clones during treatment (Landau et al., 2013).

A study by Kurtova *et al.* in bladder cancer linked resistance to treatment with cytotoxic chemotherapy to the selection of a pre-existing subpopulation (Kurtova et al.,

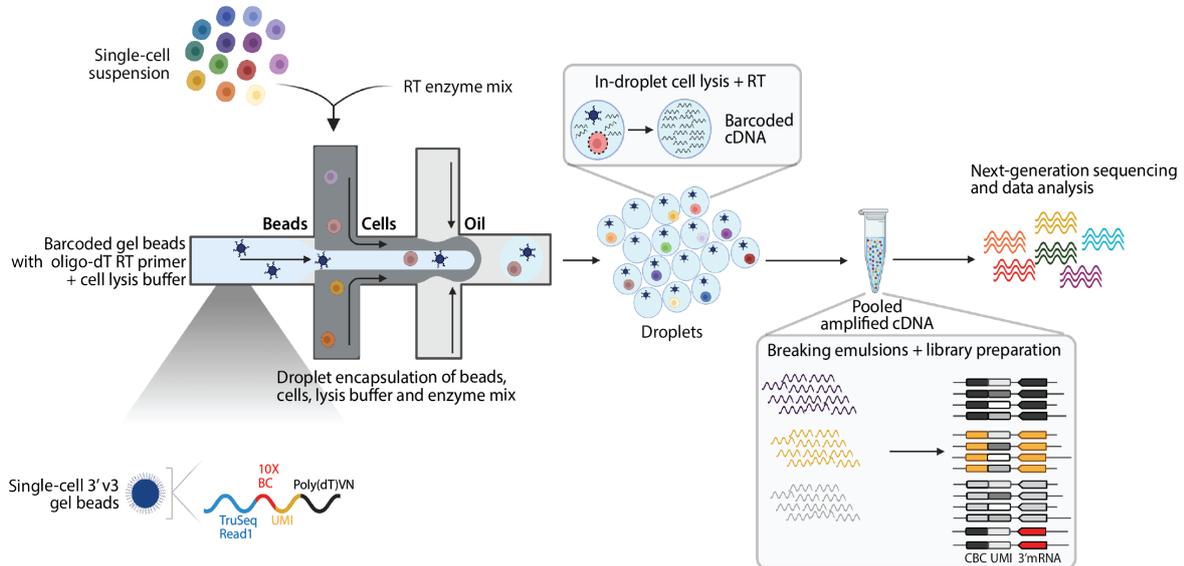
2015). Another study in breast cancer patients treated with chemotherapy suggested no change in genetic diversity, but instead an upregulation of EMT-associated genes and a selection for mesenchymal phenotypes (Almendro et al., 2014; Li et al., 2008). This is in line with the long standing concept of a treatment-induced phenotype-switch from epithelial to mesenchymal (Marine et al., 2020). Whereas most of these studies were relying on targeted markers or bulk genomic sequencing with limited resolution of clonal evolution, Kim *et al.* used a combined approach of single-cell DNA- and RNA-sequencing to analyse biopsies of TNBC patients before and after treatment. Their results revealed that resistant genotypes were pre-existing and adaptively selected in response to chemotherapy (Kim et al., 2018). Nevertheless, our understanding of how treatment impacts intratumour heterogeneity or how tumour heterogeneity impacts treatment which then in turn plays into the effectiveness of treatment, is still very sparse.

## 1.2 Single-cell RNA sequencing technologies

Single-cell RNA sequencing (scRNA-seq) has emerged as an essential tool for studying heterogeneous cell populations, including characterization of cell states, lineages and circuits. Since the first single-cell transcriptome sequencing was published in 2009, various new techniques have been developed (Tang et al., 2009). Methods for scRNA-seq use different strategies to barcode transcripts for their cell of origin and generate sequencing libraries. Low-throughput methods rely on time-consuming fluorescence-activated cell sorting (FACS) of cells into many plates that must be processed separately (Hashimshony et al., 2016; Picelli et al., 2013). High-throughput approaches encapsulate single cells in droplets (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) or wells (Gierahn et al., 2017) containing reagents and barcoded beads. By using a scalable, combinatorial indexing strategy, all mRNA in a droplet or well is assigned to their cell of origin without physically isolating single cells.

Currently, there are three droplet-based systems for scRNA-seq: inDrop (Klein et al., 2015), Drop-seq (Macosko et al., 2015) and 10X Genomics Chromium (10X) (Zheng et al., 2017). All of these technologies share similar workflows, including droplet generation and barcoded oligonucleotide coated beads, but differ in the bead and barcode design as well as their cDNA amplification method (Zhang et al., 2019). In a comparative study 10X generally performed with higher molecular sensitivity and

precision and less technical noise, while offering a user-friendly platform (Zhang et al., 2019).



**Figure 1.3 10X Genomics Platform.** Single cells are combined with reagents and encapsulated in GEMs containing barcoded gel beads. Gel beads are coated with barcoded oligonucleotides consisting of Illumina adapters, 10X barcode (cell barcode), UMIs and oligo dTs, which capture polyadenylated RNAs. RT reaction takes place inside each GEM, followed by pooling of barcoded cDNA for amplification and library preparation in bulk.

The 10X Genomics technology relies on capturing single cells on a microfluidic platform (**Figure 1.3**). Specifically, single cells are partitioned in Gel Beads in Emulsion (GEM) which further contain barcoded oligonucleotide primers and reagents for reverse transcription (RT) reactions. Each primer contains a poly(dT) sequence to capture mRNAs, a 12 nucleotide (nt) unique molecular identifier (UMI) to count individual mRNA molecules, a 16 nt 10x barcode unique to each cell and an Illumina TruSeq Read 1 primer binding site. Following GEM generation, cells are lysed, polyadenylated mRNA is captured by the poly(dT) part of the primer and reverse transcribed. All cDNA from single cells has been assigned with the same 10X barcode, allowing to map sequencing reads back to the single cell of origin. Thus, library preparation is performed in a bulk reaction followed by Illumina sequencing.

## 1.3 Lineage tracing strategies

Unravelling molecular dynamics, cell-fate decisions and spatial organisation of cells is the fundamental goal for understanding normal tissue development and homeostasis as well as disease. During differentiation, stem and progenitor cells move along a trajectory of cell fate decisions, refining their identity and function until reaching a terminal state or become quiescent until further differentiation occurs later in life. The gold standard for reconstructing the hierarchic relationship between an individual cell and its clonal descendants is lineage tracing, in which an individual cell is labelled with a heritable marker at an early time point in order to profile the fate of its clonal progeny at a later time point (Jensen and Dymecki, 2014). Pioneered in the 19th century, lineage tracing represents an essential tool for understanding mechanisms and dynamics of stem and progenitor cell fate determination. These cell fate determinations are associated with changes in the transcriptional and epigenetic landscape of cell states.

The same principles can be applied to cancer progression which also underlies an evolutionary process leaving phylogenetic traces along the route from the early transformation of normal cells into malignant cells to forming metastases at distant sites and adapting to environmental pressure through drugs or effects of the immune system (Schwartz and Schäffer, 2017).

Traditionally reliant on microscopy, prospective lineage tracing approaches have evolved to using heritable markers, including fluorescent proteins (Barker et al., 2007), mobile transposable elements (Sun et al., 2014), viral DNA barcodes (Naik et al., 2013) and Cre-mediated tissue-specific recombination (Pei et al., 2017), that were introduced into a cell and used for tracking its descendants. In contrast, retrospective lineage tracing utilises endogenous markers that naturally accumulate in the genome, such as somatic mutations (Ju et al., 2017; Zafar et al., 2017), microsatellite repeats (Frumkin et al., 2005) or epigenetic markers (Mooijman et al., 2016). Although valuable insights were gathered using these techniques, they cannot link the lineage information to a functional phenotype and are often limited by the number of cells and markers (Woodworth et al., 2017).

### 1.3.1 Lineage tracing by barcode-sequencing

Parallel advances in next-generation sequencing have opened up a new generation of genetic lineage tracing approaches (**Figure 1.4**). While the number of clones that can be tracked with fluorescent reporters is intrinsically limited, the complexity of DNA sequence barcodes scales exponentially with the length and multiplicity of the barcodes (Wagner and Klein, 2020). Each founder cell contains a unique DNA barcode, meaning the descendants of each founder cell inherit the barcode and can therefore be distinguished as a clonal unit.

In the beginning, the use of DNA barcodes relied on the identification of unique retroviral integration sites combined with Southern blot or PCR assays to read the barcode identity (Keller et al., 1985; Lemischka et al., 1986). Over the last few years, viral barcoding has been extensively used to delineate how hematopoietic stem and progenitor cells (HSPCs) differentiate and comprise blood (Gerlach et al., 2013; Gerrits et al., 2010; Lu et al., 2011; Naik et al., 2013). Moreover, Nguyen *et al.* used viral barcoding to reveal the highly complex clonal growth dynamics in serially xenograft transplantations (Nguyen et al., 2014).

Another recently developed technology relies on CRISPR-Cas9-directed genome editing to generate high-diversity DNA barcodes over time. The barcode, an array of clustered, regularly interspaced short palindromic repeats (CRISPR)/Cas9 target sites, is targeted by Cas9 resulting in a stable insertion or deletion (indel) "allele" that is inherited over subsequent generations. As cells divide, they acquire more Cas9-induced mutations at additional sites enabling the further deciphering of phylogenetic clades of cells. The first methods to demonstrate this principle were homing CRISPR (Kalthor et al., 2018, 2017) and genome editing of synthetic target arrays for lineage tracing (GESTALT) (McKenna et al., 2016). As an alternative to CRISPR-Cas9-based methods, Pei *et al.* used a Cre-loxP recombination system, termed *Polylox*, for endogenous barcoding of a mouse model (Pei et al., 2017, 2019). The *Polylox* barcode cassette consists of ten loxP sites in alternating orientations and nine intervening DNA that can be inverted or excised by Cre recombinase depending on the orientation of the flanking LoxP sites. By breeding the *Polylox* mouse model with a tissue-specific and inducible Cre mouse models, labeling of unique clones in a tissue and time-specific manner is achieved (Pei et al., 2019).

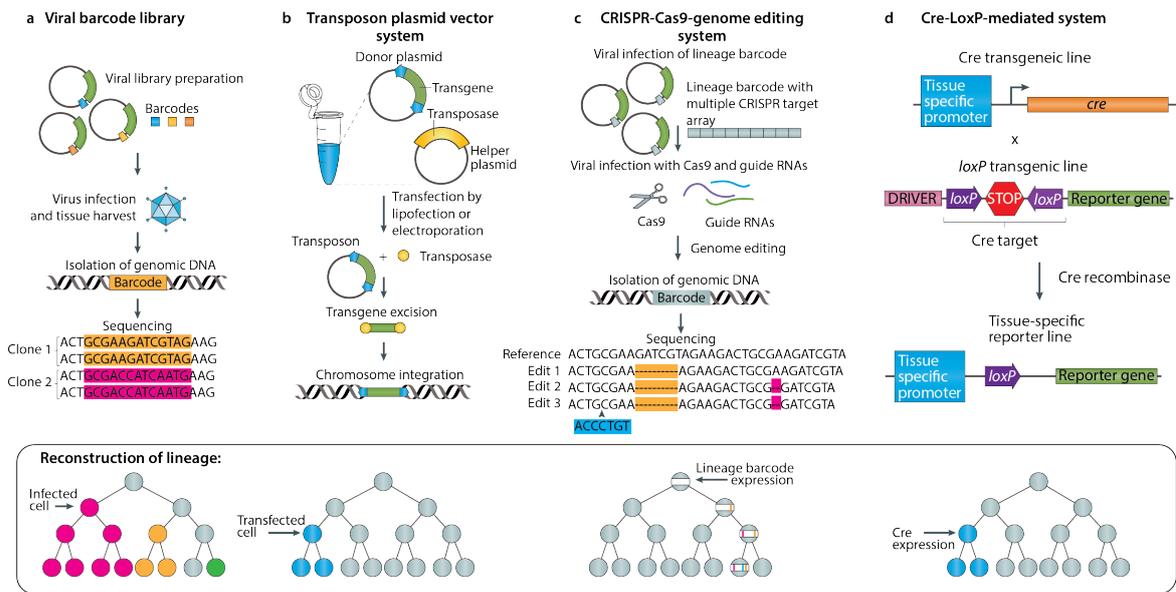


Figure 1.4 **Prospective lineage tracing approaches.** **a** | Viral barcoding systems rely on integration of a reporter transgene and a DNA barcode into the genome of the host cell. After propagation to progeny, all descendants from one founder cell share the same barcode, while clonally unrelated cells have different barcodes. Lineage tracing starts at the point of infection. **b** | Instead of using a viral system, barcodes can also be integrated into the genome using a transposon based system. **c** | Genome-editing systems use a lineage barcode with a CRISPR target array that accumulates mutations over time. Similar to retrospective approaches, phylogenetic trees are reconstructed based on the mutational patterns shared between cells. **d** | Genetic recombination systems, such as Cre-*loxP*, rely on the expression of a recombinase, controlled by a tissue-specific or cell-specific promoter. Upon Cre activation all progeny will be labeled with the reporter. *Adapted from Woodworth et al. (2017).*

### 1.3.2 Retrospective approaches

Prospective lineage tracing approaches rely on the introduction of exogenous markers into cells which marks the starting point and allows tracking of clones. However, such manipulation is not possible in the context of human development and disease. To overcome these limitations, endogenous markers, including single nucleotide variations (SNVs), copy number variations (CNVs), or other inheritable elements that accumulate over time can be used for lineage tracing. Similar to genetic barcodes, mutational marks are passed on to all progeny of the mutated founder cell. Due to the low frequency of naturally occurring mutations, high sequencing coverage is required.

CNVs are classified as DNA sequences more than 1 kb long present in various copy numbers when compared to the reference genome. Due to their relative ease of detection

in single-cell sequencing data, CNVs represent a potentially useful lineage-tracing tool. Several human diseases, especially cancer, have been linked to CNVs, and there has been evidence of the presence of CNVs also in healthy tissues, including the brain and skin (Abyzov et al., 2012; Cai et al., 2014; McConnell et al., 2013; Zarrei et al., 2015). Multiple studies have leveraged CNVs as a tracking tool to reconstruct clonal dynamics of breast tumour initiation, invasion and metastasis (Casasent et al., 2018; Navin et al., 2011; Wang et al., 2014). By combining whole-genome amplification and sequencing, Navin *et al.* identified several genetically distinct clones in breast tumours that arose during sequential clonal expansions despite a low sequencing coverage of 6% (Navin et al., 2011). Subsequent work from Navin *et al.* has led to the development of Topographical Single Cell Sequencing (TCSC), a method that combines laser capturing and single-cell RNA sequencing to preserve the spatial context while measuring CNVs. Application of this method to several ductal carcinomas showed distinct CNV profiles between ductal carcinoma *in situ* (DCIS) and invasive ductal carcinoma (IDC). Moreover, they observed that cells harboring the same CNV profile were not spatially restricted overall, confirming the migratory nature of IDC cells (Casasent et al., 2018). In a recent study, Campbell *et al.* developed "clonealign", a statistical tool to link transcriptomic profiles to genomically defined clones assuming only an effect of copy-number dosage on transcript abundance (Campbell et al., 2019). They demonstrated the power of the technique by identifying clone-specific phenotypes in a triple-negative breast cancer PDX. However, one requirement of clonealign is the presence of complex structural genomic rearrangements present in most but not all cancer types.

Another class of endogenous markers typically used for lineage tracing includes SNVs and small indels. Both frequently occur in non-coding regions in somatic cells with no phenotypic consequence. Several studies have successfully used SNVs from bulk DNA sequencing to infer phylogenetic trees (Abbosh et al., 2017; Gao et al., 2016; Ju et al., 2017). Like CNVs, SNVs can be identified in single-cell whole genome sequencing (scWGS) or single-cell whole-exome sequencing (scWES) data. However, the difficulty with detecting SNVs in single cells is due to the sparse distribution of SNVs in the genome, making it a challenge to detect the same SNVs in a large number of cells. Nevertheless, this method has been utilized to delineate the clonal evolution in healthy and tumour tissue (Leung et al., 2017; Lodato et al., 2015; Xu et al., 2012).

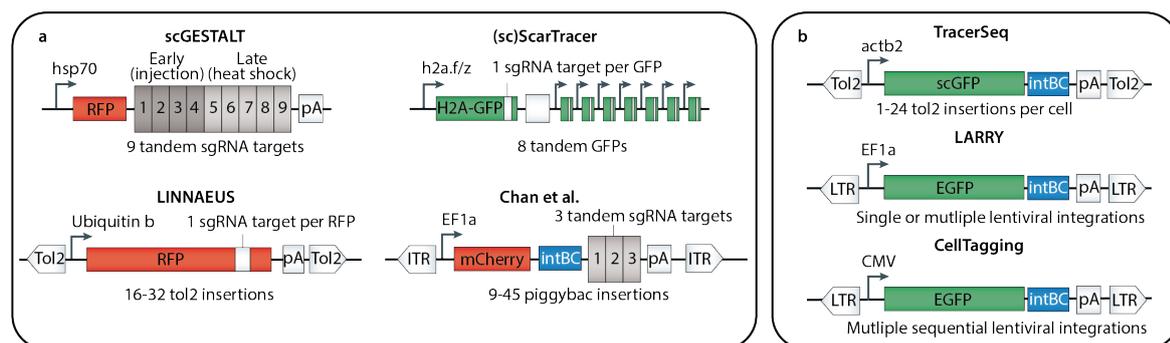
## 1.4 Single-cell transcriptomics meets lineage tracing

Neither state nor lineage information alone provide the full picture of heterogeneous cellular states and the ancestral relations between them. While scRNA-seq technologies enabled the construction of comprehensive transcriptional atlases, there is no direct capture of long-term dynamic relationships between individual cells or between cells and their progeny. Parallel advances in sequencing-based lineage tracing methods now facilitate the mapping of clonal fates onto the transcriptional cell states. These approaches are incredibly powerful and offer the opportunity to integrate complementary information about both cell state and cell lineage and the assessment of gene-expression changes as a function. To achieve barcode profiling from single-cell transcriptomes, most methods rely on a similar experimental design by embedding a barcode sequence into the 3' UTR of a constitutively transcribed fluorescent protein allowing it to be captured on single-cell RNA platforms using oligo-dT beads for library preparation (**Figure 1.5**).

The first innovations that included expressed lineage barcodes to be read from mRNA in whole-transcriptome scRNA-seq were single-cell GESTALT (scGESTALT) (Alemany et al., 2018), lineage tracing by nuclease-activated editing of ubiquitous sequences (LINNAEUS) (Spanjaard et al., 2018) and scScarTrace (Raj et al., 2018). These CRISPR-Cas9 based methods rely on the transcription of Cas9-induced mutations in a genomic target site that functions as a unique evolvable barcode allowing the simultaneous quantification of the clonal history, as well as recovering the transcriptome of single cells. Due to the activity of Cas9 over several hours, barcoding occurs sequentially resulting in complex scar patterns (Baron and van Oudenaarden, 2019). This dynamic barcoding process offers a powerful tool enabling the reconstruction of multi-branching lineages. The first three experimental approaches relying on CRISPR-Cas9 to label single cells focused on clonal dynamics in zebrafish embryogenesis (Alemany et al., 2018; Raj et al., 2018; Spanjaard et al., 2018). Chan *et al.* further implemented the CRISPR-Cas9 genetic lineage tracing technology in mice to study lineage relationships during early embryogenesis (Chan et al., 2019). They generated a transcribed and evolving molecular recorder consisting of a synthetic Cas9 target site including three cut sites and a static 8 bp integration barcode. Using a refined version of their molecular recording system, Quinn and colleagues explored clonal dynamics in a mouse model of metastasis (Quinn et al., 2021). Their findings identified driver

genes of metastatic capacity and showed that metastatic transcriptional signatures were pre-existing and shared between clonally related cells.

The cumulative editing nature of these approaches allow lineages to be reconstructed, however Cas9 degradation can restrict the time interval of the lineage tracing process. Most CRISPR-Cas9 barcoding systems rely on introducing random insertions and deletions through double-strand break repair by non-homologous end joining (NHEJ). To this point the extent and effect of potential off-target double-strand effects remains unknown. Moreover, lineage information could potentially be lost by the generation of multiple DNA double-strand breaks in close proximity resulting in the excision of the intervening sequence. In order to reconstruct the full lineage, all edits must be recovered which can be a problem considering that barcode transcripts expressed in cells could just be undetected in the mRNA profile. While CRISPR-Cas9-based technologies are the pioneer technique for whole-organism lineage tracing, these approaches suffer from the drawback of requiring the generation of transgenic animals.



**Figure 1.5 Integrative technologies for lineage tracing at single-cell level.** There are two major barcode designs: cumulative, "evolving" DNA barcodes (*a*) or integration DNA barcodes (*b*). Both experimental designs rely on embedding the barcode sequence into the 3' UTR of a constitutively transcribed fluorescent protein allowing the barcode to be expressed as mRNA and detected with scRNA-sequencing as part of each single-cell transcriptome. The methods differ in their barcode integration by using either a lentivirus (LARRY, CellTagging), transposase (TracerSeq, LINNEAUS, Chan et al.) or CRISPR-Cas9 targeting of single guide RNA arrays (scGESTALT, ScarTracer). scGESTALT, single-cell genome editing of synthetic target arrays for lineage tracing; LINNEAUS, lineage tracing by nuclease-activated editing of ubiquitous sequences; LARRY, lineage and RNA recovery. *Adapted from Marine et al. (2020).*

One alternative for CRISPR-Cas9 based methods is TracerSeq, a transposon-based barcoding approach (Wagner et al., 2018). TracerSeq makes use of the Tol2 transposase system to randomly integrate a pool of barcodes located in the GFP 3' untranslated region. While the progressive asynchronous integration of the plasmid into the genome

over cell divisions allows the reconstruction of phylogenetic trees without inducing unpaired double-strand breaks or generating identical edits as with Cas9-based methods, it still relies on injection or electroporation (Wagner et al., 2018).

The implementation of DNA barcoding has utilized an alternative approach for linking cell origin and states across time based on DNA barcoding. In contrast to CRISPR-Cas9 based technologies, virus-based cell labeling is easier to use without complex genetic manipulation. In a first example, Bidy *et al.* established a method, CellTagging, to track fibroblasts to endoderm progenitors *in vitro* using sequential rounds of cell labeling (Bidy et al., 2018). To enable the reconstruction of lineage trees, they used several indexed libraries in subsequent infection rounds to map lineage relationships. The results of the study suggest the existence of a privileged cell state shared between clonally related cells, which predefines the reprogramming potential (Bidy et al., 2018). Further analysis of the gene expression profiles identified a predictor gene, *Mettl7a1*, that correlated with increased efficacy for successful reprogramming. A similar approach was applied by Weinreb *et al.*, who also used a lentiviral-barcoding tool, LARRY (lineage and RNA recovery), to study cell fate decisions in haematopoietic progenitor cells and to correlate early gene expression profiles to later fate potentials (Weinreb et al., 2020).

Collectively, these findings highlighted that these technologies have the power to resolve different clonal origins between similar cell states and moreover, enable linking of lineage behavior to gene expression (Wagner and Klein, 2020).

## 1.5 Spatial tumour heterogeneity

### 1.5.1 Molecular profiling of tumours in space

Tissues, including tumours, are composed of a mixture of cell types and states whose spatial organisation is fundamentally related to its function and interactions. The molecular makeup of tumours highly depends on external factors, such as the tumour microenvironment and site-specific features. Previous work using multiregion sequencing strategies has provided evidence of spatial intratumour heterogeneity in multiple tumours, with spatially separated genotypes followed by evolving phenotypic intratumour diversity (Gerlinger et al., 2014, 2012). Others have highlighted

the importance of the niche on clonal fitness in the tumour, revealing intratumour heterogeneity in various cancer types across multiple spatial scales (McPherson et al., 2016; Sottoriva et al., 2015, 2013). However, most of these studies above have been using bulk genomic profiling obscuring the full picture of the existing heterogeneity. While massively parallel scRNA-seq approaches allow profiling of the full transcriptome with single-cell resolution, these techniques rely on tissue dissociation and, thus, do not contain the spatial context. Conversely, *in situ* hybridization techniques measure the spatio-temporal patterns of gene expression, but are limited in their scalability. Over the last few years, new highly multiplexed epitope imaging methods have been developed, such as serial immunofluorescence imaging, CyTOF imaging mass cytometry and multiplexed ion beam imaging (Angelo et al., 2014; Giesen et al., 2014; Lin et al., 2015). These approaches have greatly contributed to our understanding of structural tissue organisation, allowing morphological classifications especially between different cancer types (Crosetto et al., 2015). Development of spatial transcriptomics methods in parallel have given us further insights into molecular processes on subcellular level. Ultimately, simultaneous measurements of genomic, transcriptomic and proteomic networks *in situ* would need to be acquired in order to fully understand the complex cellular processes and their dependence on both the microenvironment and tissue morphology.

### 1.5.2 Spatial transcriptomic approaches

Resolving spatial transcriptomics has been a long-term challenge, starting with fluorescence *in situ* hybridisation (FISH)-based approaches to current spatial multiplexed whole transcriptome detections at single-cell resolution. Historically, spatial patterns were visualized through direct or *in situ* staining of a small subset of genes. *In situ* hybridization involved hybridisation of a labelled oligonucleotides to mRNA molecules within a cell to study gene expression in a spatial context. Recent advances of this approach resulted in the development of single-molecule fluorescence *in situ* hybridization (smFISH). This quantitative method relies on multiple fluorescently labeled oligonucleotide probes that are hybridized along the target mRNA transcript and are imaged as diffraction-limited spots, which are counted to quantify the expression level of the target gene. Moreover, smFISH has nearly 100% detection sensitivity and can be applied to whole tissue sections (Itzkovitz et al., 2012; Lyubimova et al., 2013; Raj et al., 2008). Although this approach allows robust transcript quantification in morphological intact tissue, the number of transcripts that can be simultaneously detected is limited

by the number of spectrally resolvable fluorophores. Subsequently, efforts have been made to increase the numbers of detectable mRNA species at single cell resolution with multiplexed smFISH approaches, including temporal and spectral mRNA barcoding (Lubeck and Cai, 2012; Lubeck et al., 2014). While spectral barcoding visualizes the target with a specific combination of fluorophores, temporal barcoding uses multiple cycles of smFISH hybridization and stripping to label the same RNA molecule in a predefined colour sequence with a limited pool of fluorophores. In this manner, the number of detectable transcripts scales as  $F^N$ , in which  $F$  represents the number of dyes and  $N$  the number of hybridization cycles. This approach was called sequential fluorescence *in situ* hybridization (seqFISH) (Lubeck et al., 2014). Over the years, the detection ability of seqFISH was significantly scaled up from 12 genes in fixed cells up to 10,000 cells in tissue sections, resulting in the development of seqFISH+ (Eng et al., 2019; Lubeck et al., 2014).

Since temporal barcoding is prone to errors through false-negative hybridization or tissue movement resulting in barcode switching, Chen *et al.* developed a novel error correcting barcode system, multiplexed error robust FISH (MERFISH) (Chen et al., 2015). However, MERFISH requires a large number of encoding probes to hybridize to the target molecule, and is, therefore, limited to longer transcripts only (>3 kb). A complementary method called spatially-resolved transcript amplicon readout mapping (STARmap) was introduced by Wang *et al.*, leveraging a combination of hydrogel-based chemistry, target signal amplification and *in situ* sequencing. In contrast to similar *in situ* sequencing methods, STARmap uses a novel error-reduction by dynamic annealing and ligation (SEDAL) sequencing approach in which sequencing errors in any cycle cause misdecoding and are effectively rejected.

An alternative to FISH is to sequence the RNA directly in the cells or tissue. Here, a predefined set of transcripts is targeted by specific padlock probes to create rolling-circle amplification product which can be identified in two ways: The fluorescent *in situ* sequencing (FISSEQ) method sequences the amplified cellular RNA content directly, whereas Ke *et al.* introduced a 4 nt barcode into the padlock probe and sequenced this barcode *in situ* instead (Ke et al., 2013; Lee et al., 2014).

Other technologies have introduced immobilized reverse-transcription oligo(dT) primers with positional molecular barcodes either on glass slides (Stahl et al., 2016) or

coupled to microparticles ("beads") arrayed onto rubber-coated glass coverslips (Rodrigues et al., 2019). Similar to single-cell RNA sequencing, each barcode allows spatial, cellular and molecular assignment of the mRNA molecule which is then sequenced via SOLiD (sequencing by oligonucleotide ligation and detection) chemistry. In contrast to the targeted imaging-based approaches, these methods offer unbiased capture of the transcriptomic landscape with a relatively high throughput, however they still lack subcellular resolution.

The ultimate goal is to perform multimodal spatial profiling that allows parallel measurements of genome, transcriptome and proteome. One of the first integrative technologies was recently presented by Argelaguet *et al.* who generated a single-cell multi-omics map using single-cell nucleosome, methylome and transcriptome sequencing during gastrulation in mouse embryos (Argelaguet et al., 2019). In another study, seqFISH+ was combined with DNaseq+ for genomic analysis and sequential immunofluorescence measurements for protein expression (Takei et al., 2021).

# Chapter 2

## WILDseq: Development of an integrative barcoding approach

*The work described in this chapter was the result of a collaboration with an interdisciplinary team within the IMAXT lab. The design and all experimental work was performed by me. The analysis pipelines for the barcode assignment were built by Dr. Kirsty Sawicka. The single-cell RNA sequencing analysis was performed under the guidance of Dr. Kirsty Sawicka. Fluorescent cell sorting was performed by the Flow Cytometry Core Facility and sequencing was performed by the Genomics Core, including single-cell RNA library preparation.*

### 2.1 Introduction

Tumours are composed of a heterogeneous mixture of individual cells that differ from each other in their genetic, epigenetic and transcriptional landscape. This intrinsic heterogeneity, revealed through recent advances in high-throughput sequencing, provides a fuel for cancer evolution leading to different cell fates, such as death versus survival of cancer cells upon therapeutic intervention. Thus, tumour heterogeneity represents a major clinical obstacle that results in short-lived success of most drug treatments.

Cancer progression follows evolutionary principles which result in clear phylogenetic signatures along the way. Lineage tracing approaches have the power to map these signatures and infer tumour cell ancestry from the pattern of shared features across tumour subpopulations. Prospective lineage tracing relies on labelling individual cells at an early time point in order to measure population dynamics at clonal resolution.

One limitation of lineage tracing approaches is that the main focus relies on genotypic relations of cells to each other, but lacks any information about the phenotypic state of the cell. Single-cell RNA sequencing (scRNA-seq) technologies provide the potential to explore the transcriptional states of thousands of individual cells, thereby capturing diversity in heterogeneous cell populations. Thus, combining lineage tracing and scRNA-seq would enable single-cell omic-scale profiling, while simultaneously reporting lineage information.

Here we present a single-cell resolution clonal tracking approach, which we have named "WILDseq" (Wholistic Interrogation of Lineage Dynamics by sequencing), based on a lentiviral barcoding system, permitting the parallel capture of individual cell state and cell lineage. In this context, clones refer to a population of cells which inherited the same WILDseq barcode from the founder clone. Clonal labelling starts at the point of infection. "WILDseq" integrates complementary information about cell lineage and cell state into a unified view of clonal dynamics over time. We uniquely labelled individual cells of the heterogeneous 4T1 breast cancer cell line with our heritable, expressed WILDseq barcode and analysed clonal distributions *in vitro* and *in vivo*. In a first proof-of-principle experiment, we demonstrated the ability to capture gene expression and lineage information in parallel *in vitro* using WILDseq. Our results not only exemplified the functionality of the WILDseq platform in capturing clonal identity and gene expression simultaneously, but also highlighted the importance of adding a "bottlenecking" step in the pipeline. Thus, we established and optimized a bottlenecking strategy allowing us to observe clonal dynamics in a heterogeneous model of breast cancer *in vivo*. These experiments set the basis to explore clonal changes in cancer progression under therapeutic pressure in our following experiments.

## 2.2 Material and Methods

### 2.2.1 Cell culture

The mouse mammary tumour cell lines 4T1 (ATCC) and D2A1-m2 (Jungwirth et al., 2018) were cultured at 37 °C in DMEM (Gibco), supplemented with 10% heat-inactivated fetal bovine serum (Gibco) and 50 U/mL of penicillin-streptomycin (Gibco) under 5%  $CO_2$  culture conditions. The 293FT (Thermo Fisher Scientific) packaging cell line for virus production was cultured following manufacturer's instructions. Cells were split into fresh culture medium every two to three days by trypsinization with TrypLE reagent (Gibco) quenched with complete DMEM, and maintained at cell density of 70-90%.

### 2.2.2 WILDseq library design and cloning

To generate the WILDseq library, a barcode cassette was introduced into the 3'UTR of zsGreen in the pHSW8 lentiviral construct, using PCR (Q5 High-Fidelity DNA Polymerase, NEB) and Gibson Assembly (NEB). The pHSW8 backbone was constructed in a four-way Gibson Assembly (NEB) by inserting an antiparallel cassette of a PGK promoter for the expression of zsGreen, a cloning site for high-diversity barcode libraries and a synthetic polyA signal into an empty pCCL-c-MNDU3-X backbone (Cat.No.81071, Addgene). The barcode library was designed by generating 12 nt variable sequences using the R package DNABarcodes (Buschmann, 2017) and a set Hamming distance of 5. The resulting oligo pool was purchased as a custom oligo pool (Twist Bioscience). The barcode library was built by annealing reverse complement oligos (BarcodeOligo\_Fwd/Rev) and amplifying the product by PCR for 20 cycles (using Assembly\_Fwd/Rev primers). Each oligo contains a specific PCR handle, a 12-bp variable region and 20-bp constant linker (2.1).

The amplified barcode library (5'-AGCGATTCAAAGTTCTATCCGNNNNNNNNNNNNNNtgcacggttaaccgatgcaNNNNNNNNNNNNATCGTATAGTAAACGAGCGCAT-3')

was purified by columns (Gel extraction kit, Qiagen) and the vector backbone was prepared by digestion with SmaI (NEB). (Ns denote WILDseq barcode sequences, lowercase denotes linker sequence, uppercase denotes PCR handles in the barcode cassette)

10 ng amplified WILDseq barcode library was cloned into 100 ng digested pHSW8 vector in a 50  $\mu$ L Gibson reaction (NEB), cleaned with a PCR purification kit (Qiagen)

Table 2.1 Barcode oligos

Name	Sequence
Assembly_Fwd	AAACTCTTGAGTGAAGTCCAGTGATTTTGAACCAAGCGATTC AAAGTTCT
Assembly_Rev	ccttgccctgaTAACTGGAGGCAGTAATTTACAGCCATGCGCT CGTT- TAC
BarcodeOligo_Fwd	TGAACCAAGCGATTCAAAGTTCTATCCGNNNNNNNNNNtgc atcggttaaccgatgca
BarcodeOligo_Rev	ATGCGCTCGTTTACTATACGATNNNNNNNNNNtgc atcggttaaccgatgca

and eluted in 15  $\mu$ L H<sub>2</sub>O. The entire reaction volume was transformed into 10- $\beta$  electrocompetent *E. coli* cells (NEB) and cells were expanded in liquid culture for 18 h at 37 °C. Plasmid DNA was extracted with a MaxiPrep Plasmid Plus kit (Qiagen).

### 2.2.3 Virus production and transduction

Lentiviral particles were produced by cotransfecting 293FT cells (Thermo Fisher Scientific) with transfer plasmid and the standard packaging vectors pMDL, CMV-Rev and VSV-G using either Lipofectamine 2000 (Invitrogen) or the calcium-phosphate transfection method (Wigler et al., 1978). After 16-18 h, media was changed and cells were supplied with fresh growth media. Viral supernatant was collected 48 h after transfection and applied to cells immediately following filtering through a 45  $\mu$ m syringe. Alternatively, viral supernatant was placed at 4 °C for short-term storage or -80 °C for long-term storage. When necessary, virus was concentrated using ultracentrifugation. Lentiviral titer was determined by serial dilution and measurement of fluorescence via flow cytometry. Cells were infected with lentiviral barcode libraries at a very low multiplicity of infection ( $\sim$  0.2), allowing only one barcode per cell. After two days in culture, cells expressing zsGreen (zsGreen+) were sorted using FACS Aria cell sorter (BD Bioscience).

### 2.2.4 Library complexity analysis

To obtain a comprehensive estimate of the library complexity, barcodes were amplified in four separate PCR reactions with primers containing Illumina adaptors from barcoded plasmid library. All reactions were pooled together, concentrated and purified on a column and then sequenced on one lane of an Illumina HiSeq4000. Reads that

contained the WILDseq barcode motif were identified and extracted from the FASTQ files. Reads reported for each WILDseq barcode were filtered based on a 90% percentile cut-off. The resulting whitelist was further filtered for barcodes containing the common linker region.

### 2.2.5 Bottlenecking and characterization of WILDseq pools.

4T1 or D2A1-m2 cells were infected with WILDseq library at low MOI ( $\sim 0.2-0.3$ ). Two days after infection, various numbers of zsGreen+ cells were sorted into pools. Individual cell pools were cultured for two weeks allowing the barcode clones to stabilise in culture. A whitelist was generated for each pool via reverse transcription.

### 2.2.6 Whitelist generation of WILDseq barcodes

In order to generate a representative whitelist of expressed barcodes in our pools, RNA was extracted from WILDseq transduced cells (High Pure RNA isolation kit, Roche) and reverse transcription (RT) was performed using the Superscript IV reverse transcription kit (Invitrogen) and a target site-specific primer with a unique molecular identifier (UMI) and an Illumina sample index. cDNA was amplified by PCR (Q5 High-Fidelity DNA Polymerase, NEB) using primers containing Illumina-compatible adapters and sample indices (RTWhitelist\_Fwd/Rev). Alternatively, 1  $\mu\text{g}$  gDNA was extracted from WILDseq transduced cells (Blood&Cell Culture DNA Kit, Qiagen) and the barcode amplified by PCR using primers containing Illumina-compatible adapters (gDNAWhitelist\_Fwd/Rev). PCR products were purified via gel extraction (Qiagen) and quantified by Qubit. The library was sequenced on an Illumina MiSeq with a custom sequencing primer for Read1 (CustomRead1).

Reads that contained the WILDseq barcode motif were identified and the number of unique UMIs supporting each barcode was calculated. If barcode sequences amplified from gDNA were also available an additional filtering step was introduced excluding any barcode that was not detected in both RT-PCR and gDNA library. Based on the UMI counts, the top 90th percentile of detected barcodes was taken and collapsed for PCR and sequencing error correction using hierarchical clustering and combining sequences with a Hamming distance less than 5. Based on the UMI count, true barcodes were identified and included in the whitelist.

Table 2.2 RT-PCR and gDNA PCR primers

Name	Sequence
RT primer	CAAGCAGAAGACGGCATAACGAGATNNNNNNGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCAAGCGAT TCAAAGTTCTATCCG
RTWhitelist_Fwd	AATGATACGGCGACCACCGAGATCTACACCAGCAGTATGCATG CGCTCGTTTACTATACGAT
RTWhitelist_Rev	CAAGCAGAAGACGGCATAACGA
gDNAWhitelist_Fwd	AATGATACGGCGACCACCGAGATCTACACCAGCAGTATGCATG CGCTCGTTTACTATACGAT
gDNAWhitelist_Rev	CAAGCAGAAGACGGCATAACGAGATNNNNNNGTGACTGGAGTT CAGACGTGTGCTCTTCCGATCCAAGCGATTCAAAGTTCTAT CCG
CustomRead1 primer	CCAGCAGTATGCATGCGCTCGTTTACTATACGA

### 2.2.7 Single cell library preparation

Cells were dissociated using TrypLE Express (Gibco) and counted using a hemocytometer. Cell concentration was adjusted to 500 cells/ $\mu$ L in phosphate-buffered saline (PBS) + 0.04% bovine serum albumin (BSA, Sigma Aldrich), then processed on the 10X Genomics platform using the Chromium Gene Expression 3' v3.1 dual index kit, according to the manufacturer's instructions. Tumour tissue was harvested from mice injected with a pool of barcoded 4T1 cells. Tissue was minced prior to enzymatic and mechanical dissociation using the gentleMACS<sup>TM</sup> Octo Dissociator (Miltenyi Biotec) and the respective kit (Tumour Dissociation Kit mouse). Tissue was processed into single cell suspension following manufacturer's instructions and filtered through 70  $\mu$ m filter (Miltenyi) to remove any remaining larger particles from the single-cell suspension after dissociation. For *in vivo* samples, a live-dead sort with propidium iodide (Biolegend) was included to remove dead cells and debris prior to counting. Three million cells were sorted, pelleted and resuspended in 1 mL PBS + 0.04% BSA (Sigma Aldrich) prior to counting with a hemocytometer. The final single cell suspension was diluted as required and sequenced using the Gene Expression 3' v2.1 or v3.1 dual index kit.

### 2.2.8 Enrichment library preparation

WILDseq barcodes were further amplified from cDNA libraries with WILDseq-specific primers containing Illumina-compatible adapters and sample indices (Enrich\_Fwd/Rev primers).

Table 2.3 Enrichment PCR primers

Name	Sequence
Enrich_Fwd	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNNACA CTCTTCCCTACACGACGCTC
Enrich_Rev	CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNNGTGACTGG AGTTCAGACGTGTGCTCTTCCGATCTCAGCCATGCGCTCGTT TACTATAC

### 2.2.9 Animals

Female, six to eight week-old Balb/C mice were purchased from The Charles River Laboratory. All orthotopic injections were performed using 60,000 mouse mammary tumour cells resuspended in 50  $\mu$ L of a 1:1 mix of PBS and growth-factor reduced Matrigel (Corning). Injections were done into the fourth mammary gland. Primary tumour volume was measured using the formula  $V=1/2(L \times W^2)$ , in which  $W$  is width and  $L$  is length of the primary tumour. Tumours were collected 21 days post tumour cell injection.

### 2.2.10 Bioinformatic analysis of WILDseq scRNA-seq data

#### Read alignment and generation of expression matrix.

The Cell Ranger v3.0.1 pipeline (<https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest>) was used to process data generated by the 10X Chromium platform. The pipeline relies on STAR for alignment and was used in conjunction with a custom reference genome, created by adding the sequence of the zsGreen-WILDseq barcode transgene as a new chromosome to the mm10 mouse genome. To create a new reference genome compatible with Cell Ranger, we followed the instructions from 10X Genomics on building a custom reference

([https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/tutorial\\_mr#runmkref](https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/tutorial_mr#runmkref)). In brief, we first modified the reference genome used during alignment by adding the full lentiviral plasmid sequence including the transgene above. We then created a custom gene transfer format (GTF) file, containing our custom transgene annotation, followed by indexing of the FASTA and GTF files, using the Cell Ranger mkref function.

### Single cell sequencing analysis using Seurat

The digital gene expression matrices generated were then analysed with the Seurat R package using a standard pipeline (Stuart et al., 2019). Seurat was used to perform initial filtering of digital gene expression (DGE) data files as a quality control step. We first removed cells that had over 10,000 or less than 200 unique detected genes. This process was also used to filter out cells in which the proportion of the UMI count attributable to mitochondrial genes was greater than 12%. Reads mapped to the zsGreen-WILDseq barcode transgene were removed from the count matrix allowing unbiased clustering downstream.

Following DGE filtering, data was normalised by employing a global-scaling normalization method “LogNormalize” and cell cycle scores were calculated. For each cell, cell cycle scores were generated based on canonical markers for each cell cycle phase using the Cell-Cycle vignette ([https://satijalab.org/seurat/v3.1/cell\\_cycle\\_vignette.html](https://satijalab.org/seurat/v3.1/cell_cycle_vignette.html)). The data was processed with SCTransform, a statistical approach for modeling, normalisation and variance stabilisation. During this process, calculated cell cycle scores were regressed out. The normalised values were then used as input for principal component analysis (PCA). Subsequently, PCs were used to calculate cluster of cells with the Louvain algorithm. Resulting clusters were visualised using UMAP and differential expression analysis was performed using the non-parameteric Wilcoxon rank sum test.

### Barcode demultiplexing

***WILDseq barcode demultiplexing from transcriptomics data.*** Reads mapping to the zsGreen-WILDseq barcode transgene were extracted from the processed and filtered BAM file produced by Cell Ranger. The resulting BAM file was converted into a fastq file using the tool bamtofastq from 10X Genomics

(<https://support.10xgenomics.com/docs/bamtofastq>). Reads containing the full barcode sequence (20 nt constant linker with a 12 nt region on either side) were identified and processed for further downstream analysis. The resulting barcode list was mapped to the whitelist, generated for each WILDseq cell pool, using Bowtie (Langmead et al., 2009). Next, cell barcodes and UMIs of all reads were extracted from BAM file generated above and then merged with WILDseq barcodes based on barcode ID/name using a custom R script. WILDseq barcodes were grouped by cell barcodes, followed by counting the number of unique UMIs for each cell barcode-WILDseq barcode. The matrix was filtered for WILDseq barcodes represented by at least two independent reads and more than 50% of barcode mapped reads from the cell supported assignment.

***WILDseq barcode demultiplexing from enrichment PCR data.*** Using the package UMI tools (Smith et al., 2017), a whitelist of cell barcodes was generated from the corresponding transcriptomic run to extract 10X cell barcodes and UMIs from the raw read files. The sequence corresponding to the full barcode sequence (20nt constant linker with a 12 nt variable region on either side) was extracted from each read and then mapped to the WILDseq barcode whitelist using Bowtie. A WILDseq barcode was assigned to a cell if there were at least 10 UMIs which matched the barcode to the cell and at least twice as many UMIs supporting this assignment compared to the next best.

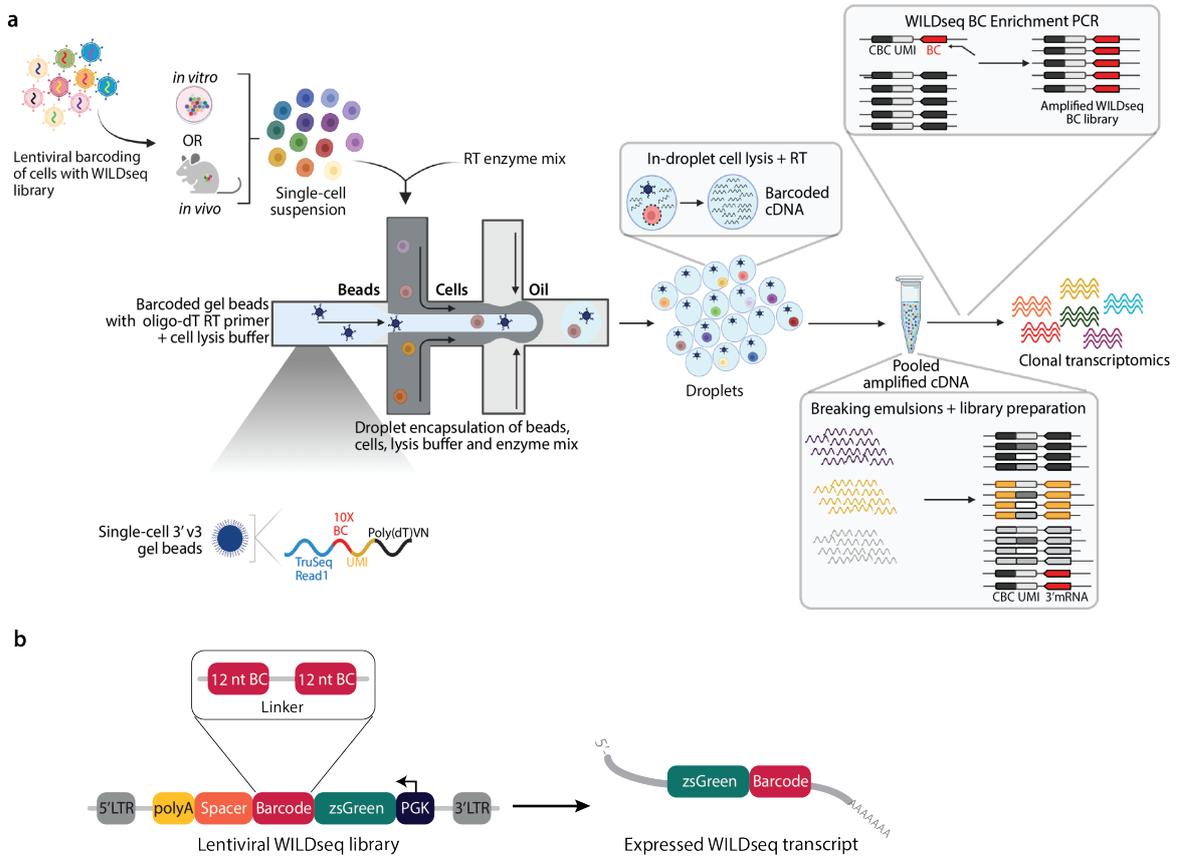
***Clone calling in 10X and enrichment PCR dataset.*** Barcode assignment from these two pipelines was compared and WILDseq barcodes identified by both methods were included in the final matrix. On the rare occasion, that the assignment did not match no clonal barcode was assigned. In the case that a WILDseq barcode was only detected by one method, a further more stringent filtering step was included. WILDseq barcodes only appearing in the 10X scRNA-seq data were required to have at least 5 UMIs, while WILDseq barcodes only observed in the enrichment PCR data needed to have at least 30 UMIs. The final barcode matrix was added to the metadata of the corresponding Seurat object.

## 2.3 Results

### 2.3.1 Design of the WILDseq platform and proof-of principle experiment

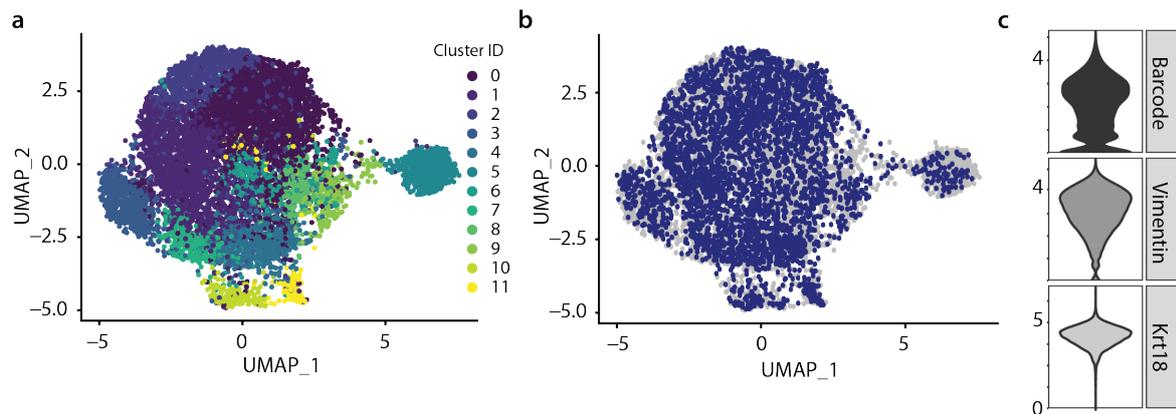
Droplet-based approaches for single-cell transcriptomics rely on two indexing strategies that allow deconvolution of pooled RNA-seq data into single cell gene expression profiles: a cell barcode (CBC) and a unique molecular identifier (UMI). In order to combine gene expression analysis with clonal lineage tracing, we built the WILDseq platform to add this additional layer of information (**Figure 2.1a**).

WILDseq is a lentiviral-based approach relying on uniquely labelling individual cells with a heritable barcode which can be captured as an expressed transcript by scRNA-seq. To deliver and capture WILDseq barcodes, we designed the "WILDseq vector", a third-generation lentiviral vector that contains a phosphoglycerate kinase (PGK)-driven expression cassette terminated with a strong synthetic polyA signal (**Figure 2.1b**). To prevent the internal polyA signal from interfering with the lentiviral particle production, the entire expression region was cloned in reverse orientation with respect to the genomic promoter. The expression cassette, moreover, contains a zsGreen reporter fluorophore allowing the cell sorting of infected cells. We chose zsGreen due to its brightness in flow cytometry and to specifically avoid fluorophores with known immunogenicity (Stripecke et al., 1999). The positioning of the barcode cassette relative to the polyA signal by  $\sim 120$ - $130$  base pairs (bp) was optimised for faithful transmission of WILDseq barcodes into scRNA-seq libraries using standard oligo-dT capture chemistry. Each barcode sequence was designed with a hamming distance of 5, where the hamming distance refers to the number of different positions between two strings of equal length, and thus allowing to correct up to 2 errors per barcode sequence. To increase the complexity of the library, we combined two barcode sequences through a common linker sequence and amplified the whole barcode cassette via PCR before inserting it into the WILDseq vector via Gibson Assembly. The final library consists of  $\sim 1,600,000$  potential barcode combinations. To ensure robust WILDseq barcode capture, we developed a one-step PCR enrichment strategy to specifically enrich WILDseq barcode transcripts out of the indexed, amplified cDNA of the transcriptome (discussed in detail later 2.3.3).



**Figure 2.1 WILDseq pipeline and barcode design.** **a** | Schematic illustration of WILDseq platform. Cells are infected with WILDseq library at a low MOI allowing only one barcode per cell. Infected cells are sorted two days post infection. WILDseq labelled cells express unique barcodes enabling clonal tracking with single-cell RNA sequencing using the 10X Genomics platform. An additional PCR step for the WILDseq barcode is included that specifically enriches for barcode sequences from the pooled cDNA and thereby, increases the assignment of WILDseq barcodes. 10x BC corresponds to cell barcode (index unique to each bead). UMI, unique molecular identifier (index unique to each bead oligo). WILDseq barcode specific for each clone (WILDseq transcript highlighted as red mRNA.) **b** | Schematic illustration of WILDseq library. Lentiviral construct containing two 12 nt WILDseq barcode sequences in the 3' UTR of zsGreen, followed by a synthetic polyadenylation signal. The spacer sequences places the barcode in the optimal distance to the polyA signal for faithful transmission into scRNA-seq library using standard oligo-dT capture. The barcode sequences were designed with a hamming distance of 5 to account for PCR and sequencing errors. Library complexity equals ~1.6 million barcodes in total.

In a pilot experiment, we performed scRNA-seq on a pool of individually labelled 4T1 breast cancer cells, analyzing ~10,000 cells in total using the WILDseq platform (**Figure 2.2a and b**).



**Figure 2.2 WILDseq barcode detection in 4T1 cells *in vitro*.** **a**| UMAP representation of scRNA-seq *in vitro* 4T1 dataset resolved in 12 individual clusters (highlighted in different colours). **b**| UMAP representation of scRNA-seq *in vitro* 4T1 dataset highlighting the expression of the WILDseq barcode transcript in blue. **c**| Violin plot of WILDseq barcode, *Vim* and *Krt18* expression in scRNA-seq *in vitro* dataset. WILDseq barcode was differentially expressed in population comparable to *Vim*.

Briefly, we infected 4T1 cells at a low multiplicity of infection (MOI) with the WILDseq barcode library such that each cell was labelled with a single, unique barcode. The library complexity is sufficient to label 10,000 cells in an experiment with less than 1% barcode overlap between clones. Two days post infection, transduced cells were sorted and recovered in culture. Overall, we assigned  $\sim 40\%$  of the cells a barcode. Gene expression levels of the barcode were highly variable within the population comparable to *Vim* (**Figure 2.2c**). For the vast majority of barcode expressing cells ( $\sim 90\%$ ), we detected only one WILDseq barcode per cell as expected. In less than 3% of the population we observed 2-3 cells with the same barcode and only one barcode was represented by six cells.

To study clonal dynamics *in vivo*, we therefore needed to bottleneck our population size - too large a starting population and the number of founding clones will be too large to record sufficient data of an individual clone using single-cell RNA sequencing, too small a starting population and it will lack the diversity to capture phenotypic heterogeneity. Based on the assumption that we capture  $\sim 8000$  single-cell transcriptomes in one scRNA-seq run, we decided to bottleneck the population to 250 cells with the hope of observing around 50-100 unique clones.

### 2.3.2 Robust detection of WILDseq barcodes *in vivo*

To demonstrate the capacity of our WILDseq technology to track clones *in vivo*, we established a WILDseq labelled population originated from 250 cells and injected 50,000 cells of this pool into the mammary fat pad of a Balb/C mouse. Specifically, we labelled 4T1 cells with our WILDseq barcode library, sorted 250 zsGreen positive cells two days post infection and stabilised the barcoded cell population in culture. Importantly, WILDseq labelled 4T1 cells were prepared such that each cell carries a distinct WILDseq barcode. 21 days after fat pad transplantation, we collected and dissociated the tumour into single cells which were processed through our WILDseq pipeline.

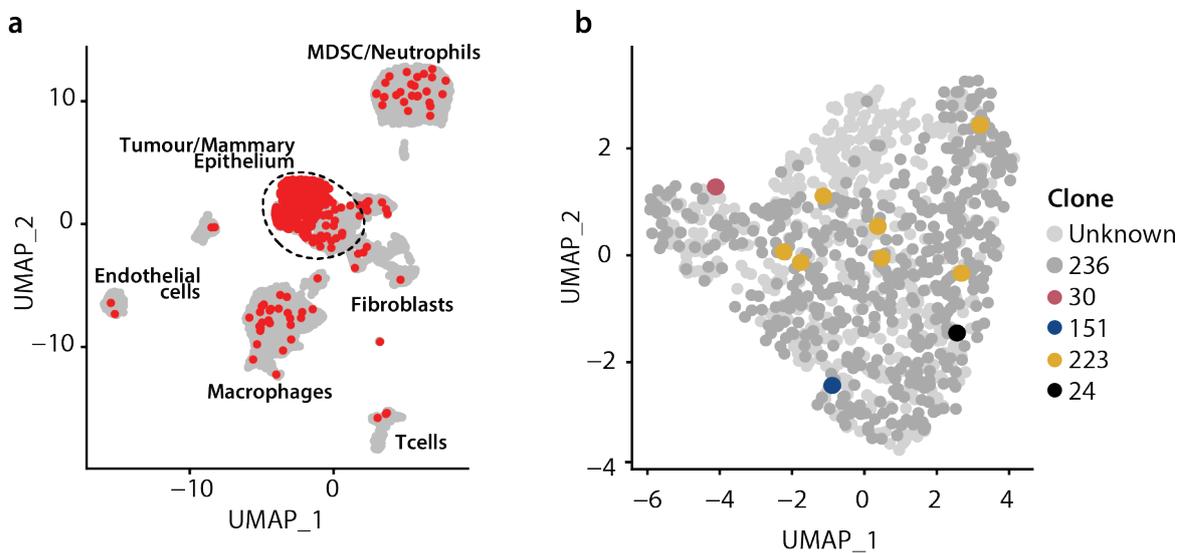
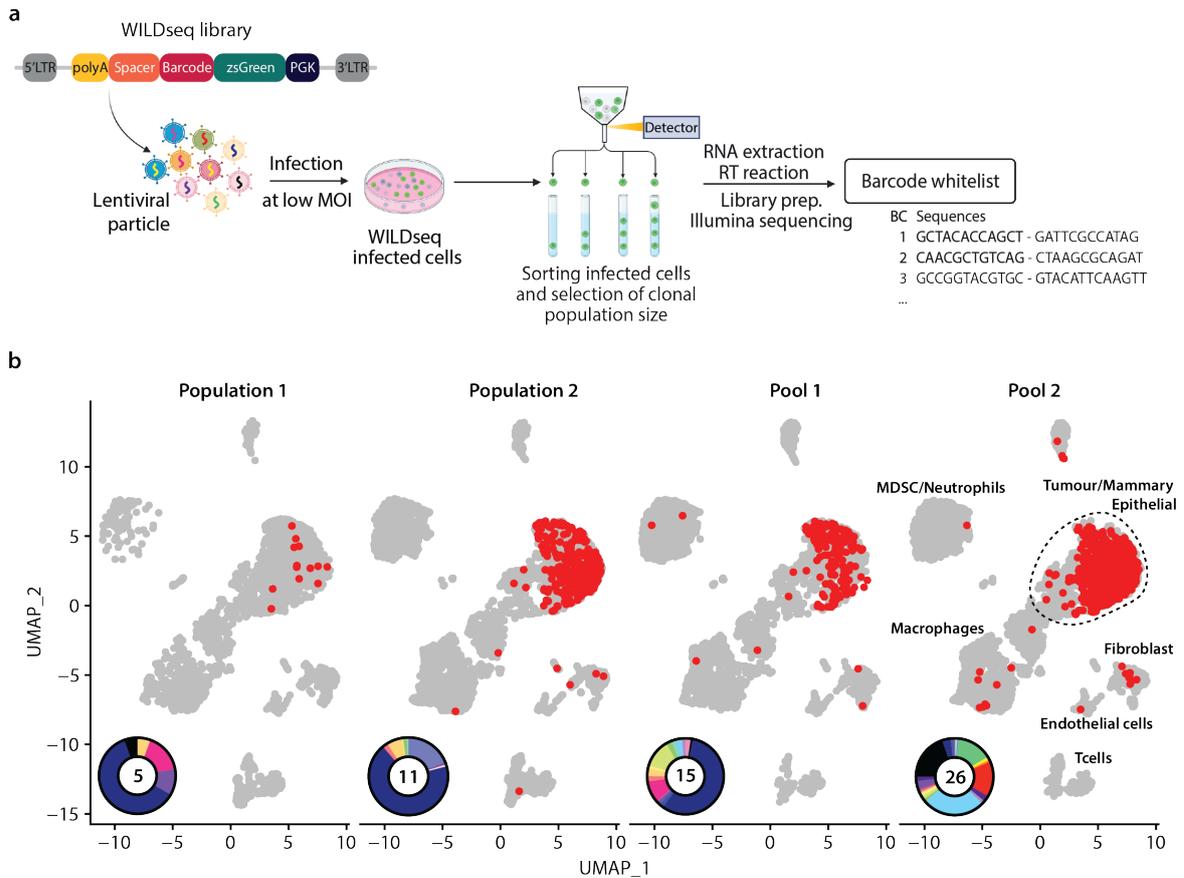


Figure 2.3 **WILDseq barcode detection *in vivo***. **a**| UMAP representation of 250 cells pool *in vivo* with barcode expression highlighted in red. **b**| UMAP representation of extracted tumour cell clusters with different WILDseq clones highlighted. Five clones were robustly detected with clone 236 dominating the tumour.

We robustly detected a WILDseq barcode in over 58% of all tumour cells. However, over 50% of the tumour was dominated by clone 236 out of the five different clones we assigned (clone 24, 30, 151, 223 and 236) (**Figure 2.3**). The other four barcodes only accounted for 10 assigned reads. Due to the lack of clonal diversity in the pool *in vivo*, we attempted to improve our bottlenecking strategy, which is a critical step to the success of the method. In parallel, we also developed a new strategy to create a reference list of barcodes, from now on referred to as a "barcode whitelist", to be able to robustly assign WILDseq barcodes to cells and to quantify the relative amount

of clones present in each of these pools (**Figure 2.4a**). Notably, we also observed a small fraction of barcodes ( $< 5\%$  per cluster) being assigned outside of the tumour clusters. We believe that this is due to technical noise, resulting in mis-assignment of reads to the wrong cell barcode. This is supported by the fact that increase of the level of stringency of the barcode assignment reduced the mis-assignment outside of the tumour clusters. In order to increase our clonal diversity, we tested different clonal pool sizes as well as different bottlenecking and pooling strategies. We generated two larger populations, consisting of 1000 (population 1) and 1250 cells (population 2), using a single sorting step. Additionally, we sorted multiple 10 cell populations and pooled them into three 120 cell pools, which we kept separate until immediately before mouse injections (pool 1). A similar approach was used to create a pool with 750 cells consisting of three independent 250 cell populations that are just combined before injection in order to increase clonal competition *in vivo* (pool 2).

Briefly, we extracted RNA from the pools which was then used for reverse transcription (RT) reaction with a custom primer detecting our WILDseq transcript. We furthermore included a unique molecular identifier (UMI) in our primer allowing the identification of PCR duplicates generated during library preparation. Barcode sequences were amplified and adapter and index sequences introduced by PCR. The library was then analysed via NGS. The resulting sequencing information was used to build a whitelist and a bowtie index to which barcode sequences from single cell RNA sequencing data could be compared, thereby allowing the robust calling of clone identities. To be able to select for the appropriately sized pool, we selected four pools based on the RT-PCR data of the *in vitro* pools and analysed them *in vivo* using scRNA-seq.



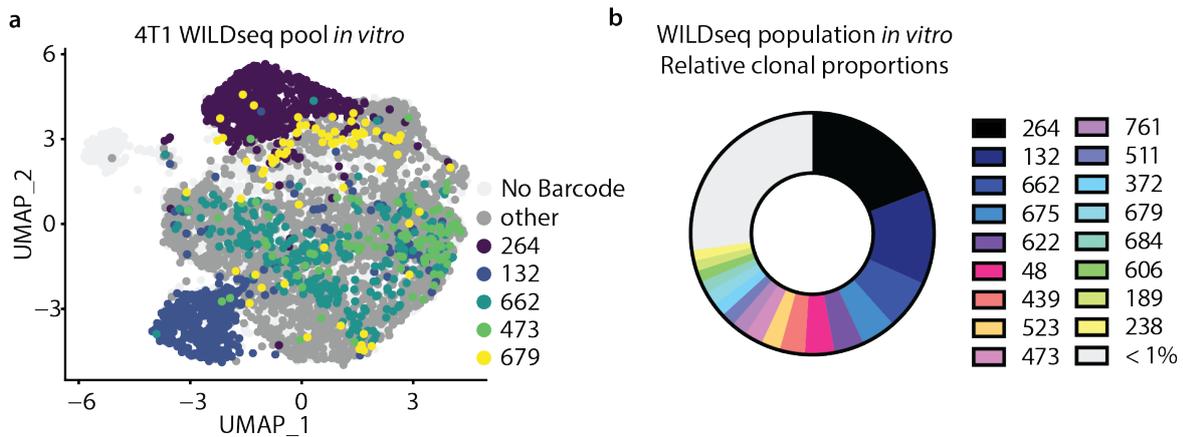
**Figure 2.4 WILDseq barcode detection *in vivo* and bottlenecking strategy of clone pools.** **a**| Schematic representation of bottlenecking strategy and generation of barcode whitelists using an RT-PCR approach. Barcode specific RT-primer included a UMI to account for PCR errors allowing robust calling of barcodes. Recovery of barcode sequences, followed by filtering and error-correction resulted in the final whitelist which was used to create a bowtie index. **b**| UMAP representations of differently sized clonal populations and pools with barcode expression highlighted in red. Populations (Population 1 = 1000 cell population, population 2 = 1250 cell population) and pools (pool 1 = 3x 120 cell populations, pool 2 = 3x 250 cell populations) were generated based on the bottlenecking pipeline illustrated in (a). Clonal distributions are illustrated in pie charts including total number of clones. Pool 2 exhibited the highest clonal diversity and number of different clones (26 clones).

Analysis of the different pools revealed that pooling three 250 cell populations (pool 2) immediately before injection, exhibited not only the highest diversity in gene expression but also in clones compared to the other pools *in vivo* (**Figure 2.4b**). This cell pool will be referred to as "WILDseq pool" and was used for all subsequent experiments. We expanded the three 250 cell populations separately for 14 days and cryopreserved them as the three parental barcoded 4T1 populations. Although we sorted more cells in the population based approach neither the expression nor the

clonal diversity were sufficient for our needs. While pool 1 had the second highest number of clones, we detected fewer barcoded cells overall in our tumour population.

### 2.3.3 Characterisation of WILDseq pool *in vitro*

To further characterise our WILDseq pool, we performed scRNA-seq on the barcoded cells *in vitro* (**Figure 2.5a**). The three 250 cell populations were stabilised in culture, sorted for their zsGreen expression and pooled to form the WILDseq pool immediately before single-cell library preparation. To increase the barcode detection and therefore the number of tumour cells for which a clonal lineage could be assigned, we prepared a barcode enrichment library using amplified, pooled cDNA from the transcriptomic library as input. Combining the barcodes detected in the transcriptomic library with those detected in the barcode enrichment library, we were able to robustly assign a WILDseq barcode to over 50% of the cells (3950 cells out of 7000 profiled cells). In total, 132 different barcodes were observed (**Figure 2.5b**).



**Figure 2.5 Characterisation of 4T1 WILDseq pool *in vitro*.** **a**| UMAP representation of WILDseq pool *in vitro*. Cells with an assigned WILDseq barcode are highlighted in dark grey or coloured spots. The main driver of clustering was clonality. **b**| Relative clonal abundance in the WILDseq pool *in vitro*. Individual clones are highlighted in different colours. Clones that are present at levels below 1% are illustrated together. In total, 132 WILDseq barcodes were detected.

We found that cells with the same barcode often clustered together. Clone 132 and 264 in particular exhibited a distinctive gene expression profile and clustered separately. This suggests that the main source of transcriptomic diversity *in vitro* is clonality. Interestingly, those clones that dominated the *in vitro* population were not the same as

those most frequently observed in the *in vivo* tumour generated from this clonal pool, indicating that although they had a growth advantage under cell culture conditions they lacked the ability to contribute to the tumour *in vivo*, possibly due to a lack of ability to engraft or to survive the conditions present in *in vivo*.

## 2.4 Discussion

Here, we introduce WILDseq, a scRNA-seq compatible clonal lineage-tracing approach that maps phenotypic cell states to clonal identities. A key advantage of WILDseq is it's simple to use and applicable to any cell culture and model system that is susceptible to viral transduction. We built WILDseq by inserting a reverse expression cassette containing two distinct error-correctable barcode sequences into a lentiviral vector backbone. Our initial experiment in which we uniquely labelled over 4000 4T1 cells with WILDseq highlighted the importance of adjusting the population size by bottlenecking. We optimized our bottlenecking strategy to obtain a heterogenous model system capable of generating sufficient data on individual clones while maximising the clonal diversity. Our final WILDseq pool consisted of three 250 4T1 cell populations that were used as parental populations for all following experiments. Transcriptomic characterisation of our cultured WILDseq pool demonstrated that clonality was the main driver of the differences in the expression profiles between clones. However, clustering of WILDseq tumours does not only depend on clonality, but also on external factors, such as the tumour microenvironment and other influences (e.g. oxygen supply, vascularisation,...) that modulate the transcriptome within a tumour cells to an equal or even greater impact than clonality. Finally, the transcriptomic natures we observed was specific to each of the clones.

Importantly, our WILDseq method offers the great advantage of integrating data across experimental treatment conditions. Current analysis methods largely rely on clear distinguishable transcriptomic markers of cell types for integration which has been proven difficult when it comes to the identification of the same subpopulation of cells, as transcriptomics signatures are often masked by transcriptomic changes induced by external factors (e.g. tumour microenvironment or drug treatment). Future optimisations of WILDseq may include alternative fluorophores or promoters and insulator sequences to overcome the limitation of silencing of lentiviral constructs in certain cell types. Moreover, this approach cannot directly provide information on spatial context and thus, another improvement could be the integration of *in situ* sequencing techniques to receive the full picture.

Currently, there are two technologies available that leverage a similar technology to WILDseq: LARRY (Weinreb et al., 2020) and Celltag (Bidy et al., 2018). While

---

both systems also rely on a lentiviral vector system to deliver a heritable, expressed barcode, neither of the two systems uses an error-correctable barcode which could potentially result in problems with barcode assignment due to PCR errors during library preparation and/or sequencing errors. Furthermore, both systems have only been applied to culture or hematological studies so far which is of comparative ease regarding sampling. Recently, Quinn *et al.* reported the use of a Cas9-based, single-cell lineage tracer for the first time in a model of lung cancer (Quinn et al., 2021). Their strategy differs from our WILDseq approach, however, in that it requires the delivery of multiple components for the CRISPR system and complicated lineage tree inference analysis due to the evolvable nature of the barcode. Moreover, we have shown that the expression of our WILDseq barcode is maintained over long periods of time *in vivo*, unlike CRISPR-Cas9 approaches that suffer from frequent single-cell barcode dropouts. Together, we established a novel approach that allows robust integration of clonal identities and transcriptomic characteristics to elucidate processes of clonal diversification and adaptation to therapy.



# Chapter 3

## Investigating therapeutic response using WILDseq in breast cancer

*The work described in this chapter was the result of a collaboration with an interdisciplinary team. All experimental work was the result of my own work with support for the animal work from Dr. Ian Gordon Cannell. The analysis pipelines for the differential gene expression and pathway analysis was performed by Dr. Kirsty Sawicka. The patient data analysis was performed by Dr. Ian Gordon Cannell. Fluorescent cell sorting was performed by the Flow Cytometry Core Facility and sequencing was performed by the Genomics Core, including single-cell RNA library preparation.*

### 3.1 Introduction

Clonal evolution is a key feature of cancer progression enabling tumour cell populations to adapt to selective pressures imposed by the tumour microenvironment or therapeutic interventions and is thought to be the underlying cause of drug resistance. Challenging a population of tumour cells with a drug could have several different outcomes on the clonal composition and diversity. Drug treatment could result in a population bottleneck of the tumour cell population followed by reconstitution of the tumour by a single or multiple resistant clones. Alternatively, the therapy could result in full remission of the tumour or have no effect on the diversity of the population. Currently, most studies have been focused on either the genetic or transcriptional mechanisms underlying therapy responses, independently lacking the ability to link functional phenotypes with individual clones. To overcome this knowledge gap, we applied our WILDseq approach to study the nature of clonal response to therapeutic pressure in

the syngeneic mammary carcinoma model 4T1. The standard of care treatment for many breast cancer patients is neoadjuvant or adjuvant chemotherapy consisting of a combination of taxanes and anthracyclines. Extensive research has been focused on identifying chemotherapy resistance mechanisms offering an ideal platform to validate our WILDseq approach. In this study, we defined resistance and sensitivity signatures of clones to the frontline chemotherapy docetaxel and the epigenetic anti-cancer drug JQ1. Moreover, we identified recurrent baseline signatures of our WILDseq clones *in vivo*. Analysis of human patient data revealed a strong association between non-responders and our docetaxel resistance genes, highlighting the clinical relevance of our findings. We further validated our WILDseq platform in a second triple-negative breast cancer cell line D2A1-m2 *in vivo* demonstrating that WILDseq can be applied to any cell culture system that is amendable to viral transduction. By applying WILDseq to study the effect of JQ1 treatment, we uncovered a potential synergy between baseline DNA damage signaling and JQ1 treatment in suppressing *MYC*-related signatures as well as a T-cell dependent resistance mechanism emphasizing the importance of a syngeneic model. Collectively, our work establishes a unique platform allowing the study of clonal response to therapy by uniquely distinguishable clones in a heterogeneous population *in vivo* and link them to their functional phenotypes.

## 3.2 Material and Methods

### 3.2.1 Tissue preparation for scRNA-seq experiments

Tumour tissue was harvested from mice injected with a pool of barcoded 4T1 cells. Tissue was minced prior to enzymatic and mechanical dissociation using the gentleMACS<sup>TM</sup> Octo Dissociator (Miltenyi Biotec) and the respective kit (Tumour Dissociation Kit mouse). Tissue was processed into single cell suspension following manufacturer's instructions and filtered through 70  $\mu\text{m}$  filter (Miltenyi) to remove any remaining larger particles from the single-cell suspension after dissociation. For *in vivo* samples, a live-dead sort with propidium iodide (Biolegend) was included to remove dead cells and debris, collecting three million live cells which were pelleted and resuspended in 1 mL phosphate-buffered saline (PBS) + 0.04% bovine serum albumin (Sigma Aldrich). Alternatively, cell concentrations of *in vitro* WILDseq pool was determined using a hemocytometer. Finally, cell suspension was submitted for scRNA-seq on the 10X Genomics platform using the Gene Expression 3' v2.1 or v3.1 dual index kit.

### 3.2.2 Animals and *in vivo* dosing

Female six to eight week-old Balb/C were purchased from The Charles River Laboratory. 60,000 tumour cells were resuspended in 50  $\mu\text{L}$  of a 1:1 mixture of PBS and growth-factor reduced Matrigel (Corning). All orthotopic injections were performed into the fourth mammary gland. Primary tumour volume was measured using the formula  $V=0.5(L \times W^2)$ , in which  $W$  is the width and  $L$  is length of the primary tumour. Tumour-bearing mice were treated with either vehicle or with different drugs seven days post transplantation. Mice received intraperitoneal injections of 75 mg/kg JQ1 (dissolved in DMSO and diluted 1:10 in 10%  $\beta$ -cyclodextrin (Sigma Aldrich)) for five consecutive days followed by two days drug holiday until tumours reached endpoint or 12.5 mg/kg docetaxel (dissolved in 1:1 mixture of ethanol and Kolliphor (Sigma) and diluted 1:4 in saline) administered three times a week for two weeks. Three animals per dose group were used. Vehicle-treated mice were sacrificed 21 days post tumour transplantation and treated animals were sacrificed when tumour volume reached that of vehicle treated animals at 21 days. D2A1-m2 tumours were generated the same way and treated with the following vehicle: 12.5% DMSO in 0.5% sodium carboxymethylcellulose.

### 3.2.3 Differential gene expression analysis

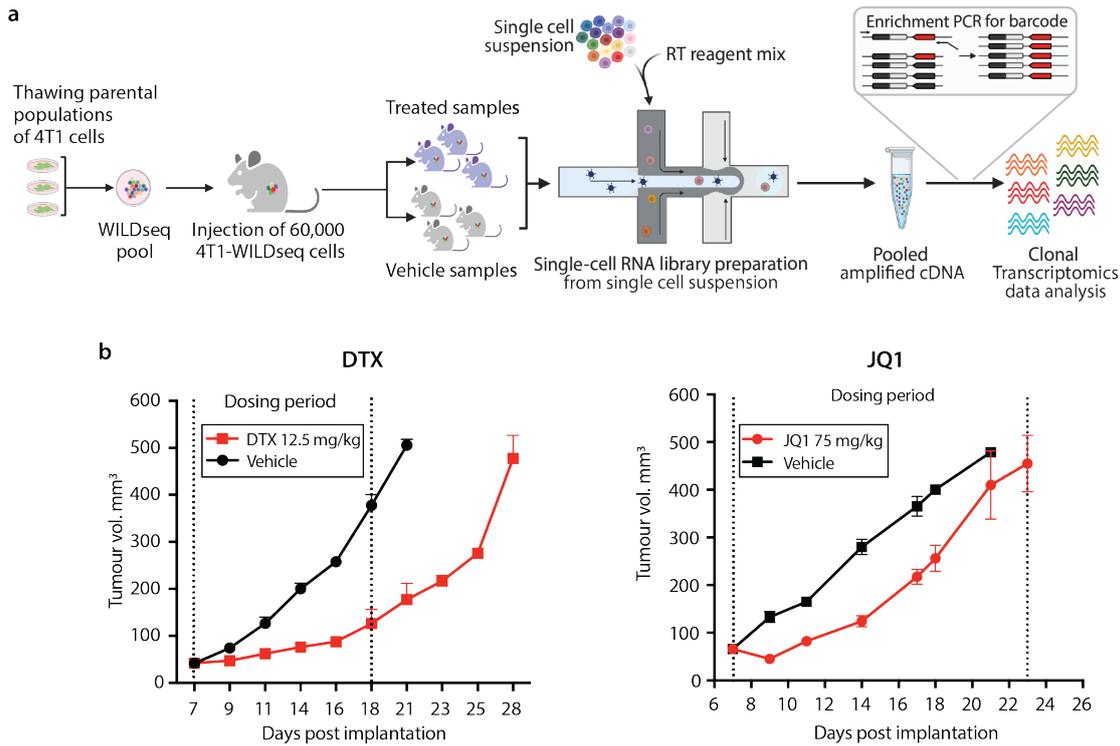
Differential gene expression between clones or conditions was assessed using Seurat's FindMarkers function with Wilcoxon rank sum test to generate  $P$  values and identify differentially expressed genes. P-values were generated on a per-sample basis and combined using the Fisher's Method. Combined p-values were adjusted for multiple comparisons using the Benjamini Hochberg correction method and filtered ( $p < 0.05$  and  $\logFC < 0$  for downregulated genes or  $\logFC > 0$  for upregulated genes). Pathway analysis was performed using the AUCell R package which enables analysis of the relative expression of a gene set (e.g. gene signature or pathway) across all the cells in the scRNA-seq dataset to calculate the enrichment of the input geneset within the expressed genes for each cell using the "Area Under the Curve" (AUC). Input genesets were taken from Molecular Signatures Database (MSigDB) C2 curated gene set collection v7.2 (Mootha et al., 2003; Subramanian et al., 2005) and filtered for signatures that contained more than 20 genes with detectable expression in our dataset. Each tumour cell got an AUCell Score for every signature in the MSig DB C2 collection assigned. AUCell scores were compared across clones or conditions using a Wilcoxon rank sum test and p-values were adjusted for multiple comparison using the Benjamini-Hochberg correction method. Combined p-values were further filtered ( $p < 0.05$  and  $\logFC < 0$  for downregulated genes or  $\logFC > 0$  for upregulated genes).

### 3.2.4 Analysis of baseline transcriptomic signatures

Baseline transcriptomic signatures were defined for the 5 major clones in 4T1 WILDseq tumours (clone 238, 350, 473, 679, and 684) using the vehicle datasets: 9 control, vehicle-treated animals administered either 10% DMSO in 10% beta-cyclodextrin for 5 days/week, 12.5% ethanol mixed with 12.5% Kolliphor diluted in saline for 3 days/week, or 12.5% DMSO in 0.5% sodium carboxymethyl cellulose for 5 days/week. Each clone was required to be represented by at least 20 cells in each sample to compare its gene expression to that of all WILDseq barcode assigned tumour cells. Comparisons were performed at gene level using Seurat or at the geneset level using AUCell. P-values of each sample were combined using the Fisher's method.

### 3.3 Results

To test our approach, we applied WILDseq to study the response of our 4T1 mammary carcinoma model to classic taxane-based chemotherapy and a first-in-class small-molecule inhibitor of BET bromodomains JQ1. We implanted approximately 300 WILDseq barcoded clones into immunocompetent mice (**Figure 3.1a**).



**Figure 3.1 WILDseq pipeline *in vivo* for docetaxel and JQ1 treatment.** **a** | Schematic illustration of WILDseq experimental workflow. **b** | Conditions of drugs were optimized to initially reduce tumour burden, but eventually results in tumour relapse and resistance to therapy. Drug treatment was started seven days post tumour initiation. Docetaxel was given three times per week at 12.5 mg/kg for a total time span of two weeks, while JQ1 was administered five times per week at 75 mg/kg until humane endpoint (n=3 per condition).

Specifically, we stabilized our three parental 250 cell populations in culture, sorted for WILDseq barcode expressing cells based on their zsGreen expression and pooled the three populations at equal ratios immediately before implantation to increase clonal competition *in vivo*. When tumours were palpable after one week, three mice were treated with either 12.5 mg/kg docetaxel for a total of two weeks or 75 mg/kg JQ1 (5 days on 2 days off) until humane end-point (**Figure 3.1b**). Control animals (n=9) were treated with vehicle formulations and tumours were collected at humane

end-point after 21 days. Whilst both treatments initially reduced tumour burden, the tumour ultimately overcame therapy and returned, indicative of selection and expansion of treatment-resistant clones. Tumour samples were dissociated, live-dead sorted to exclude debris and dead cells, and finally processed for single-cell RNA sequencing to analyse tumour composition. To increase the barcode detection, we prepared a separate barcode enrichment library from the pooled, amplified cDNA which was spiked into the transcriptomic library.

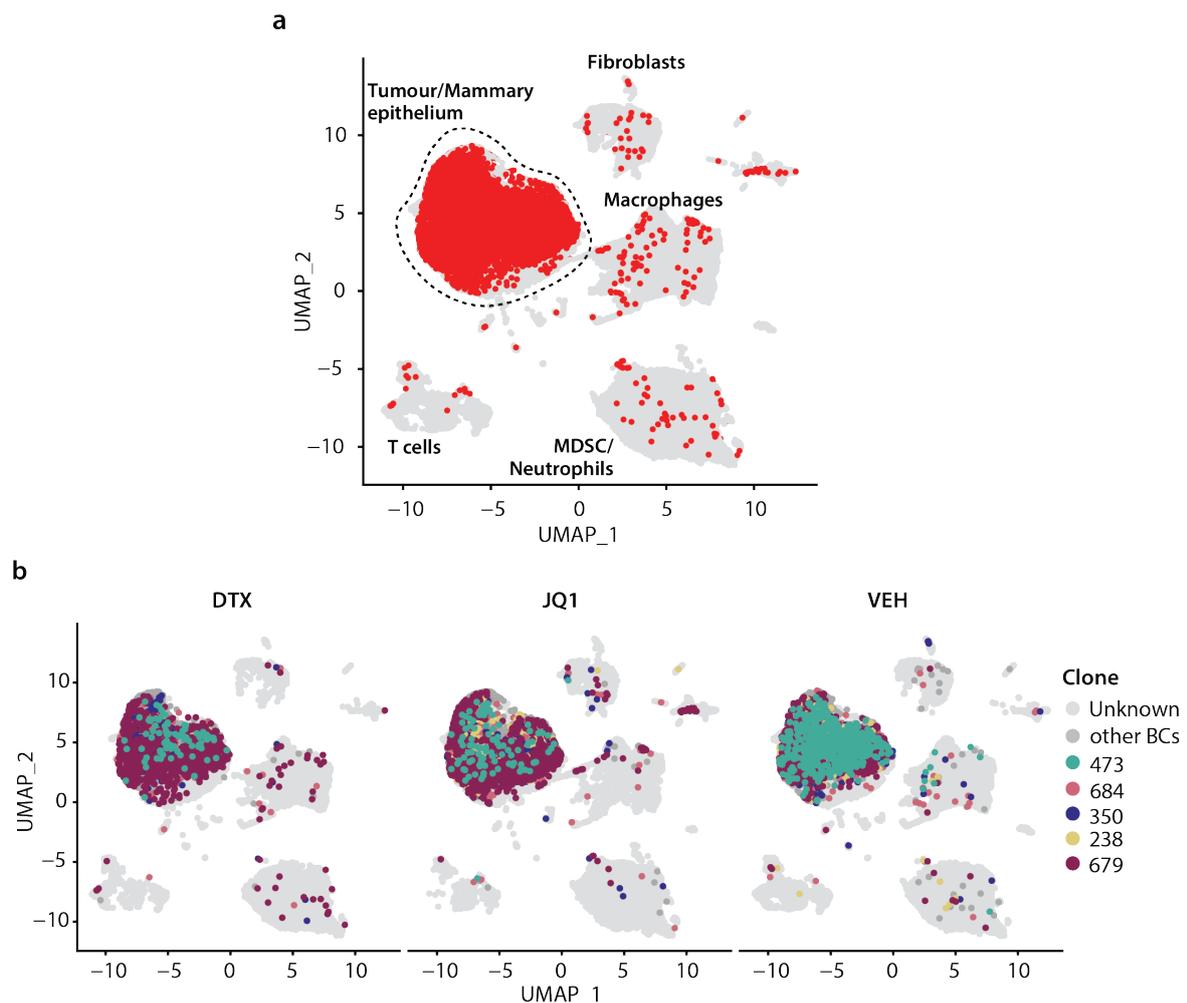


Figure 3.2 **WILDseq applied to study clonal dynamics under drug treatment *in vivo*.** **a** UMAP representation of 4T1 WILDseq tumours generated by injecting the 4T1 WILDseq pool into the mammary fatpad of Balb/C mice. WILDseq barcode expression is highlighted in red. **b** UMAP representation of WILDseq experiments clustered by treatment (VEH = all 9 vehicle treated samples) Three independent experiments were performed each involving injection into three separate host animals (docetaxel = 3 animals, JQ1 = 3 animals, Vehicle = 9 animals). Five major *in vivo* clones are highlighted in colours and other WILDseq barcode assigned cells are shown in dark grey.

In total, 64,426 single-cell transcriptomes were captured and resolved in 20 clusters of transcriptionally distinct cells including tumour cells as well as cells from the tumour microenvironment (**Figure 3.2a**). Visualization of WILDseq barcode expression was highly specific for tumour cell clusters as expected. We analysed the cell transcriptomes using the R package Seurat, including quality control, pre-processing, cell-cycle regression and the SCTransform based normalisation (Hao et al., 2021).

In order to fully characterise our 4T1 WILDseq pool, we compared the clonal distribution and gene expression profiles *in vitro* versus *in vivo* (**Figure 3.3**). We observed no correlation ( $R^2 = 0.004816$ , p-value = n.s.) between initial (*in vitro*) and final (*in vivo*) clonal abundance, suggesting that clone-intrinsic features conferring greater fitness *in vitro* do not necessarily translate in a survival benefit in the *in vivo* environment (**Figure 3.3a**).

Despite initially implanting  $\sim 132$  distinct clones, we only detected  $\sim 69$  clones *in vivo*, suggesting that only a subpopulation of cells were competent for engraftment and survival *in vivo*. While the *in vitro* pool was mostly dominated by two clones, 132 and 264, we observed five major clones *in vivo*: 238, 350, 473, 679 and 684. Downstream analysis in the tumour was focused on the characteristics of the five major clones *in vivo*.

When we compared gene expression signatures *in vitro* versus *in vivo*, we noticed great variation in the clonal gene expression profiles (**Figure 3.3b**). Most clones change dramatically with the setting and thus, are weakly correlated. In contrast, clone 679 and 684 exhibit more robust gene expression signatures across conditions reflected in a higher correlation coefficient. While clone 679 and 684 appear to have intrinsic mechanisms that define their gene expression patterns, most clonal signatures seem to be determined by external factors highlighting the importance to study clonal dynamics *in vivo* with an intact immune environment.

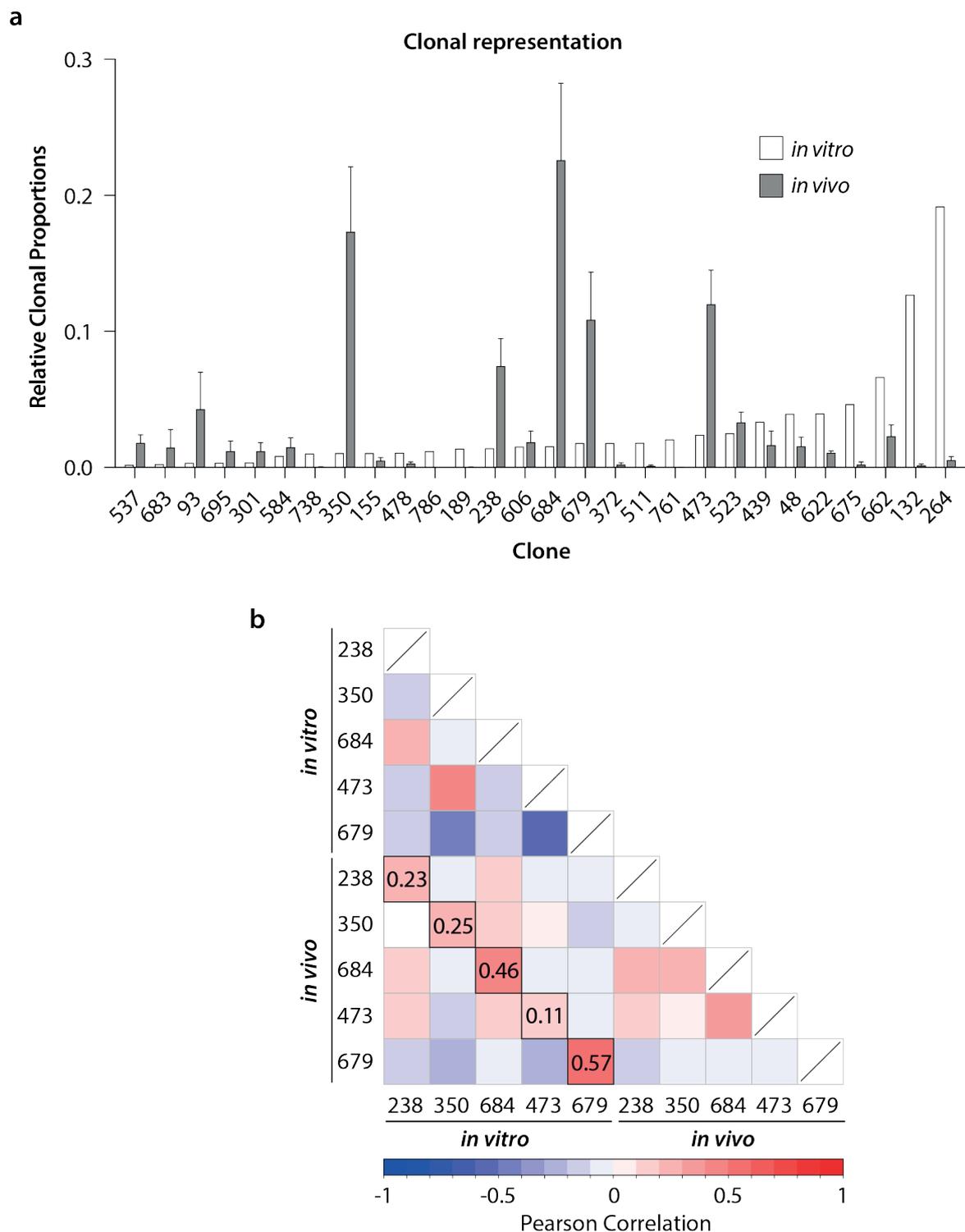


Figure 3.3 **Comparing WILDseq pool *in vitro* versus *in vivo*.** **a** | Clonal representation of WILDseq pool *in vitro* versus *in vivo*. (n = 1 for *in vitro* cultured cells; n = 9 for *in vivo* vehicle samples) **b** | Correlation of clonal gene expression signatures *in vitro* versus *in vivo*. Transcriptomic profile of each clone was compared to all other clones combined to generate an *in vitro* or *in vivo* clonal signature (logFC per gene) and illustrated as the Pearson correlation coefficient for each clone *in vitro* versus *in vivo*.

### 3.3.1 Validation of WILDseq in a the triple-negative breast cancer cell line D2A1-m2

To validate our approach in a second cell model, we selected the triple-negative breast cancer cell line D2A1-m2. We infected D2A1-m2 cells at a low MOI allowing only one WILDseq barcode per cell, followed by generating the D2A1-m2 WILDseq pool. Similarly to the 4T1 WILDseq pool, the D2A1-m2 WILDseq pool consisted of three 250 cell populations which were kept separately until pooling in equal ratios immediately before the tumour transplantation. We injected 60,000 D2A1-m2 WILDseq cells into the fourth mammary fat-pad of Balb/C mice and tumours were collected at human endpoint after 21 days. Tumours were processed into single cells, sorted for live-dead cells to get rid of debris and submitted for analysis via scRNA-seq.

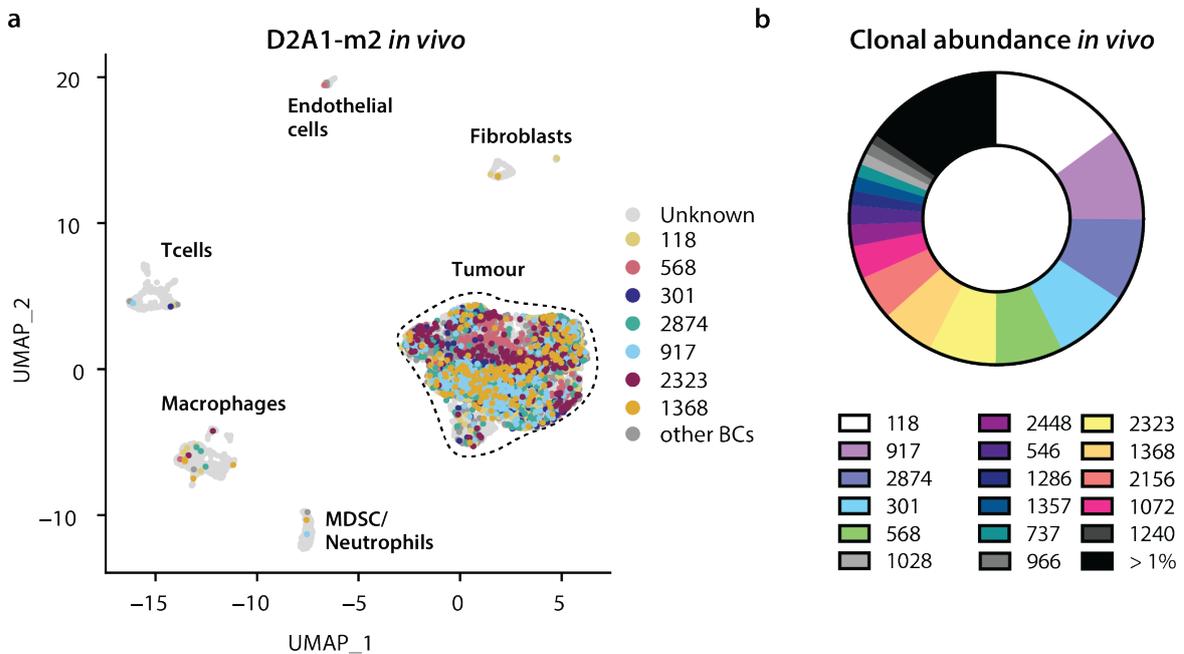


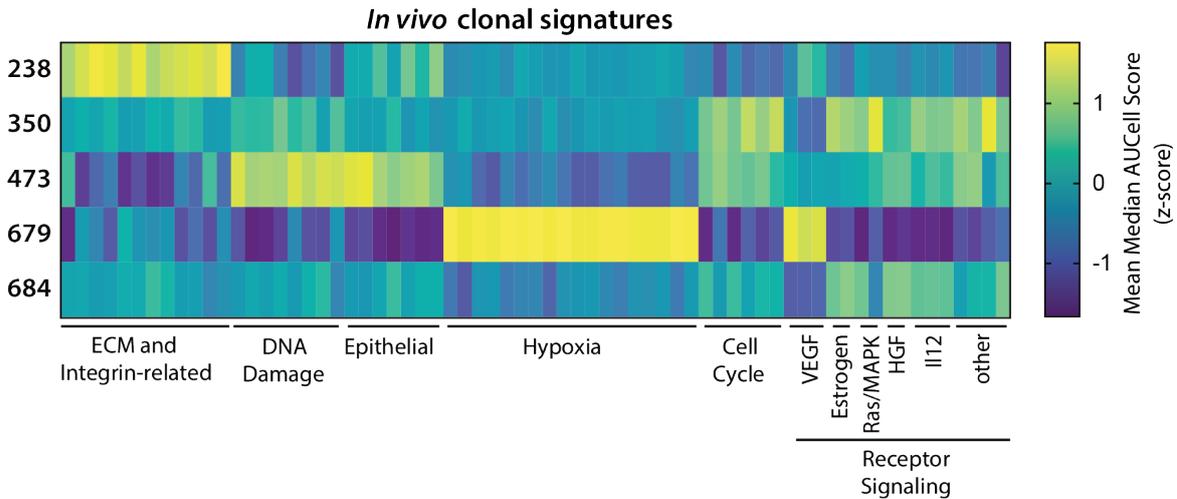
Figure 3.4 **WILDseq characterisation of D2A1 tumours.** a| UMAP representation of D2A1-m2 WILDseq tumour highlighting the seven major WILDseq expressing clones. ( $n = 2$  animals) b| Relative clonal abundance in D2A1-m2 WILDseq tumour. Individual clones are highlighted in different colours. Clones that are present at levels below 1% are illustrated together. In total, 89 different clones were observed across two independent samples with 9 clones contributing more than 70% to the total number of clones detected in the tumour. (Mean $\pm$ SEM)

We visualised the 7890 transcriptomes using the UMAP plot (**Figure 3.4a**). This revealed great differences in the tumour composition. In contrast to 4T1 WILDseq

tumours, D2A1-m2 WILDseq tumours mostly consists of tumour cells and only ~16% were assigned to cells of the tumour microenvironment. We next integrated the clonal assignment of the WILDseq barcode into the transcriptomic data: from the 3127 cells passing the threshold to support clone calling, we identified 89 unique clones on the basis of unique WILDseq barcodes (**Figure 3.4b**). This suggests that our lentiviral-labeling approach can be deployed in any organism or *in vitro* culture system that is susceptible to viral transduction to simultaneously profile transcriptomic and clonal identity at single-cell resolution.

### 3.3.2 Characterising baseline signatures of major clones *in vivo*

By analysing the transcriptome of these clones, we sought to comprehensively identify inherent transcriptional programs associated with drug sensitivity and resistance. We, therefore, performed differential expression analysis on the five most abundant clones *in vivo* using AUCell, an R tool for identifying gene signatures in scRNA-seq data. In detail, AUCell calculates the area under the recovery curve (AUC) across each single cell transcriptome to then determine whether a critical subset of the input geneset is enriched in the cell. Genes are ranked from highest to lowest value for each cell using the gene expression matrix from scRNA-seq dataset. Genes with the same expression value are randomly sorted. Based on the AUC, AUCell measures the relative biological activity of the input geneset in a given cell by estimating the proportion of genes in the input geneset and their relative expression level compared to other genes.



**Figure 3.5 Characterisation of basal transcriptomic signatures of major clones *in vivo*.** Pathway analysis of most abundant clones *in vivo* in nine vehicle samples using AUCell and 3801 input genesets. 67 genesets are highlighted and selected for their consistent and statistically significant enrichment in a specific clone across tumours. AUCell score of each clone was individually compared on per sample basis to all other clones with an assigned WILDseq barcode. The mean  $\Delta$ AUCell score across all tumour vehicle samples are plotted. A minimum of at least 20 cells of the clone of interest was set as a threshold for a vehicle sample to be included in the analysis (Clone 238: n = 5; Clone 350: n = 7; Clone 473: n = 7; Clone 679: n = 6; Clone 684: n = 8).

To define baseline signatures of our major clones *in vivo* (clone 238, 350, 473, 679, and 684), we computationally extracted all tumour cells expressing a WILDseq barcode in each vehicle sample and compared each clone individually versus the rest of the tumour to identify which gene sets were enriched or depleted (**Figure 3.5**). Overall, we analysed the expression of 3801 genesets with identifying 738 genesets as being significantly enriched and 764 genesets as being significantly depleted across our five major clones. To define consistently enriched or depleted genesets, we first calculated a p-value on a per cell basis and determined the combined p-value on a per-sample basis for each geneset. We then searched for genesets that represented the same phenotypic functional process and defined signatures for the five major clones *in vivo*. Each of the clones showed distinct gene programs that were specific to clones and for their regulation of functional pathways. For example, clone 473 was highly enriched in DNA damage-related genesets at base line. We, in particular, observed several ATM-related pathways which are activated upon DNA double-strand breaks. In contrast, clone 679 exhibited a distinct hypoxia signature, including upregulation of HIF and VEGF pathways, which might have been triggered by its tumour microenvironment *in vivo* or

represents an intrinsic mechanism. Moreover, clone 238 and 684 upregulated epithelial genesets suggesting a more epithelial, basal phenotype as opposed to clone 679 that exhibited more a mesenchymal phenotype with high levels of *Twist1*. This supports the idea that differences in 4T1 clones may be defined based their position on the epithelial-to-mesenchymal (EMT) axis and is in agreement with previous work that further identified the master regulator *Twist*, a repressor of E-cadherin expression, as being highly expressed and relevant for EMT in 4T1 cells (Yang et al., 2004).

### 3.3.3 Defining resistance and sensitivity signatures to docetaxel

A key advantage of our lineage tracer is in unambiguously cross-referencing tumour cells from the same lineage between samples and thus, allowing us to untangle transcriptional signatures of clonal subpopulations in treated versus control samples. As a proof-of-principle experiment, we examined the sensitivity of our 4T1 breast cancer model to the chemotherapeutic docetaxel, known to cause growth arrest and cell death in cancer cells by inhibiting depolarisation of microtubules. After profiling 22,771 transcriptomes by scRNA-seq, we integrated the WILDseq barcodes into the single-cell landscape. Around 30% of the sequenced tumour cells were robustly assigned with a WILDseq barcode (**Figure 3.6a**).

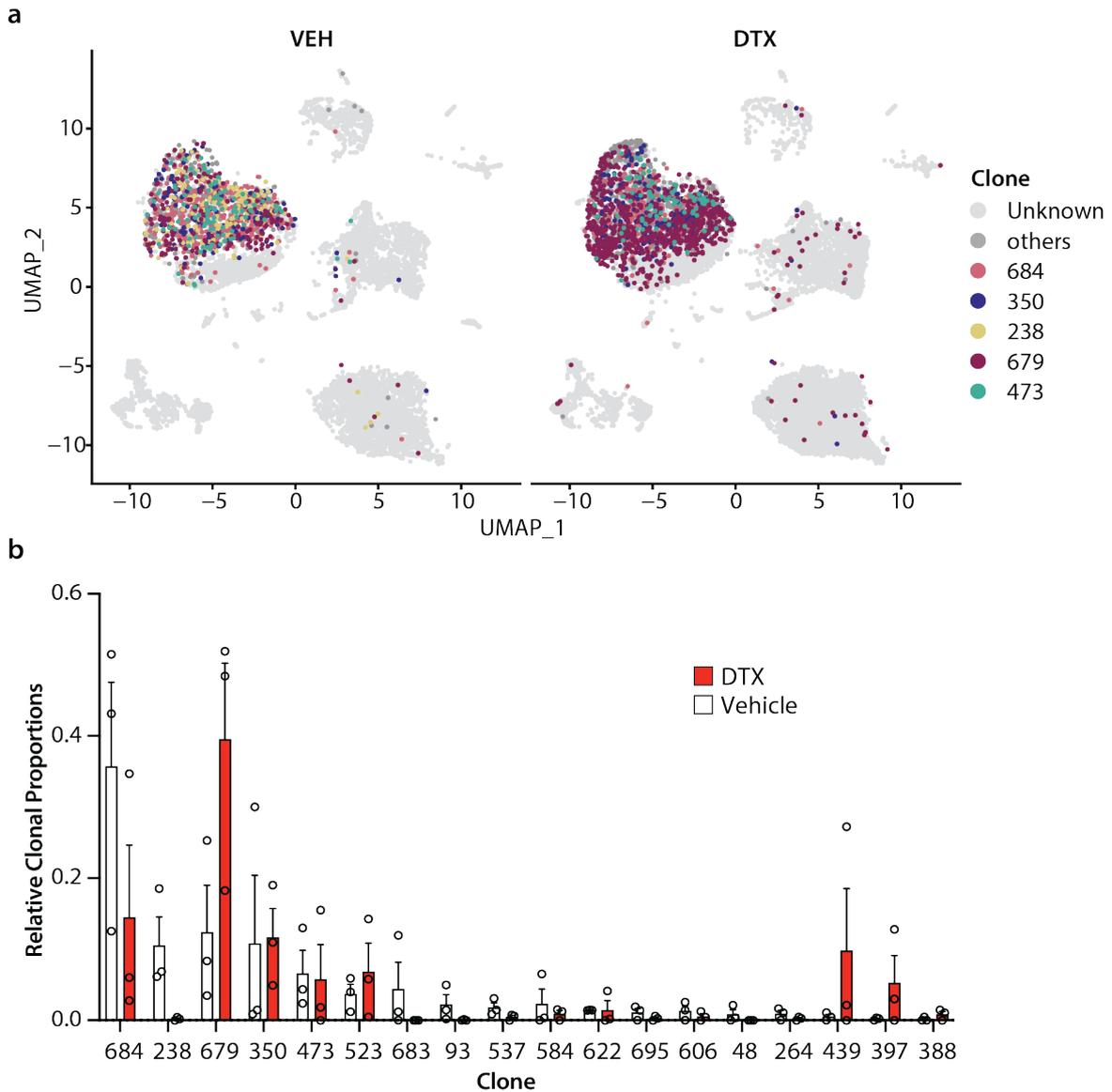
To understand what defines sensitivity and resistance mechanisms to docetaxel treatment, we pursued two analysis strategies:

1. Identification of differentially expressed genesets was performed using AUCell, and
2. Individual marker genes of resistance and sensitivity were identified using the FindMarker function in the R package Seurat.

In both cases, we compared each clone individually against all tumour cells with an assigned WILDseq barcode in each vehicle sample and calculated a combined p-value across samples to identify robust markers for resistance and sensitivity (p-value < 0.05 and  $\Delta < 0$  or  $\Delta > 0$  in every comparison). Notably, we only included vehicle samples that contained at least 20 cells of the clone of interest for the analysis.

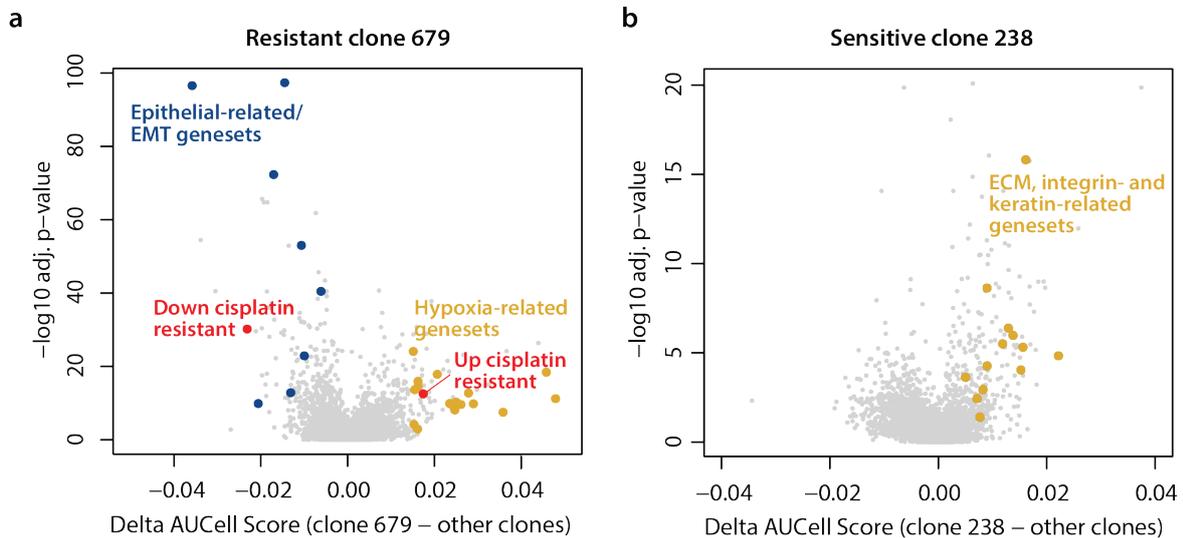
To identify sensitive and resistance clones to docetaxel, we first computationally extracted WILDseq barcode expressing cells from the tumour cell clusters and examined

clonal proportions in vehicle and treated samples. We then identified clone 238 as a docetaxel sensitive clone, while clone 679 dramatically increased upon treatment and therefore, marks an example of a resistant clone **Figure 3.6b**.



**Figure 3.6 Characterising clonal dynamics in 4T1 WILDseq tumours to docetaxel treatment.** **a**| UMAP representation of docetaxel-treated and control samples highlighting major WILDseq clones in colour and other clones with assigned WILDseq barcode in dark grey. **b**| Clonal proportions in docetaxel-treated versus control samples. Tumours cells were assigned a clonal lineage based on their WILDseq barcode expression. Threshold for clones shown was at least 20 cells across three animals in at least one condition. Clone 679 represents a resistant clone, while clone 238 was sensitive to docetaxel treatment. (Mean $\pm$ SEM)

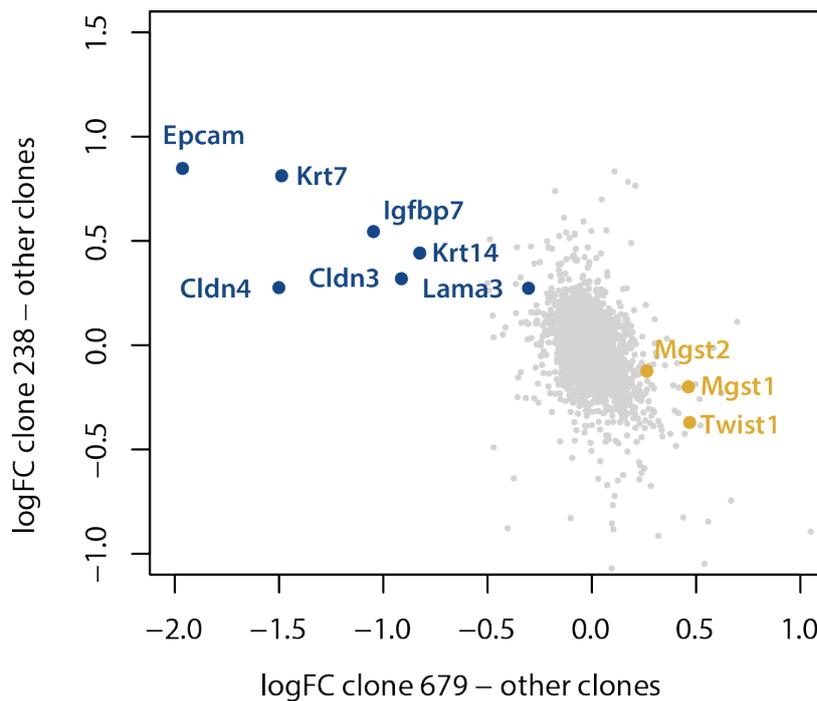
We first focused on the geneset analysis and compared the baseline signatures of clone 679 and clone 238. Clone 679 showed high levels of hypoxia-related pathways which have been associated with taxane resistance in the past. Many chemotherapeutic drugs share a similar mechanism of action eventually resulting in increased DNA damage. Resistance mechanisms, such as drug efflux pumps or increased levels of hypoxia, end up conferring resistance across various cytotoxic drugs. Interestingly, clone 679 showed upregulation and downregulation of genesets associated with cisplatin-resistance, thus further supporting our findings (**Figure 3.7a**). In contrast, clone 238 exhibited a more epithelial-basal like phenotype suggested by the downregulation of epithelial-to-mesenchymal (EMT) and epithelial-related gene sets and an upregulation of basal cytokeratins (**Figure 3.7a,b**).



**Figure 3.7 Clonal transcriptomic signatures of docetaxel resistance and sensitivity.** **a** | Geneset enrichment analysis of resistant clone 679 using AUCell. AUCell scores per cell were determined based on six control animal with at least 20 cells for clone 679. Geneset AUCell scores of clone 679 were compared to all other assigned clones within each tumour using a Wilcoxon rank sum test and combined p-values were calculated using the Fisher's method. The  $\Delta$ AUCell score was determined for each geneset comparing the median AUCell score of clone 679 to all other clones. The mean of this score is illustrated with the strongest and most significant positive and negative genesets annotated (gold and blue, respectively). Two independent genesets associated with cisplatin resistance were upregulated (red). **b** | Geneset enrichment analysis of sensitive clone 238 using AUCell. Analysis was performed as in a using data from five vehicle control animals for which at least 20 cells were identified for clone 238. Significantly changing genesets are highlighted.

Next, we performed the gene level analysis to identify specific marker genes associated with docetaxel resistance and sensitivity using Seurat. The gene level signatures

were negatively correlated between clone 679 and clone 238. Genes enriched in clone 679, including *Twist1*, *Mgst1* and *Mgst2*, represented markers for docetaxel resistance. Reciprocally, genes upregulated in clone 238, including *Krt14/17*, *Cldn3/4*, and *Epcam* were associated with docetaxel sensitivity (**Figure 3.8**). As expected, differentially expressed genes between clone 679 and clone 238 corresponded to signatures of epithelial versus mesenchymal-like phenotypes known to result in such a response to chemotherapy treatment.



**Figure 3.8 Correlation of differentially expressed genes in clone 679 and clone 238 versus all other clones.** Gene expression was compared between clone 679/clone 238 and the rest of the assigned tumour cells within each control animal with at least 20 cells of the clone of interest. The mean logFC across animals for each gene is plotted (Pearson correlation,  $R=-0.398$ ,  $p < 2.2 \times 10^{-16}$ ). Genes highlighted in gold are markers for docetaxel resistance (high in clone 679) and genes highlighted are markers for docetaxel sensitivity (high in clone 238).

Next, we confirmed our results in gene expression data of patients using ROC-plotter (Fekete and Györfy, 2019). Strikingly, the expression of our resistance and sensitivity signatures correlated with the outcome of patients treated with taxane-based chemotherapy (**Figure 3.9**). Patients that showed pathological complete response, defined as no evidence of residual disease post therapy, were high in genes associated

with our sensitive clone 238 and low in genes associated with the resistant clone 679. In contrast, non-responding patients were high in genes associated with the resistant clone. Together, these results demonstrated the power of our WILDseq pipeline in identifying signatures of resistance and sensitivity and highlighted the relevance of our findings as predictors for patient response to chemotherapy.

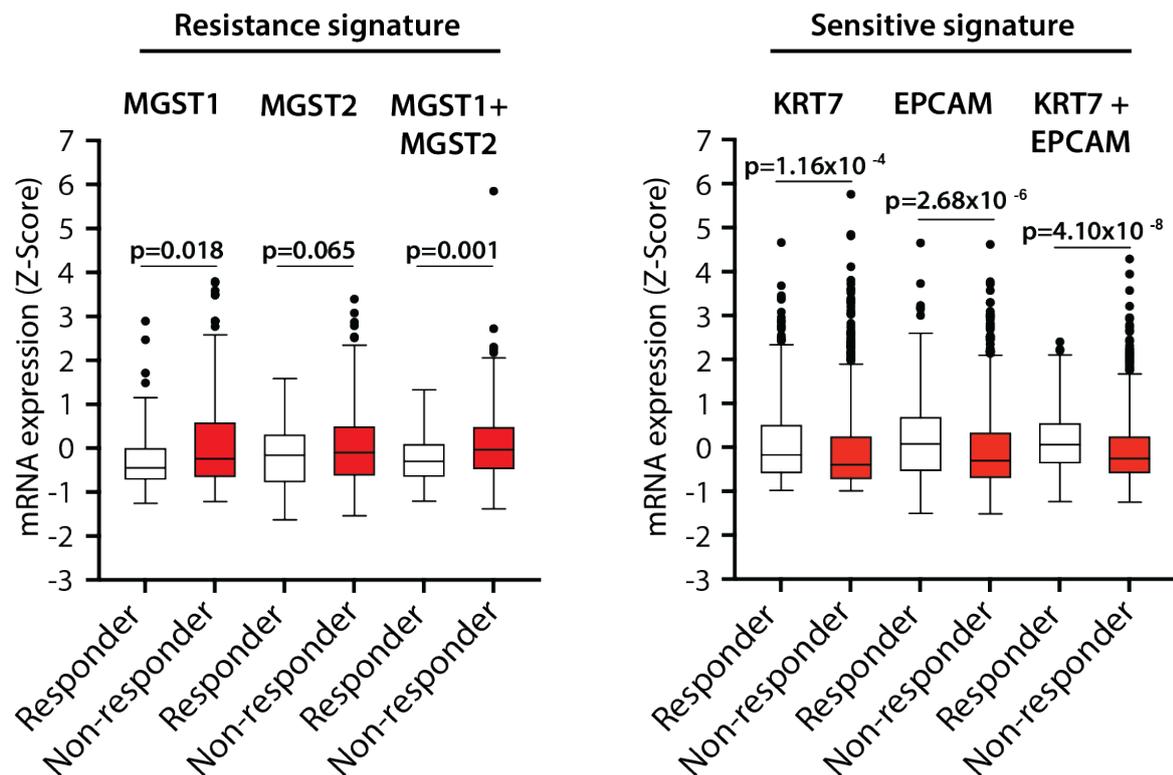


Figure 3.9 **Transcriptomic signatures of docetaxel resistance and sensitivity in patient treated with taxane-based chemotherapy.** Data set of breast cancer patients was retrieved from ROCplot.org (Fekete and Györfly, 2019) and consists of transcriptomic data of 3,104 treated breast cancer patient. Patients with response to docetaxel treatment showed upregulation in genes associated with the sensitive clone 238, while non-responding patients had high expression of genes associated with the resistant clone 679.

### 3.3.4 Defining resistance and sensitivity signatures to JQ1

Accumulating evidence suggests that a small subpopulation of cancer cells exhibiting acquired drug resistance do not necessarily require a stable heritable genetic alteration. Such findings suggest a reversible "drug-tolerant" state initiating through non-genetic mechanisms, such as epigenetic regulations. To this end, we wished to understand the anti-proliferative efficacy and the ensuing resistance to epigenetic cancer therapy with

JQ1. JQ1 is a selective small-molecule inhibitor of BET bromodomains that utilizes acetyl-lysine-competitive binding to displace BET bromodomains from chromatin, resulting in transcriptional changes and anti-proliferative effects (Delmore et al., 2011).

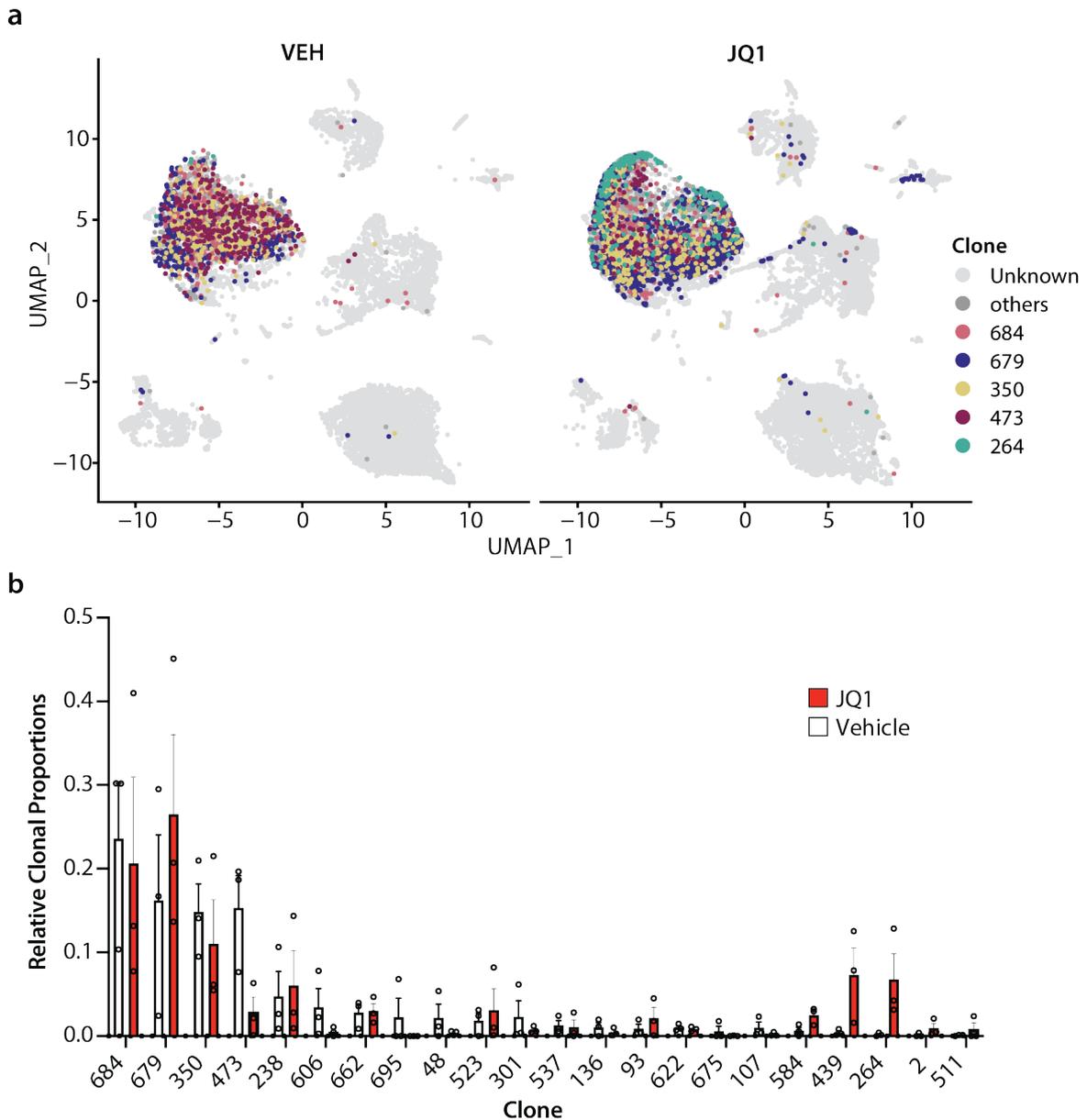
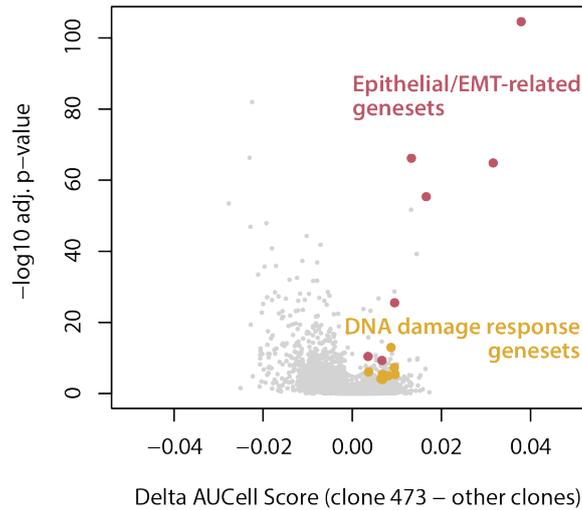


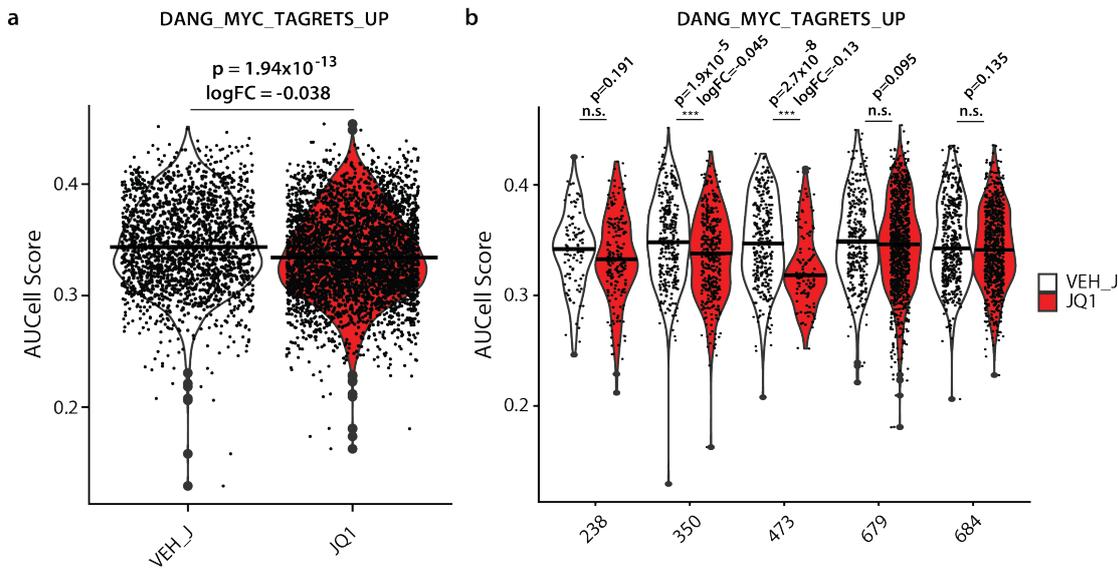
Figure 3.10 **Characterisation of clonal dynamics in JQ1-treated WILDseq tumours.** a| UMAP representation of JQ1-treated and control samples highlighting major WILDseq clones in colour and other assigned clones in dark grey. ( $n = 3$  per condition) b| Clonal proportions in JQ1-treated versus control samples. Clones with at least 20 cells across 3 animals in at least one condition are shown. Clone 264 represents a resistant clone, while clone 473 was sensitive to JQ1 treatment. (Mean $\pm$ SEM)

Single-cell profiling of JQ1 and vehicle treated tumours resulted in a total of 28,768 transcriptomes with nearly 50% of tumour cells expressing a WILDseq barcode (**Figure 3.10a**). Analysis of clonal distributions between conditions marked clone 264 as a resistant clone, while clone 473 represented an example of a sensitive clone (**Figure 3.10b**).



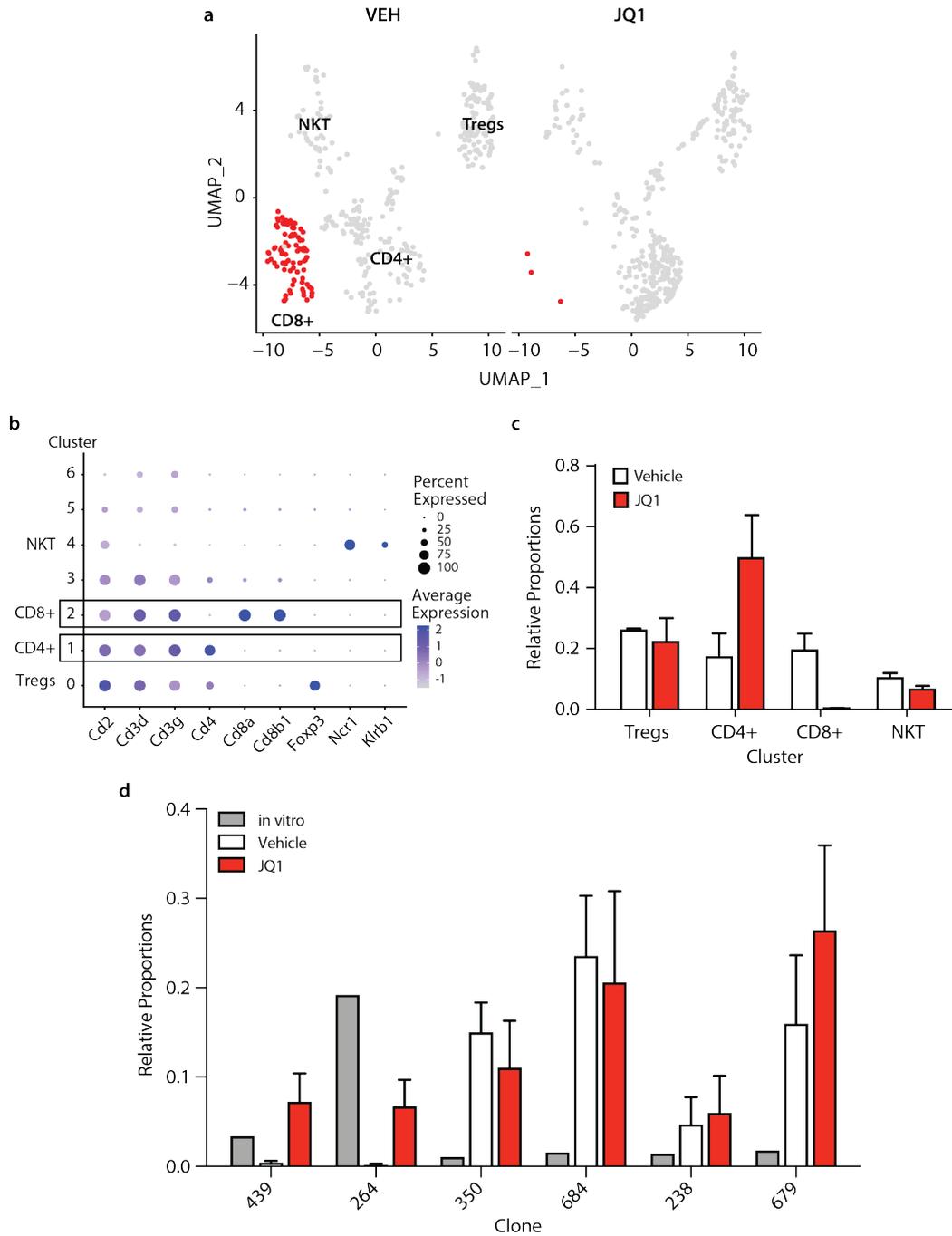
**Figure 3.11 Baseline signature of clone 473 primes it for JQ1 sensitivity.** Geneset enrichment analysis of resistant clone 473 using AUCell. AUCell scores per cell were determined based on seven control animal with at least 20 cells for clone 473. Geneset AUCell scores of clone 473 were compared to all other assigned clones within each tumour using a Wilcoxon rank sum test and combined p-values were calculated using the Fisher's method. The  $\Delta$ AUCell score was determined for each geneset comparing the median AUCell score of clone 473 to all other clones. The mean of this score is illustrated and significantly upregulated epithelial/EMT-related genesets (pink) as well as high levels of DNA damage response genesets (gold) are highlighted.

In order to understand what determines JQ1 sensitivity, we examined the baseline signature of clone 473 and found an enrichment of DNA damage repair pathways, including several ATM-mediated pathways, already in the absence of treatment (**Figure 3.11**). Clone 473 also exhibited high levels of epithelial and EMT-related genesets suggesting a more epithelial-basal phenotype.



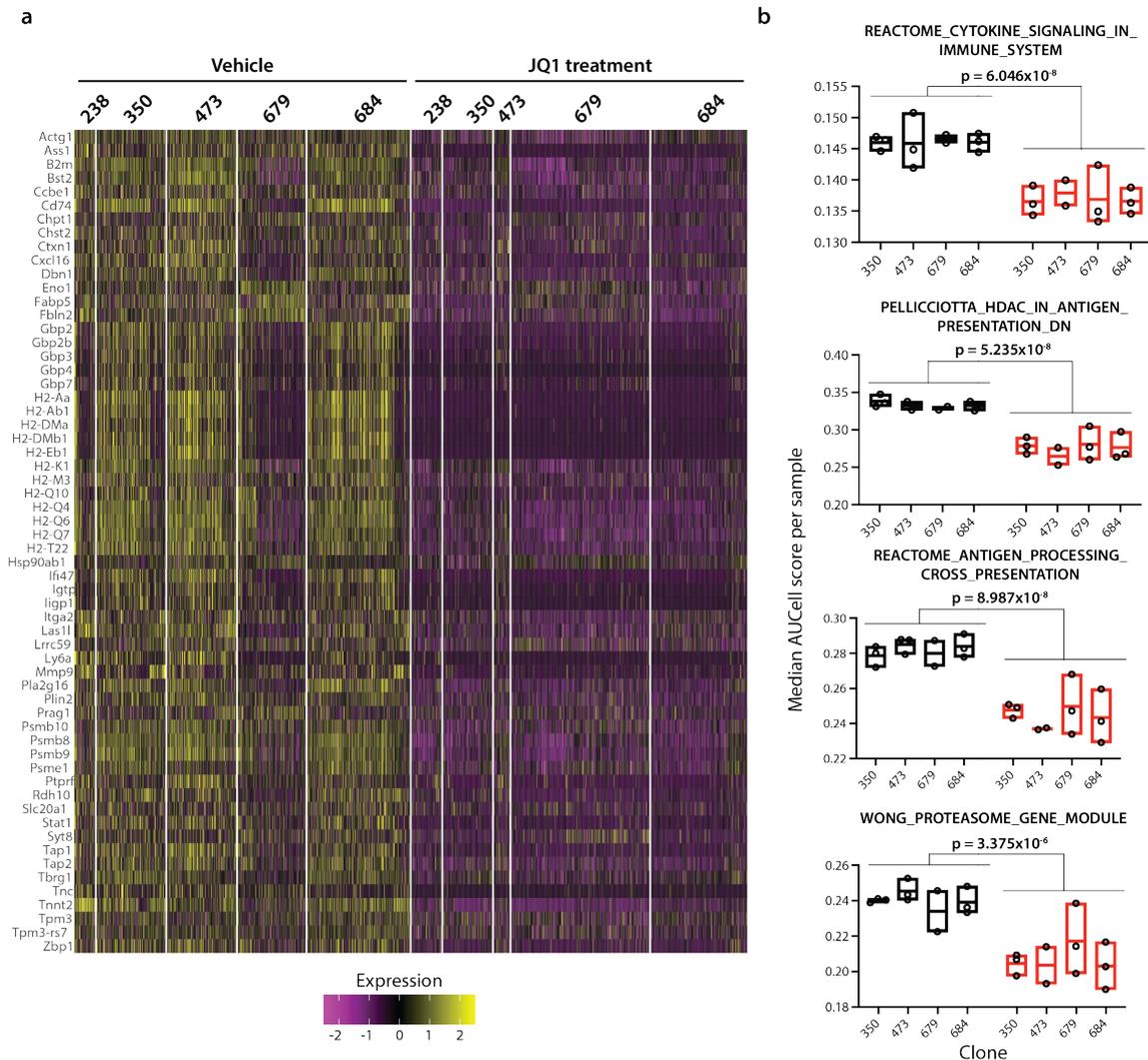
**Figure 3.12 Global downregulation of *MYC* targets in JQ1-treated tumour.** **a** | *MYC* target gene expression (in total 130 target genes in geneset) in vehicle versus JQ1-treated samples. AUCCell scores per tumour cell in vehicle versus JQ1 treated animals is plotted. LogFC from the median AUCCell score per condition is -0.038 and significance was tested using Wilcoxon Rank Sum Test. JQ1 treatment resulted in a downregulation of *MYC* target genes across all samples. **b** | Clonal response to JQ1 treatment on *MYC* target gene expression. AUCCell score per cell for assigned tumour cells in the most abundant clones are illustrated for JQ1-treated and vehicle samples. LogFC of the median AUCCell score per condition is plotted and significance was tested using Wilcoxon Rank Sum Test. Notable, detailed analysis on clonal level revealed dramatic suppression of *MYC* targets in clone 473 compared to other clones which exhibited a moderate inhibition. (n = 3 animals per condition)

Various studies have linked DNA damage as well as JQ1-treatment with *MYC* inhibition (Cannell et al., 2010; Porter et al., 2017). Further analysis of *MYC*-target gene expression revealed a global but modest downregulation of *MYC* target genes in JQ1-treated tumours (**Figure 3.12a**). Interestingly, analysis of *MYC*-target gene expression on clonal level revealed dramatic suppression of *MYC*-target genes in clone 473 specifically compared to the other major clones which showed a moderate downregulation of *MYC* target genes (**Figure 3.12b**). This observation suggested a combined effect of cells that have already been intrinsically primed to *MYC* suppression based on high levels of DNA damage which is then further increased by JQ1-induced DNA damage mediating profound *MYC* suppression (Delmore et al., 2011).



**Figure 3.13 JQ1 treatment results in specific depletion of CD8+ T-cells in breast cancer tumours.** **a** UMAP representation of T-cell cluster highlighting specific depletion of CD8+ T-cells in JQ1-treated samples. T-cell cluster was extracted from JQ1 dataset and reclustered. **b** T-cell subcluster identity was assigned based on gene expression of established markers. **c** Relative proportions of T-cell subtypes in vehicle versus JQ1-treated samples (Mean  $\pm$  SEM; n = 3 animals per condition). **d** Relative proportions of clones *in vitro*, in vehicle or JQ1-treated samples. Although clone 264 and 439 are dominant *in vitro*, they are only detected in JQ1 treated samples.(Mean  $\pm$  SEM)

Finally, we focused on unravelling the resistance mechanisms in clone 264. Despite being a dominant clone in the *in vitro* pool, clone 264 was only detected in JQ1-treated tumours *in vivo*, suggesting it is selected against *in vivo* under normal conditions. Moreover, we observed a specific depletion of T-cells upon JQ1 treatment which might be linked to the appearance of clone 264. Reclustering of the T-cell population and identifying T-cell subtypes based on gene expression of established markers revealed a specific depletion of CD8+ T-cells and an increase of CD4+ T cells in JQ1-treated samples (**Figure 3.13a-c**). The decrease in CD8+ T-cells reciprocally coincided with the appearance of clone 264 in the tumour (**Figure 3.13d**). Clone 350, 684, 238 and 679 were minor populations *in vitro*, but showed comparable levels *in vivo* in both conditions. In contrast, clone 439 and in particular clone 264 showed a higher representation *in vivo* with clone 264 being the dominant clone in the *in vitro* pool, but are only detected upon JQ1 treatment. Thus, we hypothesised that JQ1 treatment may result in transcriptional changes that prevent CD8+ T-cells from being recruited to the tumour resulting in an "immune cold" tumour that allows clone 264 to survive and expand. Different strategies have been reported that describe mechanisms of how tumour cells might prevent T-cell infiltration. Indeed, differential expression analysis of vehicle versus JQ1-treated tumours revealed global JQ1-dependent gene expression changes (**Figure 3.14a**). Many of the genes globally downregulated were part of the major histocompatibility complex (MHC) class I antigen presentation pathway, including TAP peptide transporters, proteasomal components and the beta-2 microglobulin subunit of the major histocompatibility class I complex (*B2m*) (**Figure 3.14a**). Our pathway analysis further agreed with our gene level analysis highlighting a significant downregulation of antigen representation-related pathways in JQ1-treated tumours (**Figure 3.14b**). These findings suggested a loss of antigen presentation as a possible mechanism that would clearly impact the T-cell response to a tumour.



**Figure 3.14 Global JQ1-dependent gene expression changes.** **a** | Heatmap of global JQ1-dependent gene expression changes. Gene expression of the five most abundant clones is illustrated. Analysis was performed with 1000 randomly selected cells of each clone per condition. Genes shown were significantly downregulated across all five clones in JQ1-treated animals versus vehicle animals. **b** | JQ1-dependent downregulation of antigen representation pathways. The median AUCell score per animal per clone is illustrated using tumours with at least 20 cells assigned to the clone of interest. Genesets were significantly downregulated across the four most abundant clones in JQ1-treated animals versus vehicle animals. P-values shown were calculated using a student t-test to compare 11 vehicle treated clones and 11 JQ1 treated clones.

One of the genes we were particular interested in was *B2m* which was heterogeneously expressed across the clones *in vitro* (**Figure 3.15a**). Notably, clone 439 and 264 showed the highest levels among all clones *in vitro*. We next performed pairwise differential

expression analysis of the most abundant clones *in vivo* comparing their expression in the vehicle tumours versus the JQ1-treated tumours.

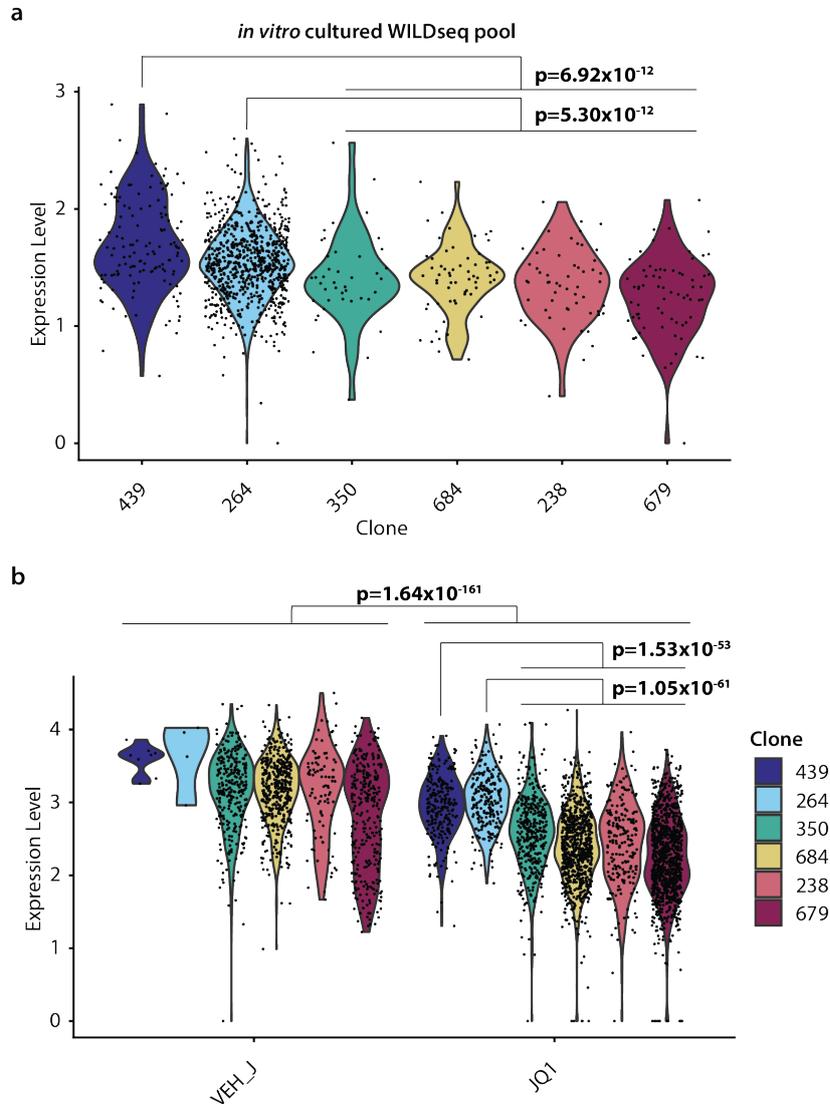


Figure 3.15 **JQ1 resistance correlates with  $\beta$ -2-microglobulin (*B2m*) expression.** a| *B2m* expression per clone *in vitro* or b| *in vivo* for most abundant clones in JQ1-treated tumours. p-values were determined using Wilcoxon rank sum test.

Consistently, clone 439 and 264 showed the highest expression levels of *B2m* in the vehicle as well as JQ1-treated samples. However, we also observed a global downregulation of *B2m* in JQ1-treated tumours across all clones. Thus, we hypothesised that the higher levels of the antigen representation machinery in clone 264 and 439 might prevent these clones from surviving under normal conditions *in vivo* due to their clearance through T-cells. In the presence of JQ1 treatment, however, antigen

representation is downregulated resulting in a cold tumour and allowing clones that would normally not be observed to grow out.

### 3.3.5 Discussion

Understanding intratumour heterogeneity, which underpins tumour evolution and chemoresistance, is a key challenge in cancer medicine. By applying our WILDseq platform to a syngeneic, heterogeneous mouse model of breast cancer, we tackled this problem and unravelled meaningful insights into the evolution of drug resistance through simultaneous single cell profiling of clonal lineage and gene expression. Among these key insights were the characterisation of clonally distinct transcriptomic signatures associated with differential responses to cytotoxic therapy, the identification of predictive biomarkers for docetaxel treatment to stratify breast cancer patients, and novel immune cell related resistance mechanism to JQ1 treatment.

We characterised distinct intrinsic gene expression programs *in vivo* linked with clonal sensitivity or resistance to treatment. Interestingly, we observed great variation of clonal transcriptomic profiles *in vitro* and *in vivo* implying a high dependency on external factors, including the tumour microenvironment. One of the strongest signatures upregulated upon docetaxel treatment in the resistant phenotype was HIF-1 regulated gene expression program induced by hypoxia. This is consistent with other studies that have reported the involvement of hypoxia in chemotherapeutic resistance in many cell lines through different mechanisms, including cell cycle arrest (Huang et al., 2010; Pucci et al., 2018), upregulation of drug efflux pumps (Samanta et al., 2014) or inhibition of apoptosis (Flamant et al., 2010). Moreover, we defined predictive biomarkers for docetaxel resistance (*Mgst1*, *Mgst2* and *Twist1*) and docetaxel sensitivity (*Krt7* and *Epcam*). *Mgst1* and *Mgst2* belong to the glutathione-S-transferase (GST) enzyme family that have been implicated in the development of drug resistance through their role in direct detoxification of oxidative stress as well as acting as a regulation of the MAP kinase pathway (Townsend and Tew, 2003). We speculate that the hypoxia signature at baseline in the resistant cells increased the levels of glutathione metabolism in order to regulate and detoxify reactive oxygen species (ROS) which in turn results in a protection mechanism from docetaxel induced damage.

Another marker for docetaxel resistance was *Twist1*, a master regulator of the epithelial-to-mesenchymal transition (EMT), a process through which cells adopt migratory and invasive behavior. Reprogramming of tumour cells through EMT has long been associated with chemoresistance (Bhang et al., 2015; Polyak and Weinberg, 2009). Moreover, enforced expression of Twist in breast cancer cell lines is sufficient to induce

an EMT and loss of E-cadherin expression suggesting that Twist might contribute to a more aggressive phenotype of invasion and metastasis (Yang et al., 2004). Conversely, chemosensitive cells expressed high levels of the canonical epithelial markers *Krt7* and *Epcam* representing the other end of the EMT spectrum. Gene expression differences in the 4T1 model are most likely to be largely attributable to their position on the EMT spectrum. Taken together, our work suggests that docetaxel resistance is mediated through intrinsic mechanisms that are associated with a more mesenchymal phenotype and ability to cope with high levels of oxidative stress, while cells with a more epithelial-basal phenotype do not survive treatment. Finally, we successfully confirmed our predictive signatures in breast cancer patient data highlighting the clinical relevance of our findings.

To our knowledge, this is the first report of using a syngeneic cancer model in an intact immune system to study the complex interplay between intrinsic clonal properties and their external factors, including the microenvironment, and how they cooperate to drive therapy resistance. One group of potential therapeutics in triple-negative breast cancer are BET bromodomain inhibitors, but acquired resistance to these inhibitors limit their potential in clinical use (Shu et al., 2016). Using WILD-Seq we found that cells sensitive to JQ1 were intrinsically primed with high levels of DNA damage, while JQ1 resistance was driven by micro-environmental factors associated with specific depletion of CD8+ T-cells. While we observed modest suppression of *MYC* signaling across clones, we observed a dramatic downregulation of *MYC* pathways specifically in JQ1 sensitive cells. We hypothesize that *MYC* target gene suppression might be a result of a combination of high levels of baseline DNA damage through double-strand breaks (Porter et al., 2017) and additional DNA damage induced by JQ1 treatment (Delmore et al., 2011), that synergise to repress *MYC* and kill cells. The nature of JQ1 resistance was mostly driven by micro-environmental factors rather than an intrinsic, primed state, such as for JQ1 sensitivity. The specific depletion of CD8+ T-cells coincided with a global decrease in antigen presentation after JQ1 treatment and the appearance of clones which have been only detected *in vitro*. These results suggest that decrease of antigen presentation in tumour cells results in alterations of T-cells and eventually lead in the depletion of CD8+ T cells, thereby allowing clones that usually are removed by the immune system to survive and expand. Future studies will focus on validating our proposed resistance mechanism by studying clonal dynamics in the absence of an intact immune system. This should allow clones that are usually

cleared out by the immune system to survive and be captured on the WILDseq platform.

These JQ1 data perfectly exemplifies a key advantage of WILDseq using a syngeneic mouse model with an intact immune system to capture the complex interplay between the tumor micro-environment and the clonal architecture and how these are remodelled by therapeutic intervention. Alternative approaches, such as CRISPR-based systems, will face certain difficulties in using syngeneic cancer models due to the need of delivering multiple components and thereby increasing the risk of immunogenicity or heavy silencing of barcode/vector expression resulting in clonal dropouts. Additionally, the extensive DNA damage caused by chemotherapy would be incompatible with a mutable barcode.



# Chapter 4

## CloneSTAR: Visualising clonal populations in space

*Due to the experimental demands of this project, this chapter would not be in my thesis without the supervision of Dr. Kirsty Sawicka. All of the experimental work presented in this chapter has been carried out by me. The CloneSTAR vector was designed and generated by me. Probe design and differential gene expression analysis for CloneSTAR clones was performed by Dr. Kirsty Sawicka. Sequencing was performed by the Genomics Core Facility. The analysis pipeline for decoding STARmap cycles was performed in collaboration with Eduardo Gonzales Solares from the Astronomy Department of the University of Cambridge.*

### 4.1 Introduction

Single cell genomics has proven to be a powerful tool for transcriptome-wide profiling of individual cells, providing information on cell types, developmental relations and gene programs. However since this method requires a dissociated single cell suspension as the starting material, it by necessity destroys information on the spatial localization of cells within the native tissue and their proximities to each other. Considering that cell signaling operates from 0 to 200  $\mu\text{m}$ , such spatial information is crucial to understand intercellular interactions and the spatial organisation of normal and diseased tissue (Longo et al., 2021). Spatial transcriptomic approaches address this challenge by physically measuring gene expression in intact tissue with the potential to achieve single cell or even subcellular resolution. Integrating scRNA-seq data and spatial transcriptomic measurements, has the potential to map phenotypic features into a

structural context. Currently, spatial transcriptomic tools are limited in their resolution and number of transcripts they can detect simultaneously in precisely localized single cells in tissues. However, they can elucidate niches enriched for distinct gene sets and cell types allowing cell subpopulations to be mapped into tissue neighbourhoods. When used in combination, scRNA-seq and spatial transcriptomics allow mapping of transcriptionally characterised single-cells to their spatial localization in the native tissue.

We apply STARmap, a novel sequencing-based technology for targeted 3D *in situ* transcriptomics, for the first time to tumour tissue. We demonstrate how scRNA-seq data can be used to guide probe design for spatial analysis of tumour heterogeneity and microenvironment and explore the possibility of incorporating lineage tracing via the WILDseq barcoding approach into our spatial transcriptomics method to simultaneously detect the spatial distribution of clones together with transcripts of interest.

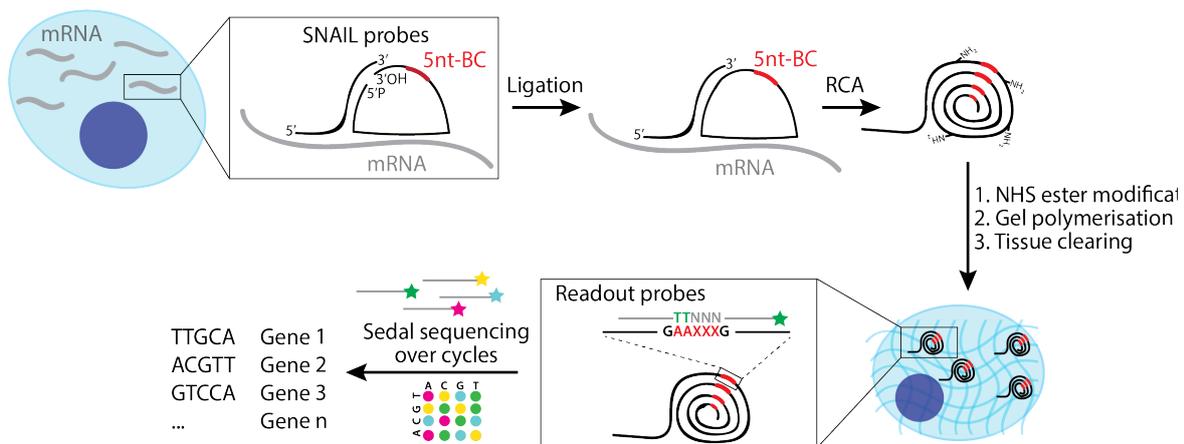


Figure 4.1 **Schematic illustration of STARmap protocol.** After tissue section or cells are prepared, custom SNAIL probes hybridize to intracellular mRNAs resulting in ligation of the padlock probe and followed by enzymatic amplification using rolling-circle-amplification (RCA) and amino-modified nucleotides. cDNA amplicons are then coupled to an acrylic acid N-hydroxysuccinimide (NHS) moiety allowing the copolymerization with acrylamid to form a hydrogel network, followed by clearance of unbound lipids and proteins. Each SNAIL probes contains a unique five nucleotide barcode which is assigned to a specific gene. Using *in situ* sequencing with two-base encoding for error correction (SEDAL) the barcode is read-out over several imaging cycles. Each imaging cycle is followed by probe stripping using 60% formamide before the next cycle begins. *Adapter from Wang et al. (2018).*

STARmap is a targeted *in situ* technology combining *in situ* sequencing and *in situ* hybridisation with the ability to measure up to 1,000 genes in tissue sections (**Figure 4.1**) (Wang et al., 2018). The method relies on the padlock probe system SNAiL (specific amplification of nucleic acids via intramolecular ligation) consisting of a primer and padlock probe pair to detect intracellular mRNA. Each padlock probe contains a five-base barcode designed as a gene-specific identifier. Notably, both probes are required to bind next to each other in order to initiate padlock probe circularization and rolling-circle amplification (RCA). Similar to other imaging-based methods, hydrogel-tissue chemistry is then used to embed the cDNA nanoball (amplicon) into a hydrogel network allowing to remove proteins and lipids for enhanced transparency. To reduce the error rates generated from single base readout, *in situ* sequencing of DNA amplicons is performed using SEDAL (sequencing with error-reduction by dynamic annealing and ligation), the sequencing-by-ligation method devised for STARmap. SEDAL relies on two kind of probes: reading probes to decode bases and fluorescent probes to convert sequence information into fluorescence signals. With each cycle the number of degenerated bases of the reading probe incrementally increase. Only if both probes are perfectly complementary to the DNA template, a stable product with high melting temperature can be formed via ligation, allowing later imaging after washes to remove unligated probes. Using 60% formamide probes are stripped after each imaging cycle for the next cycle to begin. To enable simultaneous detection of the clonal identity and transcriptomic profiles within the native tissue, we developed CloneSTAR, an inheritable barcoding system detectable with STARmap and scRNA-seq. We leverage scRNA-seq data to define transcriptional profiles corresponding to specific cell identities which we in turn used to design probe sets for STARmap. First, we established the STARmap protocol in breast cancer cells and tissue followed by testing our CloneSTAR method.

## 4.2 Material and Methods

### 4.2.1 Designing and cloning 40-mer barcodes for CloneSTAR

We generated 40-mer barcodes using the python software package FreeBarcodes (Hawkins et al., 2018). Barcodes were designed with a distance of 9 and the following experimental parameters: minimal GC content 55%, maximal GC content 70%, maximal homopolymer runs of 2 nt, and length 20 nt. The resulting barcodes were further filtered for barcodes with a melting temperature ( $T_m$ ) of  $60\text{ }^\circ\text{C} \pm 2\text{ }^\circ\text{C}$  resulting in a final pool of potential barcodes. To create 40-mer barcodes, we combined two 20-mer barcode sequences with one nucleotide in between and added common flanking regions that were homologous to the desired plasmid insertion site (Integrated DNA Technologies). The fragments were then introduced into the *Swa*I digested pHSW8 backbone using PCR and Gibson Assembly (NEB), followed by bacterial transformation and maxi-prep extraction of the plasmid DNA.

### 4.2.2 Creating barcoded clones for CloneSTAR validation experiments

The clonal cell lines 4T1-E, 4T1-F, 4T1-G, 4T1-J, 4T1-M, 4T1-R, and 4T1-T (Wagenblast et al., 2015) were cultured in DMEM high glucose and pyruvate supplemented with 10% FBS and 5% penicillin–streptomycin (Gibco). Each clonal line was infected with a lentiviral barcode and positive cells were sorted using the FACSARIA IIU cell sorter (BD Biosciences). The barcode of each individual clonal cell line was confirmed by reverse transcription of expressed barcodes and Sanger sequencing.

Table 4.1 CloneSTAR oligos

Name	Sequence
RT_primer	CAAGCAGAAGACGGCATAACGAGATCGTGATNNNNNNNNGTGACTG GAGTTCAGACGTGTGCTCTTCCGATCTCAAGCGATTCAAAGTTC TATCCG
Fwd_primer	AATGATACGGCGACCACCGAGATCTACACCAGCAGTATGCATGCG CTCGTTTACTATACGAT
Rev_primer	CAAGCAGAAGACGGCATAACGA
Seq_primer	CCAGCAGTATGCATGCGCTCGTTTACTATACGAT

### 4.2.3 Slide coating for cell and tissue experiments

Micro coverglasses or glass-bottom chambers were immersed with methacryloxypropyltrimethoxysilane (Bind-Silane; GE Healthcare) for 5 mins and washed in 100% ethanol. For tissue slices, glass surfaces were further treated with poly-L-lysine solution (Sigma Aldrich) for 5 mins and left to dry overnight. Prior to plating cells, glass surfaces were sterilized using 70% ethanol and washed three times with PBS.

### 4.2.4 Sample preparation

Cells were grown on pre-treated coverslips (#13 mm) or glass-bottom 24-well plates for 48hrs. After fixing cells in 4% PFA, cells were permeabilised in pre-chilled methanol at -20 °C for 10 mins and then placed at -80 °C for 15 mins prior to hybridisation or stored at -80 °C for up to one week. Breast tumour tissue samples were immediately embedded in O.C.T, snap-frozen with isopentane and stored at -80 °C. Tissue was cut into 16  $\mu\text{m}$  slices and mounted on pre-treated glass-bottom slides. Tissue slides were fixed with 4% PFA in PBS at room temperature for 10 mins, permeabilised with methanol at -20 °C and then placed at -80 °C for 15 mins prior to hybridisation.

### 4.2.5 STARmap procedure for cells and tissue slices

Glass-bottom plates and coverslips in plate were equilibrated to room temperature, washed with PBST-R (0.1% Tween-20 (Sigma-Aldrich), 0.1 U/ $\mu\text{L}$  SUPERaseIn (Invitrogen) in PBS) for at least 5 mins and pre-incubated with hybridisation buffer (2X SSC, 10% formamide, 1% Tween-20, 20 mM RVC and 0.1 mg/mL salmon sperm DNA) for 5 mins. SNAIL probes were dissolved at 200  $\mu\text{M}$  in UltraPure DNase/RNase-free distilled water (Invitrogen) and pooled to 25  $\mu\text{M}$  each. Probe solution was prepared by diluting 100 nM SNAIL probes in hybridisation buffer and added to the samples. The hybridisation reaction was incubated at 40 °C in a humidified oven overnight. Samples were then washed twice with PBST-R and once with 4X SSC/PBST-R for 20 mins at 37 °C in a humidified container. After rinsing the samples with PBST-R at room temperature, samples were incubated for two hours with T4 DNA ligation mixture (0.1 U/ $\mu\text{L}$  T4 DNA Ligase (Invitrogen), 1X Ligation buffer, 1X BSA (NEB) and 0.2 U/ $\mu\text{L}$  of SUPERase-In (Invitrogen)) at room temperature with gentle agitation. The samples were washed twice with PBSTR and then incubated with the RCA reaction mixture (0.1 U/ $\mu\text{L}$  Phi29 DNA polymerase (Invitrogen), 250  $\mu\text{M}$  dNTP mix, 1X BSA, 20  $\mu\text{M}$  5-(3-aminsoallyl)-dUTP and 0.2 U/ $\mu\text{L}$  of SUPERase-In) at 30 °C for two hours.

Following two washes in PBST-R, the samples were stored in PBST-R with 2  $\mu\text{L}/\text{mL}$  SUPERase-In overnight at 4 °C.

The next day, samples were washed with PBS and treated with 20 mM acrylic acid NHS ester (Stock concentration: 300 mM dissolved in anhydrous DMSO) in PBS for 2 hrs at room temperature. The samples were briefly washed twice with PBST and incubated with monomer buffer (4% acrylamide and 0.2% bis-acrylamide in 2X SSC) for 30 mins at room temperature. The monomer buffer was removed from the samples and the polymerisation mix was added to the centre of the sample and immediately covered with repel silane coated coverslips. The solution was then allowed to polymerise for 1 hr at room temperature. The coverslips were then gently removed and the samples were washed twice in PBST prior to incubation with digestion buffer consisting of 0.2 mg/mL Proteinase K and 1% SDS in 2X SSC. The sample was digested for one hour at 37 °C and then washed three times with PBST. The protocol was either continued at this stage or the samples were stored in PBS at 4 °C overnight.

#### 4.2.6 STARmap imaging and *in situ* sequencing

For single-gene/barcode detection, samples were incubated with a 19-nt fluorescent oligo at a concentration of 500 nM in 1X SSC/PBST for 30 mins at room temperature, followed by 3 washes with 10% formamide (Sigma Aldrich) in 2X SSC. Each sequencing cycle started with a stripping process: two rounds of stripping were performed with 60% formamide and 0.1% Triton-X-100 for 10 mins at room temperature, followed by three washes in PBST. The sample was then placed into the sequencing reaction (0.1 U/ $\mu\text{L}$  T4 DNA Ligase (Invitrogen), 1X T4 DNA ligase buffer, 1X BSA, 10  $\mu\text{M}$  reading probe and 5  $\mu\text{M}$  fluorescent oligos) for 3 hrs at room temperature. The sample was washed three times with 2X SSC and 10% formamide gently agitating and placed into fresh buffer prior to imaging. DAPI staining was performed following manufacturer's instructions every second cycle for the purpose of image registration. Images were acquired using Leica WLL SP8 confocal microscopy with a 405 diode, white light laser, 40x and 60x oil-immersed objective (NA 1.3).

#### 4.2.7 STARmap and CloneSTAR probe design

The appropriate transcript sequence per gene was selected from bulk RNA-seq data from 4T1 cells. In cases of multiple isoform expression, probes were designed against a shared sequence between expressed isoforms. Probes were designed using the program

Picky 2.2 (Wang et al., 2018) by dividing each transcript into two halves and up to five target sequences were identified for each half. Downstream analysis and sequence selection was performed in R. Each target sequence is 40-46 nt in length and was split into two oligos with a gap of 0-2 nt between probe sequences by selecting the pair with the best melting temperature match. Sequences were further filtered for GC content, repetitive sequences, closest non-target  $T_m$  and spacing along the transcript resulting in four final probe pairs per transcript. CloneSTAR barcodes were targeted by a single probe pair. Reading probes including six reading probes (R1 to R6) and twelve 2-base encoding fluorescent probes labelled with either Alexa546, 596 or 647 were purchased from IDT. For detection with a single probe, 19 nt detection probes (detect1) labelled with either Alexa546, 596 or 647 were purchased from IDT.

#### 4.2.8 Single-cell RNA sequencing and marker selection

***Cell preparation for scRNA-seq*** Cells and tissue samples were prepared as described in section 2.2.7.

***Bioinformatic analysis*** Read alignment, generation of expression matrix and downstream analysis in Seurat was performed as described in section 2.2.10.

***CloneSTAR barcode demultiplexing.*** Reads mapping to the CloneSTAR barcode sequence were extracted from the processed and filtered BAM file produced by Cell Ranger. The resulting BAM file was filtered for reads that show at least a 20 nt overlap with the 40-mer CloneSTAR barcode and exactly mapped to the expected sequence. Using a customised R script, only CloneSTAR barcode assignments supported by at least 2 UMIs and 80% of the barcode sequences from the cell were included for downstream analysis. The final barcode list was added to the metadata of the corresponding Seurat object for clone identification.

***CloneSTAR marker selection.*** Cluster identity was assigned based on CloneSTAR expression. Significantly enriched genes per clone were defined with a adjusted p value  $< 0.05$  and a log fold change  $> 0.1$ . Markers were selected from the resulting list using the following selection criteria: (1) Significantly enriched in clone cluster;(2) not significantly enriched in any other clone cluster; and (3) Log fold change  $> 0.3$ . Universally expressed genes were selected based on their expression in  $>70\%$  of cells across all clusters and for not being significantly enriched in any cluster.

***Tumour microenvironment marker selection.*** Significantly enriched genes per cluster were determined with a log fold change  $> 0.25$ . Markers were selected based on the intersect of significantly enriched genes and tumour microenvironment cell type

markers reported in literature (Bach et al., 2017; Weinreb et al., 2020; Zhang et al., 2019).

## 4.3 Results

In order to map the clonal composition of tumours in a spatial context, three possible approaches can be used:

1. Detection of genetic clonality via DNAFISH or transcriptional start sites
2. Detection of expressed barcode transcripts through spatial transcriptomics or
3. Detection of distinct gene expression signatures assigned to clonal identities

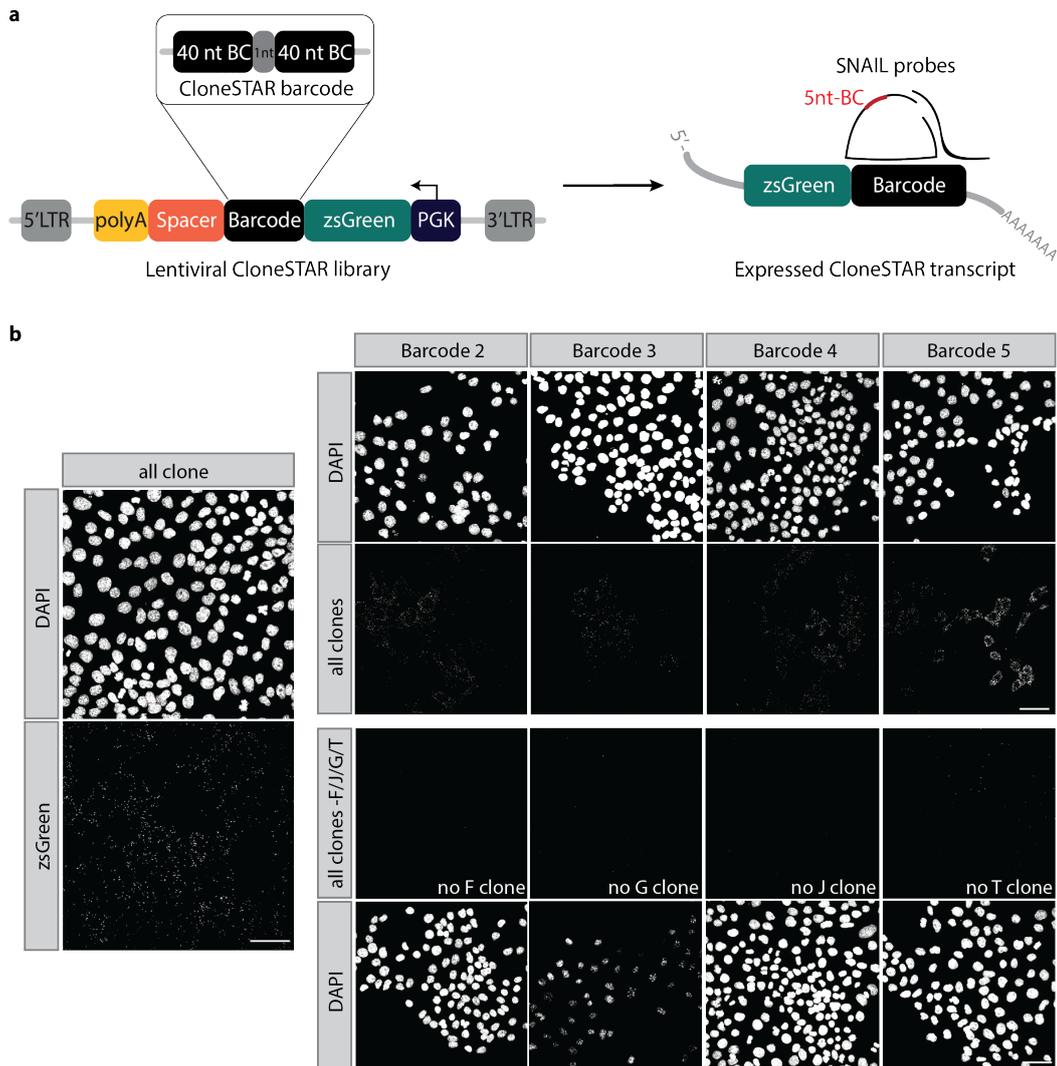
Here, we focused on establishing the identification of clonal niches via an expressible barcoding system in combination with STARmap. As an alternative, we explored the option of defining gene expression signatures specific to distinct clones using scRNA-seq to map clonal niches in tissue. We, therefore, leveraged our expressible barcoding system as a ground-truth dataset.

### 4.3.1 Design of CloneSTAR vector and detecting of CloneSTAR barcodes *in vitro*

To enable direct visualisation of clones and gene expression while preserving the spatial context, we adapted our WILDseq technology and made it compatible with STARmap. STARmap detection relies on detecting the transcript of interest using a pair of primer and padlock probes. This approach is termed SNAIL, for specific amplification of nucleic acids via intramolecular ligation. Each probe targets an area between 18-20 nucleotides (nt) and only when both probes hybridise adjacent to each other on the same transcript, the padlock probe can be circularised through ligation and amplified using RCA. In order to make our WILDseq barcode compatible with the SNAIL method, we designed the "CloneSTAR" barcode (**Figure 4.2a**). We replaced the WILDseq barcode cassette with a 40-mer barcode sequence consisting of two unique 20 nt sequences separated by one nucleotide allowing enough sequence space for SNAIL probes to bind. Each barcode was detected with one specific pair of SNAIL probes in contrast to genes which were detected using four probe pairs. Furthermore, we designed a single probe complementary to each barcode DNA amplicon allowing barcode detection without SEDAL. CloneSTAR barcodes are expressed in the 3'UTR of the zsGreen reporter fluorophore (CloneSTAR transcript).

We first set out a system with known transcriptomic differences in order to test CloneSTAR. We analysed bulk RNA sequencing data of 23 clonal lines derived from the heterogeneous 4T1 breast cancer model, previously generated in our laboratory and selected four distinct 4T1 clones - F, G, J, and T - based on their transcriptomic differences (Wagenblast et al., 2015). The four clones were most transcriptomically diverse and thus, should capture the full transcriptomic range allowing us to establish our CloneSTAR system in the context of the 4T1 model. Besides establishing CloneSTAR barcode detection, we also used the system to test whether we could define specific gene expression patterns to identify clones as an alternative approach.

We infected each of the clones with a unique CloneSTAR barcode - F clone with BC2, G clone with BC3, J clone with BC4 and T clone with BC5 - and sorted based on the zsGreen expression. To test whether we can detect individual clones, we applied CloneSTAR to a mixed population of all clones (**Figure 4.2b**). Briefly, cells were plated in equal ratios as a mixture into Bind-Silane coated 24-well plates and grown for 48 hrs prior to fixation. We performed the STARmap protocol using SNAIL probes designed to target either a constant region in the CloneSTAR transcript constant across barcodes allowing to visualize all cells expressing our barcode transcript, or a specific SNAIL probe pair against each CloneSTAR barcode. We further included a condition for each clone-probe pair, where we omitted the corresponding clone to rule out cross-reactivity and check for specificity of our SNAIL probes.



**Figure 4.2 CloneSTAR barcode design and testing barcode detection *in vitro* in 4T1 clones - F, G, J, and T.** **a** | Schematic illustration of CloneSTAR vector. The 40mer CloneSTAR barcode, which consists of two 20mer cassettes separated by one nucleotide was cloned into the pHSW8 vector backbone. **b** | Testing barcode detection in 4T1 clones F, G, J, T. In a 24-well plate, cell clones were plated and cultured together. After two days, cells were fixed and CloneSTAR barcodes were detected using STARmap. A control condition was included for each clone-probe pair where the concatenated clone was omitted to rule out cross-reactivity of our SNAIL probes. All four barcodes were detected and no cross-reactivity between barcodes was observed. Probes targeting zsGreen visualised all BC expressing clones independent of their barcode sequence. Scale bar represents 50  $\mu\text{m}$ .

We found that the CloneSTAR barcode transcript was detectable in the majority cells when using a single probe targeting a constant region of the barcode transcription. We observed various expression levels of CloneSTAR transcripts between clones. Moreover, we were able to identify the different clones based on barcode detection without

observing any cross-reactivity between barcode and probe pairs using our CloneSTAR technology.

### 4.3.2 Visualising differential gene expression profiles of clones with STARmap

To define marker genes specific for each clone, we analysed scRNA-seq data derived from a mixed population of the four clones (**Figure 4.3a**). CloneSTAR barcodes are expressed and readily captured within each single-cell transcriptome, enabling clonal identification. We assigned clusters to the individual clones (F, G, J and T) based on their barcode expression (**Figure 4.3b**).

As expected, cultured cells clustered based on their clonality. Consistent with our prior CloneSTAR experiment, we observed variable levels of barcode expression and/or detection between the clones (**Figure 4.3b**) with highest expression being detected in the T clone (31.7%), F (34.7%) and J clone clusters (30.1%), while the G clone cluster showed the lowest barcode expression (10.5% barcodes). Overall, barcodes were assigned to 23.7% of the cells. Using the assigned clusters, we selected specific marker genes for each clone. We defined genes significantly enriched per cluster and selected specific markers for each clone based on the following criteria: Genes (1) significantly enriched in clusters of clone of interest, (2) not significantly enriched in any other cluster, and with (3) a log-fold change  $> 0.3$  between clusters of clone of interest and the rest. From the genes matching these criteria we selected a final list of clonal markers, which defined gene expression signatures specific to each clone, consisting of 16 genes (**Figure 4.3c**) and four clonal barcodes. We further included four genes (Rack1, Hsp90b1, Eif2f and Cdk4) universally expressed across all clones and not significantly enriched in any clone.

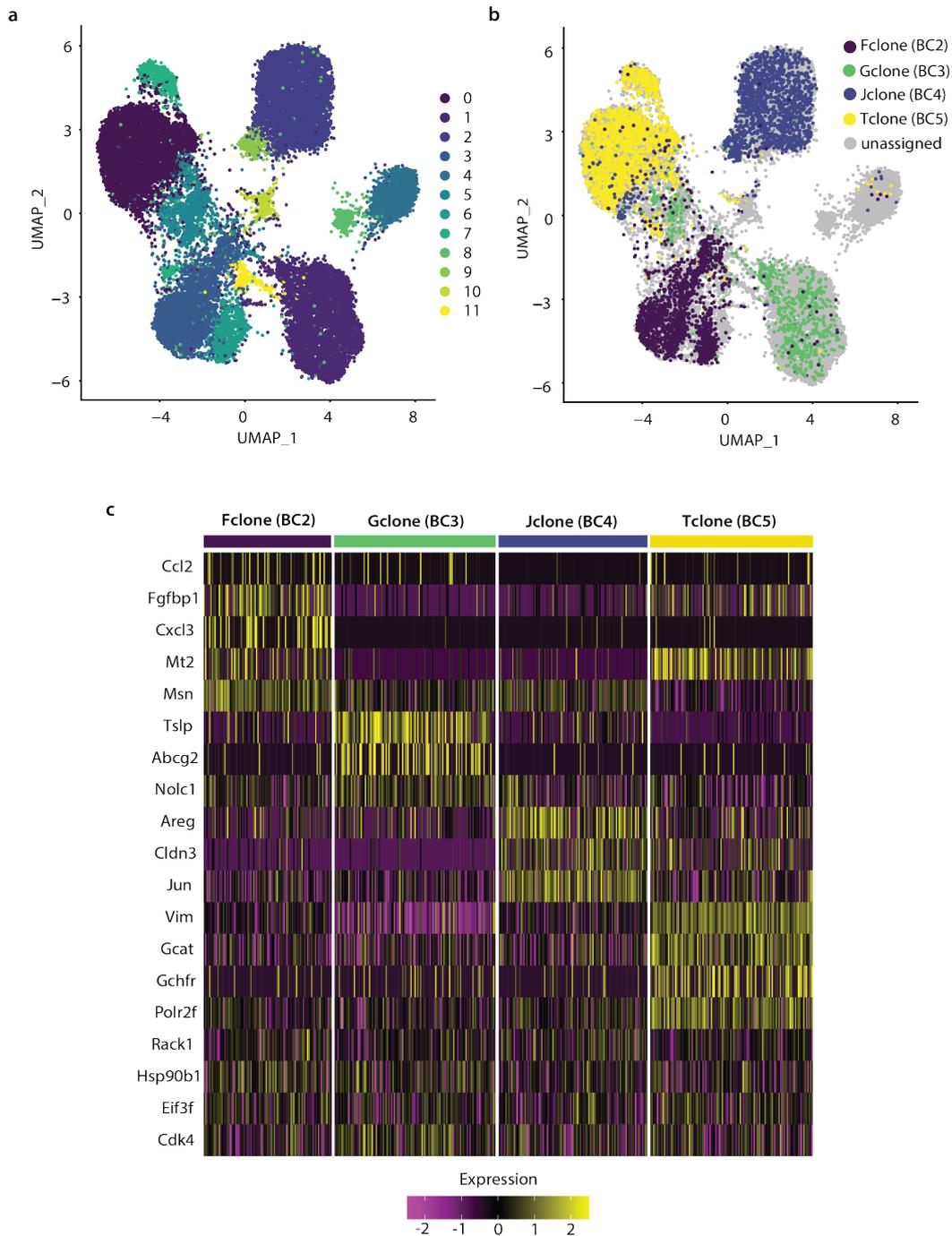
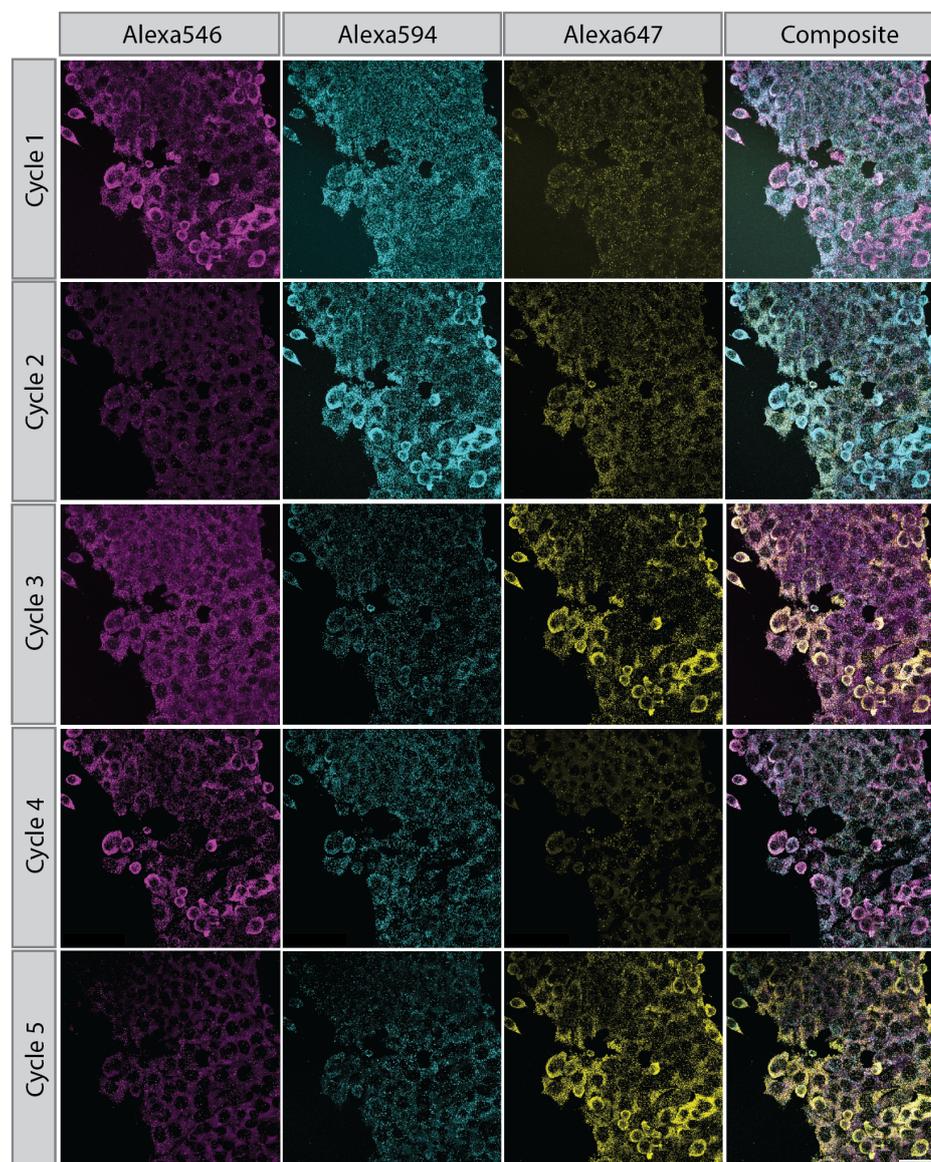


Figure 4.3 ScRNA-seq of CloneSTAR labelled 4T1 cells *in vitro* and selection of **clonal markers**. UMAP representation of CloneSTAR scRNA-seq dataset highlighting the different clusters (**a**) and the expression of the CloneSTAR barcode transcripts (**b**). **c** Gene expression heatmap representation of markers for each clone. Genes shown were selected based on only being significantly enriched in the cluster of interest and not in any other cluster with a  $\log_{2}FC > 0.3$ . Rack1, Hsp90b1, Eif3f and Cdk4 were selected as universally expressed genes present in more than 70% of the cells and not significantly enriched in any cluster.

To investigate whether the gene expression correlated with the barcode expression, we generated a probe library of the 16 clonal marker genes and four CloneSTAR barcodes and performed CloneSTAR (**Figure 4.4**).



**Figure 4.4 Detection of 16-gene library and CloneSTAR barcodes *in vitro*.** Fluorescent images of STARmap across all six cycles. All four clones were plated in a mixture and full six cycles of STARmap were applied using a 16 gene library with clonal marker genes. Barcodes were detected with single probes at the end as well as through SEDAL. Gene transcripts were read out over five cycles using SEDAL. Scale bar represents 50  $\mu\text{m}$ .

Following the five readout cycles in three channels to detect the 16-gene library, CloneSTAR barcodes were visualised with single probes. The identify of each RNA

species is encoded as a five-base barcode manifested in the DNA amplicon and read out through several cycles of hybridisation - ligation - imaging - stripping. Ligation of reading and decoding probes only takes place if both probes are perfectly complementary to the DNA template resulting in a stable product, allowing unligated probes to be washed away. After each imaging cycle, probes are stripped from the tissue-hydrogel for the next cycle to begin. In contrast to the original STARmap protocol, we only used three channels (546, 594, 647) due to high background fluorescence in channel 488. To align the images acquired in different cycles and for cell segmentation, we stained the nuclei with DAPI.

We are currently optimising the analysis pipeline in order to decode the gene expression across the five cycles (**Figure 4.5c**). Nonetheless, we clearly demonstrated that we can do multiple cycles of hybridizing and stripping without any bleed-through from previous rounds or between channels. To detect 16 genes plus four barcodes in three colours, seven rounds would have been necessary using a standard smFISH scaling linearly. In contrast, CloneSTAR only requires three rounds to readout 27 genes of interest with no redundancy. Our current approach, using three channels and five cycles, has the potential to measure the gene expression of 240 genes of interest. Notably, we did not observe an appreciable decrease in signal through the cycles. Preliminary analysis revealed distinct expression of genes in different clones, as expected. For example, T clone markers, all encoded in channel 546 in cycle 1, nicely aligned with the barcode expression (**Figure 4.5**). This offers the potential of using pre-defined gene expression signatures to identify clones in a spatial context.

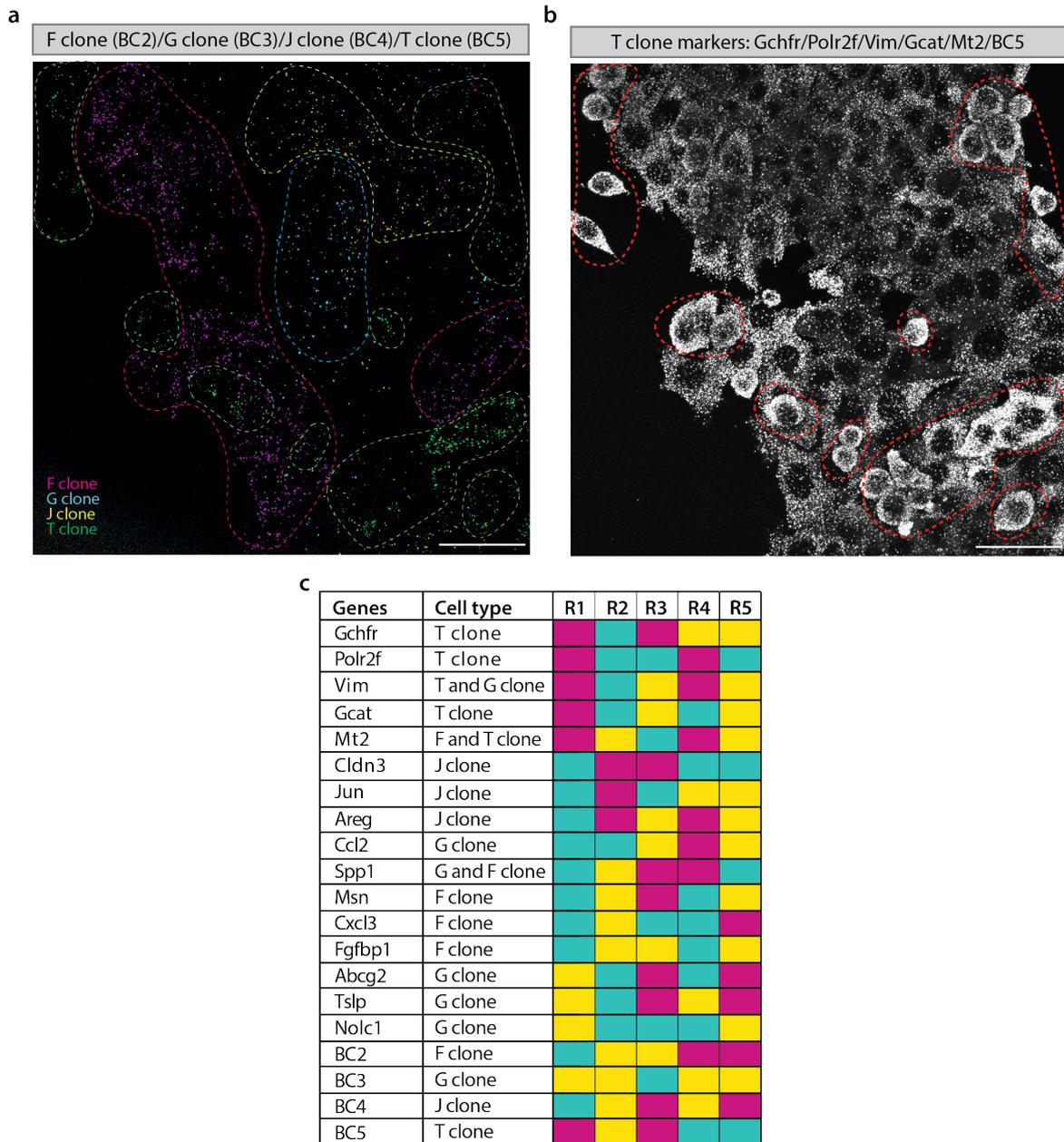


Figure 4.5 **CloneSTAR barcode detection aligns with clonal marker genes *in vitro*.** **a|** Image shows barcode detection highlighting clonal populations. **b|** T clone marker gene expression (Gchfr, Polr2f, Vim, Gcat, Mt2 and BC5) in the first cycle of 16-gene set library in the 546 channel. T clone barcode expression aligns with expression of markers genes. Scale bar represents 50  $\mu\text{m}$ . **c|** Code book for 16-gene library and barcodes. Colours highlight the different channels: 546 = magenta, 594 = cyan and 647 = yellow.

### 4.3.3 Establishing STARmap in breast cancer tissue

We next wished to establish the protocol *in vivo* in 4T1 breast cancer tumours. Tumours harbour a highly heterogeneous tumour microenvironment with complex, spatially restricted interactions with the immune system. Cells of the tumour microenvironment have been extensively studied using scRNA-seq and *in situ* hybridisation, yielding well-defined transcriptional markers to identify cellular identities (**Figure 4.6a**).

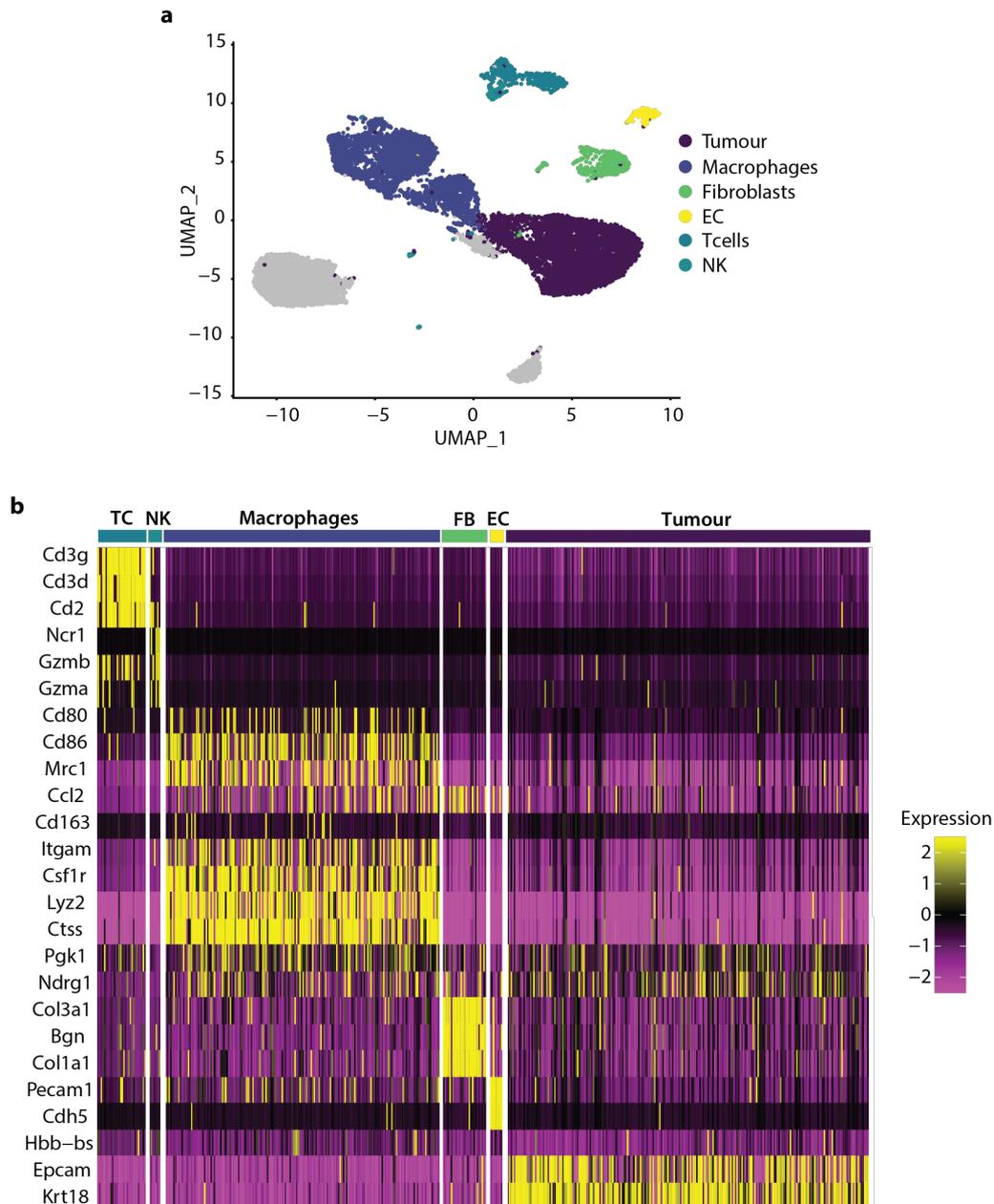


Figure 4.6 **Selection of tumour microenvironment markers.** a| UMAP representation of 4T1 tumour scRNA-seq dataset highlighted by cell phenotype. b| Gene expression heatmap representation of key marker genes used to identify cells of the tumour microenvironment. EC=Endothelial cells, FB=Fibroblasts, NK=Natural killer cells, TC=T-cells.

To this end, we performed scRNA-seq of our 4T1 breast tumours and compared lineage marker expression in order to assign cell identities in 4T1 breast tumours. The resulting cell phenotypes fell broadly into the categories of tumor, stroma and immune cells. We selected specific markers for T-cells, macrophages, endothelial cells,

fibroblasts as well as tumour cells (**Figure 4.6b**) and also included markers for hypoxia, which is a prominent feature in 4T1 tumours.

Table 4.2 Tissue pre-treatment conditions tested on 4T1 breast tumour tissue

Name	Step1	Step2	Step3
None	100% Methanol for 10 mins at -20 °C and 15 mins at -80 °C	None	None
SDS	100% Methanol for 10 mins at -20 °C and 15 mins at -80 °C	None	4% SDS in PBS for 2 mins at RT
Short Pepsin	100% Methanol for 10 mins at -20 °C and 15 mins at -80 °C	None	1 mg/ml pepsin in 0.1N HCl for 10 mins at RT
Long Pepsin	100% Methanol for 10 mins at -20 °C and 15 mins at -80 °C	None	1 mg/ml pepsin in 0.1N HCl for 30 mins at RT
Low Prot K	100% Methanol for 10 mins at -20 °C and 15 mins at -80 °C	None	1 $\mu$ g/ml proteinase K in 1XPBS for 10 mins at RT
High Prot K	100% Methanol for 10 mins at -20 °C and 15 mins at -80 °C	None	20 $\mu$ g/ml proteinase K in 1XPBS for 10 mins at 37 °C
Triton + SDS	100% Methanol for 10 mins at -20 °C and 15 mins at -80 °C	0.5% Triton for 15 mins	4% SDS in 1XPBS for 2 mins at RT
Ethanol + Triton	70% Ethanol for 10 mins at RT and 1 hr at 4 °C	0.5% Triton for 15 mins	None
Ethanol + Triton + SDS	70% Ethanol for 10 mins at RT and 1 hr at 4 °C	0.5% Triton for 15 mins	4% SDS in 1XPBS for 2 mins at RT

To test the resulting 24-gene set library on tumour tissues using STARmap, we collected fresh frozen 4T1 tumours sections onto treated glass-slides or into wells of a 24-well chamber. Breast tumours are though tumours with a high percentage of fat and connective tissue which represents a challenge for tissue clearing techniques. Therefore, we optimized the original STARmap protocol and included a pre-treatment

step. We tested various pre-treatment conditions (**Table 4.2**) and found that using 4% SDS in 1XPBS was most effective in preserving the detection signal.

Following the initial clearing process using 4% SDS in 1XPBS, we applied the STARmap protocol and decoded the gene-specific identifiers over 5 cycles. Nuclei were stained with DAPI for cell segmentation and image alignment. We specifically designed our probe set such that all markers for a specific cell type could be identified unambiguously in at least one of the sequencing cycles (**Figure 4.7b**). For example, channel 546 showed all T-cell markers in cycle 1 while all macrophage markers were visualised in cycle 2. We are currently still optimizing the analysis in order to decode the full geneset.

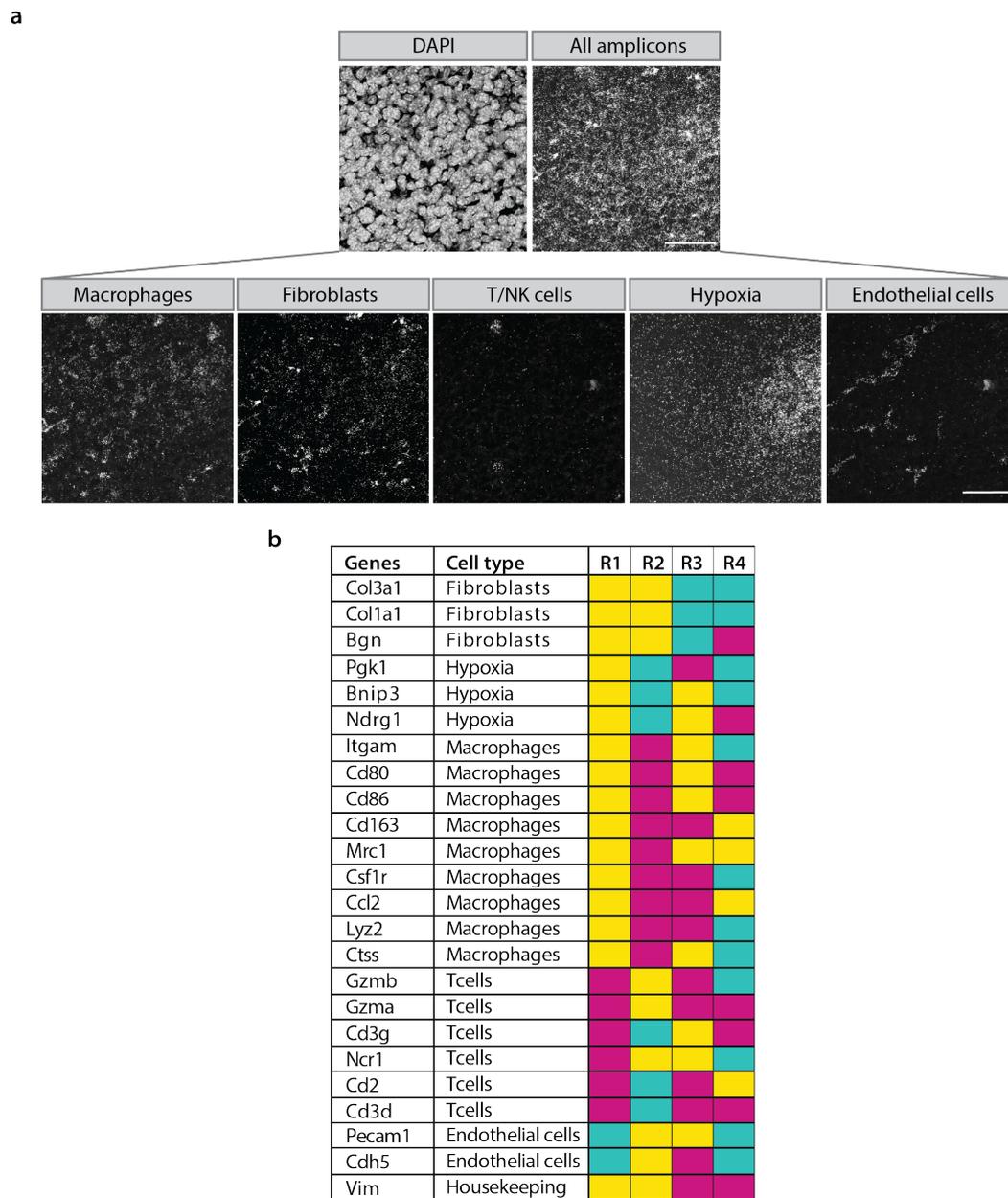
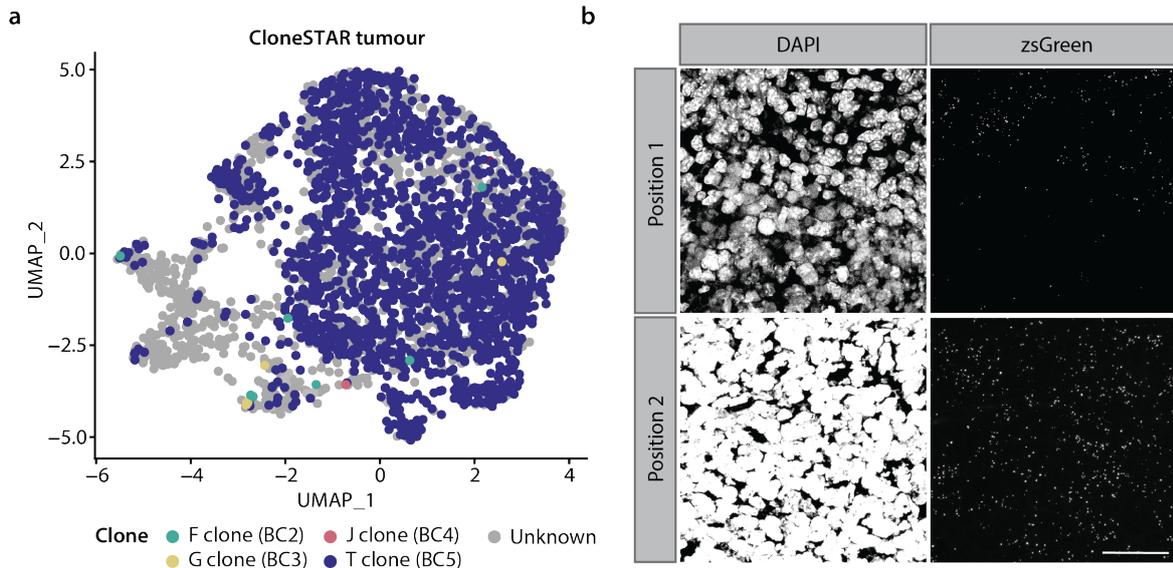


Figure 4.7 **Identification of cell types *in vivo* using STARmap** a| Visualisation of the tumour microenvironment in 4T1 tumours. Nuclei were stained with DAPI and cell type identification was performed using a 24-gene set library and STARmap. Scale bar represents 50  $\mu\text{m}$ . b| *In situ* encoding scheme. Magenta=Alexa547, Cyan=Alexa594, Yellow=647.

#### 4.3.4 Visualisation of clones *in vivo*

In order to test whether we can identify clones *in vivo* using the CloneSTAR barcode, we generated a tumour by pooling the four 4T1 CloneSTAR clones at equal ratios and

transplanting 60,000 cells bilaterally into the mammary fat-pad of a Balb/C mouse. To confirm for the expression of the transcriptional signatures of the clones *in vivo*, we performed scRNA-seq on one of the tumours and processed the other tumour for downstream analysis with the CloneSTAR protocol.



**Figure 4.8 CloneSTAR barcode detection *in vivo* using scRNA-seq and STARmap.** **a** | UMAP representation of CloneSTAR tumour cluster highlighting the different clones. 4T1 clones F (BC2), G (BC3), J (BC4) and T (BC5) were pooled at equal ratios and 60,000 cells were transplanted on both sides into the mammary fat-pad of a Balb/C mouse. Tumours were collected after 21 days and split between scRNA-seq and CloneSTAR analysis. ScRNA-seq revealed that tumour was mostly dominated by the T clone, while only a few other clones were detected. **b** Spatial distribution of CloneSTAR clones injected into the mammary fat-pad of a Balb/C animal. Tumours were collected after 21 days and fresh frozen.

Analysis of the scRNA-seq data revealed that the majority of the tumour consisted of the 4T1-T clone (**Figure 4.8a**). Overall, the CloneSTAR barcode transcript was detected in 57% of all tumour cells with the T clone alone accounting for 56% of barcode assignment. We observed seven barcodes of the F clone, six for the G clone and only one for the J clone. Based on the low expression of most barcodes in our CloneSTAR tumour, we decided to target a common region of the barcode transcript first to increase our chances of detecting the RNA. We used two probes targeting the 3'UTR of zsGreen in the CloneSTAR vector (**Figure 4.8b**). Although we observed specific signal in the tumour sample, the expression was very low indicated by the low number of spots in relation to the amount of cells detected with the nuclear stain and was further highly dependent on the area visualised. The low number of detected

transcripts could be due to the low expression of the CloneSTAR barcode or sterical hindrance which blocks SNAIL probes from binding.

## 4.4 Discussion

Single-cell RNA sequencing (scRNA-seq) characterises the transcriptome of individual cells and has the power to reveal rare subpopulation within a given tissue. However, this comes at the cost of spatial information and thus, limiting our understanding of intercellular communication and organisation in the tumor microenvironment. Using a spatial transcriptomic approach to interrogate a tissue meets this challenge by measuring genes expression of pre-selected markers with detailed positional information in an intact tissue.

To bridge this gap, we developed CloneSTAR, an integrative approach of scRNA-seq and STARmap to study transcriptionally characterised clones within their native tissue context. The method begins with marker characterisation of cells types and clonal populations present in a tissue by scRNA-seq, followed by identification of niches enriched for distinct transcriptomic signatures using spatial transcriptomics and the pre-defined markers from the sequencing analysis. We adapted our WILDseq approach (presented in Chapter 2&3) and generated a novel expressed barcoding system allowing the detection of clonal barcodes with STARmap. Our results demonstrated the power of CloneSTAR in detecting different clones *in vitro* using four transcriptionally distinct clones of our 4T1 mouse model. First, we showcased that we can robustly identify different clones in a mixed population of cells in culture. Secondly, transcriptomic characterisation of the clones confirmed distinct expression signatures and allowed the selection of clonal marker genes. Our CloneSTAR results showed that clonal identities aligned with gene expression signatures *in vitro*.

The widest application of spatial transcriptomic approaches has been to study the tumour microenvironment. To this end, we successfully established STARmap *in vivo* in targeting the tumour microenvironment in 4T1 breast tumours. Based on the low number of barcode assignments in the scRNA-seq *in vitro* data, which has a higher sensitivity than CloneSTAR, we anticipated that detection of barcode *in vivo* using CloneSTAR will be challenging. The signal was weak in the dense tumour tissue and we only observed a low number of detected transcripts. Future studies will focus on improving barcode detection by using a stronger promoter to increase barcode expression or using padlock probes instead of SNAIL probes requiring only one probe to bind instead of two. Alternatively, we will explore the possibility to detect genetic clonality using DNAFISH or transcription sites to detect clones *in vivo*. Furthermore, distinct gene expression patterns of clones might also allow clonal identification. Our

---

preliminary results revealed specific alignment of distinct gene expression patterns and clonal identities *in vitro*. It will be interesting to test the transcriptomics signatures *in vivo* as an alternative approach to identify clones. To this end, the multiplexing capacity of STARmap represents a key feature allowing us to encode enough genes to identify clones *in vivo*. CloneSTAR has the potential to provide a new perspective into the organisation of distinct clonal populations and their interactions in local tissue niches allowing to link the phenotype with the causality.



# Chapter 5

## Conclusion

Molecular differences between individual cells can translate into marked differences in cell fate, especially in determining survival versus death of tumour cells during therapeutic intervention. Identifying the origin of these differences remains a challenge due to the hidden nature of the molecular causes that lead to a distinct cellular fate in bulk data. Single-cell RNA sequencing methods have emerged as a powerful tool to resolve this underlying tumour heterogeneity and provided insights into phenotypes of stromal and tumour cells in different cancers. Triple-negative breast cancer (TNBC) is a highly heterogeneous and aggressive disease that frequently develops therapy resistance. An unresolved question is what precisely distinguishes distinct cellular outcomes to treatment and can we predict cellular fate based on differences in the initial states of cells. Resistance might be caused by the selection of rare pre-existing clones or alternatively through the adaptation of clonal intrinsic and extrinsic processes. The aim of this thesis was to uncover functionally heterogeneous subpopulations of tumour cells with different treatment sensitivity and understand how clonal lineage-dependent transcriptomic diversity contributes to these phenotypes.

Here we developed WILDseq (Wholistic Interrogation of Lineage Dynamics by sequencing), a high-complexity expressed barcode library for simultaneous mapping of each cells' clonal identity and transcriptional states. The barcode is constitutively expressed within the 3' untranslated region of a polyadenylated zsGreen transcript, enabling sequencing via standard oligo dT-capturing chemistry. To mine tumour heterogeneity and couple it to scRNA-seq, we created and optimized a bottlenecking strategy enabling the identification of the appropriate pool size of labelled cells. This step is crucial for the success of the method - too large a starting population of cells

and the number of founding clones/barcodes will be too large to generate sufficient data on an individual clone, too small a starting population and it will lack the diversity required to capture phenotypic variation. We generated our WILDseq pool from 750 barcoded cells which resulted in the detection of 132 clones *in vitro* after stabilisation in culture.

We applied WILDseq to a syngeneic mouse model of breast cancer yielding 64,000 single-cell profiles and 69 distinct clones, indicating that over 50 % of all injected clones successfully survived and expanded. When we compared transcriptional signatures for the most abundant clones *in vivo* versus *in vitro* we observed only a weak correlation indicating that clonal gene expression mostly depends on external factors, such as tumour microenvironment or growth factor supply.

Non-genetic mechanisms, such as transcriptional and epigenetic changes, have recently emerged as important drivers of drug resistance in cancer (Salgia and Kulkarni, 2018). Our results revealed two different mechanisms of treatment resistance. The gene signatures associated with chemoresistance included hypoxia and EMT, the latter relating to the transition of tumour cells to mesenchymal phenotypes in response to cytotoxic therapy. This observation has previously been reported in several studies performed in post treatment samples from breast cancer patients (Almendro et al., 2014; Kim et al., 2018). Similarly, hypoxia has been shown to enhance chemoresistance in tumour cells via HIF-1 (Doktorova et al., 2015; Petit et al., 2016). Our results are consistent with a previous study in TNBC patients that reported that a small subpopulation of tumour cells were primed through their transcriptional programs for a resistant phenotype (Kim et al., 2018). Future work will focus on functional studies to validate our chemoresistant gene signatures and understand their mechanistic roles in conferring resistant phenotypes. Using the CRISPR-Cas9 toolkit, we will induce loss of function mutations in target genes reversing therapy resistance. Alternatively, we will overexpress candidate genes and thereby, reinstating a chemoresistant phenotype in chemosensitive cells. Moreover, our data offers the possibility to overcome chemoresistance by using a combination therapy with inhibiting hypoxia through HIF-1 inhibitors (Hu et al., 2013) or targeting EMT signaling to re-sensitise tumours cells to chemotherapy (Marcucci et al., 2016).

BET bromodomain inhibition has demonstrated efficacy in triple-negative breast cancer (TNBC) and other cancer types, but inherent and acquired resistance limit

---

their potential use in the clinic. We leveraged our WILDseq platform to explore the clonal drug response to the BET bromodomain inhibitor JQ1. Our results revealed a cell state that intrinsically primed tumour cells to be sensitive to JQ1 treatment, while JQ1 resistance was mediated through external factors. Sensitive cells exhibited high levels of DNA damage repair (DDR)-related gene sets at baseline. Previous studies have shown that DNA damage induced a p53-mediated reduction of *MYC* expression (Porter et al., 2017). JQ1 has been reported to downregulate *MYC* transcription, followed by global suppression of *MYC*-dependent target genes (Delmore et al., 2011). Thus, the combined effect of *MYC* suppression resulted in a dramatic decrease of *MYC* in JQ1-sensitive cells. While JQ1-sensitivity was primed through an intrinsic cell state, JQ1-resistance was mediated by micro-environmental factors. We found a global downregulation of gene sets involved in antigen presentation accompanied with a specific depletion of CD8+ T-cells resulting in an immune "cold" tumour. Clonal populations with higher levels of antigen representation via MHC class I showed high abundance *in vitro*, but were only able to survive and expand in the presence of JQ1 treatment. Our data contrast with previous studies that reported the suppression of PD-L1 expression by BET bromodomain inhibition resulting in an increase in the activity of anti-tumor cytotoxic T cells (Hogg et al., 2017; Zhu et al., 2016). One of the studies further demonstrated a synergistic response of JQ1 and immunotherapy with anti-PD1 in a *Myc*-driven lymphoma model (Hogg et al., 2017). Future experiments are required to confirm our hypothesis, starting with analysing the clonal distribution of our WILDseq pool in an immunocompromised animal. Alternatively, we will test whether overexpression of *Myc* can rescue JQ1-sensitive cells.

Further use of integrative tools like WILDseq, in combination with *in situ* sequencing approaches, may help to fully reveal the tumour and its surrounding landscape and directly map clonal transcriptomic differences to their causality. Thus, we aimed to develop a version of WILDseq compatible with the spatial transcriptomics method STARmap. Our resulting CloneSTAR approach has the ability to visualise clonal subpopulations *in vitro* combined with measuring gene expression in a spatial context. We leveraged the expressible nature of the barcode to first perform a detailed transcriptomic characterisation of our clones in order to map and align transcriptomic profiles and clonal identities. Furthermore, we successfully implemented STARmap on breast cancer tissue. However, identification of clones *in vivo* through detection of the CloneSTAR barcode will require further optimisation. In future work, exchang-

ing the promoter to enhance expression or generating a longer barcode to have more sequencing space may improve this technique in order to be utilised in an *in vivo* setting.

Overall, our data have several important clinical implications. First, the pre-existence of chemoresistant phenotypes in the tumour mass indicates that there might be diagnostic possibilities for detecting biomarkers for resistance and sensitivity in patients prior to treatment allowing to anticipate treatment benefits. Second, we found a novel resistance mechanism to BET inhibition demonstrating the importance of having a functional host immune system present. This is a key advantage of WILDseq over CRISPR-Cas9 lineage tracing approaches that rely on the delivery of multiple components resulting in an increased risk of immunogenicity and vector silencing. Due to its compatibility with syngeneic models, WILDseq would be well suited to study clonal response to immunotherapy and increase our understanding of therapy response and resistance to immunotherapy. Third, integrating transcriptional profiles and clonal identities with the spatial information has the potential to directly map clonal transcriptomic differences to their causality completing the picture of the tumour and its surrounding landscape.

Overall, we have developed a novel integrative platform, called WILDseq, that can enable high-resolution mapping of clonal identities and cell states. This technology will increase our knowledge about how complex and heterogeneous tumour populations change and adapt to perturbations such as drug treatment, providing a novel and valuable tool for the study of tumour heterogeneity and evolution. Ultimately, understanding (epi)genomic, transcriptomic, and phenotypic changes that occur during tumour evolution upon therapeutic intervention may inform treatment strategies and lead to better clinical outcomes.

# References

- Abbosh, C., Birkbak, N. J., Wilson, G. A., Jamal-Hanjani, M., Constantin, T., Salari, R., Le Quesne, J., Moore, D. A., Veeriah, S., Rosenthal, R., Marafioti, T., Blackhall, F., Summers, Y., Hafez, D., Naik, A., Ganguly, A., Kareht, S., Shah, R., Joseph, L., Marie Quinn, A., Crosbie, P. A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Oukrif, D., Akarca, A. U., Hartley, J. A., Lowe, H. L., Lock, S., Iles, N., Bell, H., Ngai, Y., Elgar, G., Szallasi, Z., Schwarzh, R. F., Herrero, J., Stewart, A., Quezada, S. A., Peggs, K. S., Van Loo, P., Dive, C., Lin, C. J., Rabinowitz, M., Aerts, H. J. W. L., Hackshaw, A., Shaw, J. A., Zimmermann, B. G. and Swanton, C. (2017). Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* *545*, 446–451.
- Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M. S., Tomasini, L., Ferrandino, A. F., Rosenberg Belmaker, L. A., Szekely, A., Wilson, M., Kocabas, A., Calixto, N. E., Grigorenko, E. L., Huttner, A., Chawarska, K., Weissman, S., Urban, A. E., Gerstein, M. and Vaccarino, F. M. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* *492*, 438–442.
- Aleman, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. and van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature* *556*, 108–112.
- Almendro, V., Cheng, Y.-K., Randles, A., Itzkovitz, S., Marusyk, A., Ametller, E., Gonzalez-Farre, X., Muñoz, M., Russnes, H., Helland, A., Rye, I., Borresen-Dale, A.-L., Maruyama, R., van Oudenaarden, A., Dowsett, M., Jones, R., Reis-Filho, J., Gascon, P., Gönen, M., Michor, F. and Polyak, K. (2014). Inference of Tumor Evolution during Chemotherapy by Computational Modeling and In Situ Analysis of Genetic and Phenotypic Cellular Diversity. *Cell Reports* *6*, 514–527.
- Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P. and Maley, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine* *22*, 105–113.
- Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., Levenson, R. M., Lowe, J. B., Liu, S. D., Zhao, S., Natkunam, Y. and Nolan, G. P. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature Medicine* *20*, 436–442.

- Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P. J., Kelsey, G., Dean, W., Nichols, J., Stegle, O., Marioni, J. C. and Reik, W. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* *576*, 487–491.
- Aslakson, C. J. and Miller, F. R. (1992). Selective Events in the Metastatic Process Defined by Analysis of the Sequential Dissemination of Subpopulations of a Mouse Mammary Tumor. *Cancer Research* *52*, 1399 LP – 1405.
- Bach, K., Pensa, S., Grzelak, M., Hadfield, J., Adams, D. J., Marioni, J. C. and Khaled, W. T. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications* *8*, 2128.
- Barker, N., van Es, J. H., Kuipers, J., Kujala, P., van den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P. J. and Clevers, H. (2007). Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* *449*, 1003–1007.
- Baron, C. S. and van Oudenaarden, A. (2019). Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nature Reviews Molecular Cell Biology* *20*, 753–765.
- Bhang, H.-e. C., Ruddy, D. A., Krishnamurthy Radhakrishna, V., Caushi, J. X., Zhao, R., Hims, M. M., Singh, A. P., Kao, I., Rakiec, D., Shaw, P., Balak, M., Raza, A., Ackley, E., Keen, N., Schlabach, M. R., Palmer, M., Leary, R. J., Chiang, D. Y., Sellers, W. R., Michor, F., Cooke, V. G., Korn, J. M. and Stegmeier, F. (2015). Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature Medicine* *21*, 440.
- Biddy, B. A., Kong, W., Kamimoto, K., Guo, C., Waye, S. E., Sun, T. and Morris, S. A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. *Nature* *564*, 219–224.
- Burrell, R. A. and Swanton, C. (2014). Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular Oncology* *8*, 1095–1111.
- Buschmann, T. (2017). DNABarcodes: an R package for the systematic construction of DNA sample tags. *Bioinformatics* *33*, 920–922.
- Cai, X., Evrony, G., Lehmann, H., Elhosary, P., Mehta, B., Poduri, A. and Walsh, C. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Reports* *8*, 1280–1289.
- Calabrese, C., Davidson, N. R., Demircioğlu, D., Fonseca, N. A., He, Y., Kahles, A., Lehmann, K.-V., Liu, F., Shiraishi, Y., Soulette, C. M., Urban, L., Greger, L., Li, S., Liu, D., Perry, M. D., Xiang, Q., Zhang, F., Zhang, J., Bailey, P., Erkek, S., Hoadley, K. A., Hou, Y., Huska, M. R., Kilpinen, H., Korbel, J. O., Marin, M. G., Markowski, J., Nandi, T., Pan-Hammarström, Q., Peadamallu, C. S., Siebert, R., Stark, S. G., Su, H., Tan, P., Waszak, S. M., Yung, C., Zhu, S., Awadalla, P., Creighton, C. J.,

- Meyerson, M., Ouellette, B. F. F., Wu, K., Yang, H., Brazma, A., Brooks, A. N., Göke, J., Rättsch, G., Schwarz, R. F., Stegle, O. and Zhang, Z. (2020). Genomic basis for RNA alterations in cancer. *Nature* *578*, 129–136.
- Campbell, K. R., Steif, A., Laks, E., Zahn, H., Lai, D., McPherson, A., Farahani, H., Kabeer, F., O’Flanagan, C., Biele, J., Brimhall, J., Wang, B., Walters, P., Consortium, I., Bouchard-Côté, A., Aparicio, S. and Shah, S. P. (2019). clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biology* *20*, 54.
- Cannell, I. G., Kong, Y. W., Johnston, S. J., Chen, M. L., Collins, H. M., Dobbyn, H. C., Elia, A., Kress, T. R., Dickens, M., Clemens, M. J., Heery, D. M., Gaestel, M., Eilers, M., Willis, A. E. and Bushell, M. (2010). p38 MAPK/MK2-mediated induction of miR-34c following DNA damage prevents Myc-dependent DNA replication. *Proceedings of the National Academy of Sciences* *107*, 5375 LP – 5380.
- Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca, G., Troester, M. A., Tse, C. K., Edmiston, S., Deming, S. L., Geradts, J., Cheang, M. C., Nielsen, T. O., Moorman, P. G., Earp, H. S. and Millikan, R. C. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* *295*.
- Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M. E. and Navin, N. E. (2018). Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* *172*, 205–217.
- Caswell-Jin, J. L., McNamara, K., Reiter, J. G., Sun, R., Hu, Z., Ma, Z., Ding, J., Suarez, C. J., Tilk, S., Raghavendra, A., Forte, V., Chin, S.-F., Bardwell, H., Provenzano, E., Caldas, C., Lang, J., West, R., Tripathy, D., Press, M. F. and Curtis, C. (2019). Clonal replacement and heterogeneity in breast tumors treated with neoadjuvant HER2-targeted therapy. *Nature Communications* *10*, 657.
- Chan, M. M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T. M., Adamson, B., Jost, M., Quinn, J. J., Yang, D., Jones, M. G., Khodaverdian, A., Yosef, N., Meissner, A. and Weissman, J. S. (2019). Molecular recording of mammalian embryogenesis. *Nature* *570*, 77–82.
- Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* *348*.
- Crosetto, N., Bienko, M. and van Oudenaarden, A. (2015). Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics* *16*, 57–66.
- Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Graf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerod, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowitz, F., Murphy, L., Ellis, I., Purushotham, A., Borresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C. and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* *486*, 346–352.

- Das Thakur, M., Salangsang, F., Landman, A. S., Sellers, W. R., Pryer, N. K., Levesque, M. P., Dummer, R., McMahon, M. and Stuart, D. D. (2013). Modelling vemurafenib resistance in melanoma reveals a strategy to forestall drug resistance. *Nature* *494*, 251–255.
- Delmore, J., Issa, G., Lemieux, M., Rahl, P., Shi, J., Jacobs, H., Kastritis, E., Gilpatrick, T., Paranal, R., Qi, J., Chesi, M., Schinzel, A., McKeown, M., Heffernan, T., Vakoc, C., Bergsagel, P., Ghobrial, I., Richardson, P., Young, R., Hahn, W., Anderson, K., Kung, A., Bradner, J. and Mitsiades, C. (2011). BET Bromodomain Inhibition as a Therapeutic Strategy to Target c-Myc. *Cell* *146*, 904–917.
- Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., Abbott, R. M., Hoog, J., Dooling, D. J., Koboldt, D. C., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wendl, M. C., McMichael, J. F., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Appelbaum, E., DeSchryver, K., Davies, S., Guintoli, T., Lin, L., Crowder, R., Tao, Y., Snider, J. E., Smith, S. M., Dukes, A. F., Sanderson, G. E., Pohl, C. S., Delehaunty, K. D., Fronick, C. C., Pape, K. A., Reed, J. S., Robinson, J. S., Hodges, J. S., Schierding, W., Dees, N. D., Shen, D., Locke, D. P., Wiechert, M. E., Eldred, J. M., Peck, J. B., Oberkfell, B. J., Lolofo, J. T., Du, F., Hawkins, A. E., O’Laughlin, M. D., Bernard, K. E., Cunningham, M., Elliott, G., Mason, M. D., Thompson Jr, D. M., Ivanovich, J. L., Goodfellow, P. J., Perou, C. M., Weinstock, G. M., Aft, R., Watson, M., Ley, T. J., Wilson, R. K. and Mardis, E. R. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* *464*, 999–1005.
- Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., McMichael, J. F., Wallis, J. W., Lu, C., Shen, D., Harris, C. C., Dooling, D. J., Fulton, R. S., Fulton, L. L., Chen, K., Schmidt, H., Kalicki-Weizer, J., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Wendl, M. C., Heath, S., Watson, M. A., Link, D. C., Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., Graubert, T. A., Mardis, E. R., Wilson, R. K. and DiPersio, J. F. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* *481*, 506–510.
- Doktorova, H., Hrabeta, J., Khalil, M. A. and Eckschlager, T. (2015). Hypoxia-induced chemoresistance in cancer cells: The role of not only HIF-1. *Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia* *159*, 166–177.
- Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulina, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C. and Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* *568*, 235–239.
- Fekete, J. T. and Györffy, B. (2019). ROCplot.org: Validating predictive biomarkers of chemotherapy/hormonal therapy/anti-HER2 therapy using transcriptomic data of 3,104 breast cancer patients. *International Journal of Cancer* *145*, 3140–3151.
- Flamant, L., Notte, A., Ninane, N., Raes, M. and Michiels, C. (2010). Anti-apoptotic role of HIF-1 and AP-1 in paclitaxel exposed breast cancer cells under hypoxia. *Molecular Cancer* *9*, 191.

- Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. and Shapiro, E. (2005). Genomic Variability within an Organism Exposes Its Cell Lineage Tree. *PLOS Computational Biology* *1*, e50.
- Gao, R., Davis, A., McDonald, T. O., Sei, E., Shi, X., Wang, Y., Tsai, P.-C., Casasent, A., Waters, J., Zhang, H., Meric-Bernstam, F., Michor, F. and Navin, N. E. (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics* *48*, 1119–1130.
- Gerlach, C., Rohr, J. C., Perié, L., van Rooij, N., van Heijst, J. W. J., Velds, A., Urbanus, J., Naik, S. H., Jacobs, H., Beltman, J. B., de Boer, R. J. and Schumacher, T. N. M. (2013). Heterogeneous Differentiation Patterns of Individual CD8+ T Cells. *Science* *340*, 635 LP – 639.
- Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C. R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P. A., Stamp, G., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Futreal, P. A., Stewart, A. and Swanton, C. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics* *46*, 225–233.
- Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., ... and Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* *366*, 883–892.
- Gerrits, A., Dykstra, B., Kalmykova, O. J., Klauke, K., Verovskaya, E., Broekhuis, M. J. C., de Haan, G. and Bystrykh, L. V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* *115*, 2610 LP – 2618.
- Gerstung, M., Jolly, C., Leshchiner, I., Dentre, S. C., Gonzalez, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vázquez-García, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., Malikic, S., Donmez, N., Livitz, D. G., Cmero, M., Demeulemeester, J., Schumacher, S., Fan, Y., Yao, X., Lee, J., Schlesner, M., Boutros, P. C., Bowtell, D. D., Zhu, H., Getz, G., Imielinski, M., Beroukhi, R., Sahinalp, S. C., Ji, Y., Peifer, M., Markowitz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q. D., Spellman, P. T., Wedge, D. C. and Van Loo, P. (2020). The evolutionary history of 2,658 cancers. *Nature* *578*, 122–128.
- Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Love, J. C. and Shalek, A. K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods* *14*, 395–398.
- Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D. and Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods* *11*, 417–422.
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B.,

- Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P. and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell* *184*, 3573–3587.
- Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K. J., Rozenblatt-Rosen, O., Dor, Y., Regev, A. and Yanai, I. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology* *17*, 77.
- Hawkins, J. A., Jones, S. K., Finkelstein, I. J. and Press, W. H. (2018). Indel-correcting DNA barcodes for high-throughput sequencing. *Proceedings of the National Academy of Sciences* *115*, E6217 LP – E6226.
- Heppner, G. H., Dexter, D. L., DeNucci, T., Miller, F. R. and Calabresi, P. (1978). Heterogeneity in Drug Sensitivity among Tumor Cell Subpopulations of a Single Mammary Tumor. *Cancer Research* *38*, 3758 LP – 3763.
- Hogg, S. J., Vervoort, S. J., Deswal, S., Ott, C. J., Li, J., Cluse, L. A., Beavis, P. A., Darcy, P. K., Martin, B. P., Spencer, A., Traunbauer, A. K., Sadovnik, I., Bauer, K., Valent, P., Bradner, J. E., Zuber, J., Shortt, J. and Johnstone, R. W. (2017). BET-Bromodomain Inhibitors Engage the Host Immune System and Regulate Expression of the Immune Checkpoint Ligand PD-L1. *Cell Reports* *18*, 2162–2174.
- Holohan, C., Van Schaeybroeck, S., Longley, D. B. and Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer* *13*, 714–726.
- Hu, Y., Liu, J. and Huang, H. (2013). Recent agents targeting HIF-1 $\alpha$  for cancer therapy. *Journal of Cellular Biochemistry* *114*, 498–509.
- Huang, L., Ao, Q., Zhang, Q., Yang, X., Xing, H., Li, F., Chen, G., Zhou, J., Wang, S., Xu, G., Meng, L., Lu, Y. and Ma, D. (2010). Hypoxia induced paclitaxel resistance in human ovarian cancers via hypoxia-inducible factor 1 $\alpha$ . *Journal of Cancer Research and Clinical Oncology* *136*, 447–456.
- Itzkovitz, S., Lyubimova, A., Blat, I. C., Maynard, M., van Es, J., Lees, J., Jacks, T., Clevers, H. and van Oudenaarden, A. (2012). Single-molecule transcript counting of stem-cell markers in the mouse intestine. *Nature Cell Biology* *14*, 106–114.
- Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., Lee, S.-M., Forster, M. D., Ahmad, T., Hiley, C. T., Abbosh, C., Falzon, M., Borg, E., Marafioti, T., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P. A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Dentro, S., Tanriere, P., O’Sullivan, B., Lowe, H. L., Hartley, J. A., Iles, N., Bell, H., Ngai, Y., Shaw, J. A., Herrero, J., Szallasi, Z., Schwarz, R. F., Stewart, A., Quezada, S. A., Le Quesne, J., Van Loo, P., Dive, C., Hackshaw,

- A. and Swanton, C. (2017). Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* *376*, 2109–2121.
- Jensen, P. and Dymecki, S. M. (2014). *Essentials of Recombinase-Based Genetic Fate Mapping in Mice* BT - *Mouse Molecular Embryology: Methods and Protocols*. Springer US, Boston, MA.
- Ju, Y. S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L. B., Rahbari, R., Wedge, D. C., Davies, H. R., Ramakrishna, M., Fullam, A., Martin, S., Alder, C., Patel, N., Gamble, S., O’Meara, S., Giri, D. D., Sauer, T., Pinder, S. E., Purdie, C. A., Borg, A., Stunnenberg, H., van de Vijver, M., Tan, B. K. T., Caldas, C., Tutt, A., Ueno, N. T., van ’t Veer, L. J., Martens, J. W. M., Sotiriou, C., Knappskog, S., Span, P. N., Lakhani, S. R., Eyfjoerd, J. E., Borresen-Dale, A.-L., Richardson, A., Thompson, A. M., Viari, A., Hurles, M. E., Nik-Zainal, S., Campbell, P. J. and Stratton, M. R. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* *543*, 714–718.
- Jungwirth, U., van Weverwijk, A., Melake, M. J., Chambers, A. F., Gao, Q., Fivaz, M. and Isacke, C. M. (2018). Generation and characterisation of two D2A1 mammary cancer sublines to model spontaneous and experimental metastasis in a syngeneic BALB/c host. *Disease Models & Mechanisms* *11*.
- Juric, D., Castel, P., Griffith, M., Griffith, O. L., Won, H. H., Ellis, H., Ebbesen, S. H., Ainscough, B. J., Ramu, A., Iyer, G., Shah, R. H., Huynh, T., Mino-Kenudson, M., Sgroi, D., Isakoff, S., Thabet, A., Elamine, L., Solit, D. B., Lowe, S. W., Quadt, C., Peters, M., Derti, A., Schegel, R., Huang, A., Mardis, E. R., Berger, M. F., Baselga, J. and Scaltriti, M. (2015). Convergent loss of PTEN leads to clinical resistance to a PI(3)K $\alpha$  inhibitor. *Nature* *518*, 240–244.
- Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P. and Church, G. M. (2018). Developmental barcoding of whole mouse via homing CRISPR. *Science* *361*, eaat9804.
- Kalhor, R., Mali, P. and Church, G. M. (2017). Rapidly evolving homing CRISPR barcodes. *Nature Methods* *14*, 195–200.
- Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C. and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature methods* *10*, 857.
- Keller, G., Paige, C., Gilboa, E. and Wagner, E. F. (1985). Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. *Nature* *318*, 149–154.
- Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T. and Navin, N. E. (2018). Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* *173*, 879–893.
- Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. and Kirschner, M. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* *161*, 1187–1201.

- Kurtova, A. V., Xiao, J., Mo, Q., Pazhanisamy, S., Krasnow, R., Lerner, S. P., Chen, F., Roh, T. T., Lay, E., Ho, P. L. and Chan, K. S. (2015). Blocking PGE2-induced tumour repopulation abrogates bladder cancer chemoresistance. *Nature* *517*, 209–213.
- Landau, D., Carter, S., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S., Vartanov, A., Fernandes, S., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., Hacohen, N., Meyerson, M., Lander, E., Neuberg, D., Brown, J., Getz, G. and Wu, C. (2013). Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* *152*, 714–726.
- Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* *10*, R25.
- Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalthor, R., Yang, J. L., Ferrante, T. C., Terry, R., Jeanty, S. S. F., Li, C., Amamoto, R., Peters, D. T., Turczyk, B. M., Marblestone, A. H., Inverso, S. A., Bernard, A., Mali, P., Rios, X., Aach, J. and Church, G. M. (2014). Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science* *343*, 1360 LP – 1363.
- Lelekakis, M., Moseley, J. M., Martin, T. J., Hards, D., Williams, E., Ho, P., Lowen, D., Javni, J., Miller, F. R., Slavin, J. and Anderson, R. L. (1999). A novel orthotopic model of breast cancer metastasis to bone. *Clinical & Experimental Metastasis* *17*, 163–170.
- Lemischka, I. R., Raulet, D. H. and Mulligan, R. C. (1986). Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* *45*, 917–927.
- Leung, M. L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., Vilar, E., Maru, D., Kopetz, S. and Navin, N. E. (2017). Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome research* *27*, 1287–1299.
- Li, X., Lewis, M. T., Huang, J., Gutierrez, C., Osborne, C. K., Wu, M. . F., Hilsenbeck, S. G., Pavlick, A., Zhang, X., Chamness, G. C., Wong, H., Rosen, J. and Chang, J. C. (2008). Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *J Natl Cancer Inst* *100*.
- Lin, J.-R., Fallahi-Sichani, M. and Sorger, P. K. (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature Communications* *6*, 8390.
- Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D’Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J. and Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* *350*, 94 LP – 98.
- Longo, S. K., Guo, M. G., Ji, A. L. and Khavari, P. A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics* *1*.

- Lu, R., Neff, N. F., Quake, S. R. and Weissman, I. L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature Biotechnology* *29*, 928.
- Lubeck, E. and Cai, L. (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods* *9*, 743–748.
- Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods* *11*, 360–361.
- Lyubimova, A., Itzkovitz, S., Junker, J. P., Fan, Z. P., Wu, X. and van Oudenaarden, A. (2013). Single-molecule mRNA detection and counting in mammalian tissue. *Nature Protocols* *8*, 1743–1758.
- Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A. and McCarroll, S. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* *161*, 1202–1214.
- Marcucci, F., Stassi, G. and De Maria, R. (2016). Epithelial–mesenchymal transition: a new target in anticancer drug discovery. *Nature Reviews Drug Discovery* *15*, 311–325.
- Marine, J.-C., Dawson, S.-J. and Dawson, M. A. (2020). Non-genetic mechanisms of therapeutic resistance in cancer. *Nature Reviews Cancer* *20*, 743–756.
- McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R. S., Vermeesch, J. R., Hall, I. M. and Gage, F. H. (2013). Mosaic Copy Number Variation in Human Neurons. *Science* *342*, 632 LP – 637.
- McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F. and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* *353*, aaf7907.
- McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., Zhang, A. W., ... and Shah, S. P. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics* *48*, 758–767.
- Mooijman, D., Dey, S. S., Boisset, J.-C., Crosetto, N. and van Oudenaarden, A. (2016). Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature Biotechnology* *34*, 852–856.
- Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. and Groop, L. C. (2003). PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* *34*, 267–273.

- Naik, S. H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R. J. and Schumacher, T. N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* *496*, 229–232.
- Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J. and Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* *472*, 90–94.
- Nguyen, L. V., Cox, C. L., Eirew, P., Knapp, D. J. H. F., Pellacani, D., Kannan, N., Carles, A., Moksa, M., Balani, S., Shah, S., Hirst, M., Aparicio, S. and Eaves, C. J. (2014). DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. *Nature Communications* *5*, 5871.
- Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J. and Reed, J. (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clinical cancer research* *16*, 5222–5232.
- Parker, J., Mullins, M., Cheang, M., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J., Stijleman, I., Palazzo, J., Marron, J., Nobel, A., Mardis, E., Nielsen, T., Ellis, M., Perou, C. M. and Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* *27*.
- Pei, W., Feyerabend, T. B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., Chen, W., Sauer, S., Wolf, S., Höfer, T. and Rodewald, H.-R. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* *548*, 456–460.
- Pei, W., Wang, X., Rössler, J., Feyerabend, T. B., Höfer, T. and Rodewald, H.-R. (2019). Using Cre-recombinase-driven Polylox barcoding for in vivo fate mapping in mice. *Nature Protocols* *14*, 1820–1840.
- Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R. and Sammut, S.-J. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications* *7*, 1–16.
- Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. . L., Brown, P. O. and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature* *406*.
- Petit, C., Gouel, F., Dubus, I., Heuclin, C., Roget, K. and Vannier, J. P. (2016). Hypoxia promotes chemoresistance in acute lymphoblastic leukemia cell lines by modulating death signaling pathways. *BMC Cancer* *16*, 746.
- Picelli, S., Björklund, A. s. K., Faridani, O. R., Sagasser, S., Winberg, G. and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* *10*, 1096–1098.

- Polyak, K. and Weinberg, R. A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer* 9.
- Porter, J. R., Fisher, B. E., Baranello, L., Liu, J. C., Kambach, D. M., Nie, Z., Koh, W. S., Luo, J., Stommel, J. M., Levens, D. and Batchelor, E. (2017). Global Inhibition with Specific Activation: How p53 and MYC Redistribute the Transcriptome in the DNA Double-Strand Break Response. *Molecular Cell* 67, 1013–1025.
- Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X. and Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research* 12, R68.
- Pucci, P., Rescigno, P., Sumanasuriya, S., de Bono, J. and Crea, F. (2018). Hypoxia and Noncoding RNAs in Taxane Resistance. *Trends in Pharmacological Sciences* 39, 695–709.
- Pulaski, B. A. and Ostrand-Rosenberg, S. (1998). Reduction of Established Spontaneous Mammary Carcinoma Metastases following Immunotherapy with Major Histocompatibility Complex Class II and B7.1 Cell-based Tumor Vaccines. *Cancer Research* 58, 1486 LP – 1493.
- Quinn, J. J., Jones, M. G., Okimoto, R. A., Nanjo, S., Chan, M. M., Yosef, N., Bivona, T. G. and Weissman, J. S. (2021). Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* 371, eabc1944.
- Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* 5, 877–879.
- Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A. and Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology* 36, 442–450.
- Rodrigues, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F. and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463 LP – 1467.
- Salehi, S., Kabeer, F., Ceglia, N., Andronescu, M., Williams, M. J., Campbell, K. R., Masud, T., Wang, B., Biele, J., Brimhall, J., Gee, D., Lee, H., Ting, J., Zhang, A. W., Tran, H., O’Flanagan, C., Dorri, F., Rusk, N., de Algara, T. R., Lee, S. R., Cheng, B. Y. C., Eirew, P., Kono, T., Pham, J., Grewal, D., Lai, D., Moore, R., Mungall, A. J., Marra, M. A., Hannon, G. J., Battistoni, G., Bressan, D., Cannell, I. G., Casbolt, H., Fatemi, A., Jauset, C., Kovačević, T., Mulvey, C. M., Nugent, F., Ribes, M. P., Pearsall, I., Qosaj, F., Sawicka, K., Wild, S. A., Williams, E., Laks, E., Li, Y., O’Flanagan, C. H., Smith, A., Ruiz, T., Lai, D., Roth, A., Balasubramanian, S., Lee, M., Bodenmiller, B., Burger, M., Kuett, L., Tietscher, S., Windhager, J., Boyden, E. S., Alon, S., Cui, Y., Emenari, A., Goodwin, D., Karagiannis, E. D., Sinha, A., Wassie, A. T., Caldas, C., Bruna, A., Callari, M., Greenwood, W., Lerda, G., Eyal-Lubling, Y., Rueda, O. M., Shea, A., Harris, O., Becker, R., Grimaldi, F., Harris, S., Vogl, S. L., Weselak, J., Joyce, J. A., Watson, S. S., Vázquez-García,

- I., Tavaré, S., Dinh, K. N., Fisher, E., Kunes, R., Walton, N. A., Sa'd, M. A., Chornay, N., Dariush, A., González-Solares, E. A., González-Fernández, C., Yoldas, A. K., Millar, N., Whitmarsh, T., Zhuang, X., Fan, J., Lee, H., Sepúlveda, L. A., Xia, C., Zheng, P., McPherson, A., Bouchard-Côté, A., Aparicio, S. and Shah, S. P. (2021). Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature* *595*, 585–590.
- Salgia, R. and Kulkarni, P. (2018). The Genetic/Non-genetic Duality of Drug ‘Resistance’ in Cancer. *Trends in Cancer* *4*, 110–118.
- Samanta, D., Gilkes, D. M., Chaturvedi, P., Xiang, L. and Semenza, G. L. (2014). Hypoxia-inducible factors are required for chemotherapy resistance of breast cancer stem cells. *Proceedings of the National Academy of Sciences* *111*, E5429 LP – E5438.
- Schroers, B., Boegel, S., Albrecht, C., Bukur, T., Bukur, V., Holtsträter, C., Ritzel, C., Manninen, K., Tadmor, A. D., Vormehr, M., Sahin, U. and Löwer, M. (2020). Multi-Omics Characterization of the 4T1 Murine Mammary Gland Tumor Model.
- Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* *18*, 213–229.
- Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., ... and Aparicio, S. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* *486*.
- Shu, S., Lin, C. Y., He, H. H., Witwicki, R. M., Tabassum, D. P., Roberts, J. M., Janiszewska, M., Jin Huh, S., Liang, Y., Ryan, J., Doherty, E., Mohammed, H., Guo, H., Stover, D. G., Ekram, M. B., Peluffo, G., Brown, J., D’Santos, C., Krop, I. E., Dillon, D., McKeown, M., Ott, C., Qi, J., Ni, M., Rao, P. K., Duarte, M., Wu, S.-Y., Chiang, C.-M., Anders, L., Young, R. A., Winer, E. P., Letai, A., Barry, W. T., Carroll, J. S., Long, H. W., Brown, M., Shirley Liu, X., Meyer, C. A., Bradner, J. E. and Polyak, K. (2016). Response and resistance to BET bromodomain inhibitors in triple-negative breast cancer. *Nature* *529*, 413–417.
- Smith, T. S., Heger, A. and Sudbery, I. (2017). UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research* *27*.
- Sørli, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lønning, P. and Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* *98*.
- Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D. and Curtis, C. (2015). A Big Bang model of human colorectal tumor growth. *Nature Genetics* *47*, 209–216.
- Sottoriva, A., Spiteri, I., Piccirillo, S. G. M., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C. and Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences* *110*, 4009 LP – 4014.

- Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N. and Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nature Biotechnology* *36*, 469–473.
- Stahl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, A., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J. and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* *353*, 78 LP – 82.
- Stripecke, R., del Carmen Villacres, M., Skelton, D. C., Satake, N., Halene, S. and Kohn, D. B. (1999). Immune response to green fluorescent protein: implications for gene therapy. *Gene Therapy* *6*, 1305–1312.
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* *177*, 1888–1902.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* *102*, 15545 LP – 15550.
- Sun, J., Ramos, A., Chapman, B., Johnnidis, J. B., Le, L., Ho, Y.-J., Klein, A., Hofmann, O. and Camargo, F. D. (2014). Clonal dynamics of native haematopoiesis. *Nature* *514*, 322–327.
- Takei, Y., Yun, J., Zheng, S., Ollikainen, N., Pierson, N., White, J., Shah, S., Thomassie, J., Suo, S., Eng, C.-H. L., Guttman, M., Yuan, G.-C. and Cai, L. (2021). Integrated spatial genomics reveals global architecture of single nuclei. *Nature* *590*, 344–350.
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* *6*, 377–382.
- The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* *490*, 61–70.
- Townsend, D. M. and Tew, K. D. (2003). The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* *22*, 7369–7375.
- Wagenblast, E., Soto, M., Gutierrez-Angel, S., Hartl, C. A., Gable, A. L., Maceli, A. R., Erard, N., Williams, A. M., Kim, S. Y., Dickopf, S., Harrell, J. C., Smith, A. D., Perou, C. M., Wilkinson, J. E., Hannon, G. J. and Knott, S. R. V. (2015). A model of breast cancer heterogeneity reveals vascular mimicry as a driver of metastasis. *Nature* *520*, 358–362.
- Wagner, D. E. and Klein, A. M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics* *21*, 410–427.

- Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G. and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* *360*, 981 LP – 987.
- Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G. P., Bava, F.-A. and Deisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* *361*.
- Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F. and Navin, N. E. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* *512*, 155.
- Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. and Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* *367*, eaaw3381.
- Wigler, M., Pellicer, A., Silverstein, S. and Axel, R. (1978). Biochemical transfer of single-copy eucaryotic genes using total cellular DNA as donor. *Cell* *14*, 725–731.
- Woodworth, M. B., Girskis, K. M. and Walsh, C. A. (2017). Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nature Reviews Genetics* *18*, 230–244.
- Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., ... and Wang, J. (2012). Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. *Cell* *148*, 886–895.
- Yang, J., Mani, S. A., Donaher, J. L., Ramaswamy, S., Itzykson, R. A., Come, C., Savagner, P., Gitelman, I., Richardson, A. and Weinberg, R. A. (2004). Twist, a Master Regulator of Morphogenesis, Plays an Essential Role in Tumor Metastasis. *Cell* *117*, 927–939.
- Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., ... and Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine* *21*, 751–759.
- Zafar, H., Tzen, A., Navin, N., Chen, K. and Nakhleh, L. (2017). SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology* *18*, 178.
- Zarrei, M., MacDonald, J. R., Merico, D. and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics* *16*, 172–183.
- Zhang, J., Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., ... and Futreal, P. A. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* *346*, 256 LP – 259.
- Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y. and Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell* *73*, 130–142.

- Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications* *8*, 14049.
- Zhu, H., Bengsch, F., Svoronos, N., Rutkowski, M., Bitler, B., Allegranza, M., Yokoyama, Y., Kossenkov, A., Bradner, J., Conejo-Garcia, J. and Zhang, R. (2016). BET Bromodomain Inhibition Promotes Anti-tumor Immunity by Suppressing PD-L1 Expression. *Cell Reports* *16*, 2829–2837.

