Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M. Sc. Maja Dorota Starostecka

born in Jelenia Góra, Poland

Oral examination: 2.05.2023

Characterization of cell-type-specific responses to doxorubicin treatment in murine breast cancer organoids using a single-cell multi-omics approach

Referees:

Prof. Dr. Michael Boutros

Prof. Dr. Oliver Stegle

Abstract

Doxorubicin is a DNA-damaging agent, widely used as a chemotherapeutic in clinics to treat HER2-positive breast cancer patients. Despite the overall cytotoxic effect, some cancerous cells may survive the treatment by successfully repairing DNA breaks or converting the original damage into mutations. In the long term, their sustained presence may lead to therapy resistance, cancer relapse, and secondary malignancies. The interaction between DNA damage induced by drugs and repair pathways leaves characteristic patterns, called mutational signatures. While base substitutions or small indel signatures have been annotated for some chemotherapies, the pattern and impact of structural variants (SVs) have remained underexplored. In this thesis, I investigated the SV burden and transcriptomic changes promoted by doxorubicin, at the single-cell level, in a murine mammary gland organoid model of HER2-positive breast cancer.

Single-cell transcriptomic profiling revealed the three main cell types present in the murine mammary gland organoids, as well as both global and cell-type-specific, changes induced by doxorubicin.

In addition, single-cell DNA template strand sequencing (Strand-seq) was further developed and applied here for the first time to characterize doxorubicin-associated genomic instability in murine organoids. Thanks to a novel computational single-cell multi-omic method, scNOVA, genomic data were integrated with nucleosome occupancy (NO) measurements to enable simultaneous SV detection and cell-type classification in the same cell. Strand-seq and scNOVA integration showed that doxorubicin is associated with a higher SV burden and increased frequency of sister chromatid exchanges in all three cell types of murine mammary gland organoids. Deletions and complex rearrangements emerged as candidate mutational patterns of doxorubicin.

Taken together, the results presented in this thesis exemplify the synergistic integration of organoid models with single-cell multi-omic readouts for the systematic study of heterogenous populations and demonstrate the necessity to expand the search for therapy-associated mutational signatures to SVs.

Zusammenfassung

Doxorubicin ist ein DNA-schädigender Wirkstoff, der in Kliniken häufig als Chemotherapeutikum zur Behandlung von HER2-positiven Brustkrebspatientinnen eingesetzt wird. Trotz der insgesamt zytotoxischen Wirkung der Behandlung können einige Krebszellen überleben, indem sie DNA-Brüche erfolgreich reparieren oder die ursprünglichen Schäden in Mutationen umwandeln. Langfristig kann ihre anhaltende Präsenz zu Therapieresistenz, Krebsrückfällen und sekundären Malignomen führen. Die Wechselwirkung zwischen den durch Medikamente verursachten DNA-Schäden und den Reparaturwegen hinterlässt charakteristische Muster, die so genannten Mutationssignaturen. Während für einige Chemotherapeutika Basensubstitutionen oder kleine Indelsignaturen beschrieben wurden, sind die Muster und Auswirkungen von Strukturvarianten (SVs) noch nicht ausreichend erforscht. In dieser Arbeit untersuchte ich die SV-Belastung und die transkriptomischen Veränderungen, die durch Doxorubicin gefördert werden, auf Einzelzellebene in einem murinen Brustdrüsenorganoidmodell von HER2-positivem Brustkrebs.

Einzelzell-Transkriptom-Untersuchungen zeigten die drei wichtigsten Zelltypen, die in den Brustdrüsenorganoiden der Maus vorkommen, sowie globale als auch zelltypspezifische Veränderungen, die durch Doxorubicin ausgelöst werden.

Darüber hinaus wurde die Einzelzell-DNA-Template-Strangsequenzierung (Strand-seq) weiterentwickelt und hier zum ersten Mal angewandt, um die Doxorubicin-assoziierte genomische Instabilität in murinen Organoiden zu charakterisieren. Dank einer neuartigen computergestützten Multi-Zell-Methode, scNOVA, wurden genomische Daten mit Messungen der Nukleosomenbelegung (NO) integriert, um eine gleichzeitige SV-Erkennung und Zelltyp-Klassifizierung in derselben Zelle zu ermöglichen. Die Integration von Strand-seq und scNOVA zeigte, dass Doxorubicin mit einer höheren SV-Belastung und einer erhöhten Häufigkeit von Schwesterchromatidaustauschen in allen drei Zelltypen von Brustdrüsenorganoiden der Maus verbunden ist. Deletionen und komplexe Rearrangements erwiesen sich als mögliche Mutationsmuster von Doxorubicin.

Zusammengenommen veranschaulichen die in dieser Arbeit vorgestellten Ergebnisse die synergistische Integration von Organoidmodellen mit multizellulären Einzelzellmessungen für die systematische Untersuchung heterogener Populationen und machen deutlich, dass bei der Suche nach therapieassoziierten Mutationssignaturen SVs einzuschliessen sind.

Acknowledgments

Doing a PhD is hardly ever an easy and straightforward process, but it is also one of the reasons why it is a truly unique and shaping experience. While writing this very last part of my thesis and reflecting on the 4.5 years at EMBL with the all ups and downs, I am proud of how much I have grown as a scientist and a person.

I would like to thank my supervisor Prof. Dr. Jan Korbel for giving me the opportunity to pursue PhD in his group, and showing me the value of patience and persistence, both important for impactful research. I am grateful for all the feedback that I got from the members of my thesis advisory committee: Prof. Dr. Michael Boutros, Prof. Dr. Oliver Stegle, Dr. Martin Jechlinger, and especially Dr. Simone Köhler, who took additional time and effort to discuss with me my results. I would also like to thank Prof. Dr. Stefan Wiemann and Dr. Kyung-Min Noh who agreed to join my defense committee.

I was very fortunate to be a part of the Korbel group, and I would like to thank all the present and past members with whom I shared the daily joys, and from whom I got support and encouragement in the low moments. Throughout this PhD I had a chance to work closely with extremely talented yet humble scientists. Finishing this project would not be possible without the help of Dr. Hyobin Jeong, truly the ideal collaborator. Hyobin is not only a skilled scientist with a strong work ethic but also a great friend. We spent lots of time together outside of the lab (mainly eating very good food) and I will always remember our semi-illegal dinners during lockdowns while ‘trapped in Bahnstadt’. I was very lucky to share the desk with Dr. Karen Grimes. I probably never said it out loud or thanked her directly, but I learned tons from Karen, both when it comes to single-cell analysis and also (maybe even more importantly) crucial skills needed to ‘survive’ PhD. My scientific progress would not be the same without the support of Dr. Marco Cosenza. I am impressed with Marco’s knowledge (and a bit jealous of it) and I am grateful that he always had time to answer my questions and provide some clarity when I was confused. Having the chance to meet and learn from Hyobin, Karen, and Marco will be definitely one of the highlights of my time at EMBL. I would like to thank Dr. Ashley Sanders who introduced me to the world of Strand-seq and provided supervision during the first year of my PhD. I would like to highlight the work of the technicians from the group: Dr. Eva Benito Garagorri, Catherine Stober Brasseur, Patrick Hasenfeld, Dr. Maise Gomes Queiroz, and Benjamin Raeder. There were moments I asked you for something seemingly impossible, and yet you were able to deliver that.

I would like to thank the members of the Jechlinger group: Martin, Dr. Sylwia Gawrzak, and Marta Garcia Montero for the successful collaboration. Marta, a lab manager and a technician with golden hands, taught me the organoid culture and provided much-needed encouragement at the beginning of my PhD. I admire Sylwia's rigorous approach to science and I am thankful for all the career advice that I got from her.

During the PhD I relied heavily on the outstanding service offered by the members of EMBL facilities: Flow Cytometry Core Facility and Genomics Core Facility, as well as the lab kitchen. Your work made many of my experiments much easier.

I would not be here without all the help that I got from my family. My parents, Dorota and Jerzy Starosteccy, made sure that I had access to excellent education and encouraged me to learn foreign languages. This is the best investment in my future that I could have ever asked for. I am also extremely grateful for the honest advice that I have always gotten from my sister Ewa (I should actually say Dr. Ewa Czerwińska), regardless of how far we are from each other. You really set the bar high for yourself and I am very proud of you.

I would like to thank my friend Pavle Boskovic. I am glad that we shared the experience of doing a Masters and PhD together (which very often included exchanging countless memes and gossip). Pavle is one of the smartest people that I know and a very dedicated scientist. I am excited to hear about his next discoveries in the future.

I find it surprisingly difficult to express in words my gratitude to Joseph Christian Somody. So many fantastic moments and adventures, so much support and so much to look forward to. Thank you, Chris.

Table of contents

List of abbreviations

ANOVA	analysis of variance
B	basal
bp	base pair(s)
C	Crick strand
CFS	common fragile site
CNN	convolutional neural network
DAPI	4’,6-diamidino-2-phenylindole
DC	ductal carcinoma
DCIS	ductal carcinoma *in situ*
Del	deletion
DMSO	dimethyl sulfoxide
dox	doxycycline
DSB	double-strand break
dxr	doxorubicin
ER	estrogen receptor
F	fibroblast
FACS	fluorescence-activated cell sorting
FBS	fetal bovine serum
FISH	fluroescence *in situ* hybridization
FPKM	fragments per kilobase of exon per million mapped fragments
FSC	forward scatter
GO	gene ontology
GSEA	gene set enrichment analysis
GST	glutathione-S-transferase
H	haplotype
HER2	human epidermal growth factor receptor 2
HPV	human papillomavirus
HR	hormone receptor
ICGC	International Cancer Genome Consortium
IGV	Integrative Genomics Viewer
indel	insertion or deletion mutation
Inv	inversion

InvDup	inverted duplication
kb	kilo base pairs (1x10^3 bp)
LB	lobular carcinoma
LCIS	lobular carcinoma *in situ*
LP	luminal progenitor
MaSC	mammary stem cell
Mb	mega base pairs (1x10^6 bp)
MEBM	mammary epithelial cell basal medium
MEF	mouse embryonic fibroblast
MEpiCGS	mammary epithelial cell growth supplement
mESC	mouse embryonic stem cell
ML	mature luminal
MMP	matrix metalloproteinase
MMTV-LTR	mouse mammary tumor virus long terminal repeat
MNase	micrococcal nuclease
MRD	minimal residual disease
My	myoepithelial
NO	nucelosome occupancy
ns	non-significant
PARP	poly-(ADP-ribose) polymerase
PBS	phosphate-buffered saline
PCAWG	Pancancer Analysis of Whole Genomes
PLS-DA	Partial Least Squares Discrimination Analysis
PR	progesteron receptor
QC	quality control
RNA-seq	RNA sequencing
ROS	reactive oxygen species
RT-qPCR	reverse transcription quantitative real-time PCR
rtTA	reverse tetracycline-dependent transcriptional activator
SA-β-gal	senescence-associated β-galactosidase
SASP	senescence-associated secretory phenotype
scATAC-seq	single-cell sequencing assay for transposase-accessible chromatin
SCC	side scatter
scDNA-seq	single-cell DNA sequencing
SCE	sister chromatid exchange

scMNase-seq	single-cell micrococcal nuclease sequencing
scNOVA	single-cell Nucleosome Occupancy and Genetic Variation Analysis
scRNA-seq	single-cell RNA sequencing
scTRIP	single-cell tri-channel processing
sn	single-nucleus
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SV	structural variant
Strand-seq	single-cell DNA template strand sequencing
TetO	tetracycline operator
TF	transcription factor
TOP2A	topoisomerase IIα
TOP2B	topoisomerase IIβ
TOPBP1	DNA topoisomerase 2-binding protein 1
TSS	transcription start site
TTS	transcription termination site
UMAP	Uniform Manifold Approximation and Projection
VIP	Variable Importance in Projection
W	Watson strand

List of figures

List of tables

Chapter 1 Introduction

1.1 Breast cancer: disease characteristics

The extraordinary importance of mammary glands in the evolutionary success of mammals is best illustrated by the fact that this entire group of animals is named after *mamma*, a Latin word for breast. The mammary glands are present in both females and males but the male breast tissue is residual, and the other aspects and functions of this organ are clearly sexually dimorphic[1]. The female mammary gland is a branching structure composed of ducts and alveoli (clustered into lobules) that undergo development and differentiation until adulthood to enable the production of milk during lactation[2] (Figure 1A). Although all these sophisticated morphological changes are tightly regulated over time by different hormones and signaling pathways, in certain cases the mammary epithelial cells would start dividing uncontrollably resulting in cancer. Breast cancer is the most frequent cancer type affecting women worldwide (11.7% of all cancer cases reported in 2020[3]). Breast cancer in men is very rare with approximately 1% of all breast cancers[4]. Because of its high prevalence in females, breast cancer has got sufficient attention in developed countries to better detect and manage the disease through, for example, mammography and genetic counseling screening programs, or the development of targeted therapy. Early-diagnosed breast cancer with no detectable metastasis can be potentially curable but up to 30% of all breast cancer patients will experience relapse or metastasis, months or years after the initial treatment[5]. Metastatic breast cancer remains incurable with currently available methods which makes breast cancer the leading cause of cancer deaths in females in the majority of countries (15.5% of the total cancer deaths in females; much higher fatality rates are reported in underdeveloped regions[3]). The challenge to cure breast cancer continues mostly because the disease is highly heterogeneous with different subtypes, molecular characteristics, treatment options, and survival chances[6].

1.1.1 The clinical significance of tumor heterogeneity

By definition, heterogeneity is a state of being diverse in content. In cancer biology, heterogeneity includes interpatient heterogeneity, intrapatient heterogeneity, and intratumor heterogeneity. Interpatient heterogeneity or intertumor heterogeneity refers to differences between different patients diagnosed with presumably the same cancer type based on its morphological features[7]. Intrapatient heterogeneity extends to the tumors in the same patient as cancer cells forming the primary tumor are different from the ones in metastases[8].

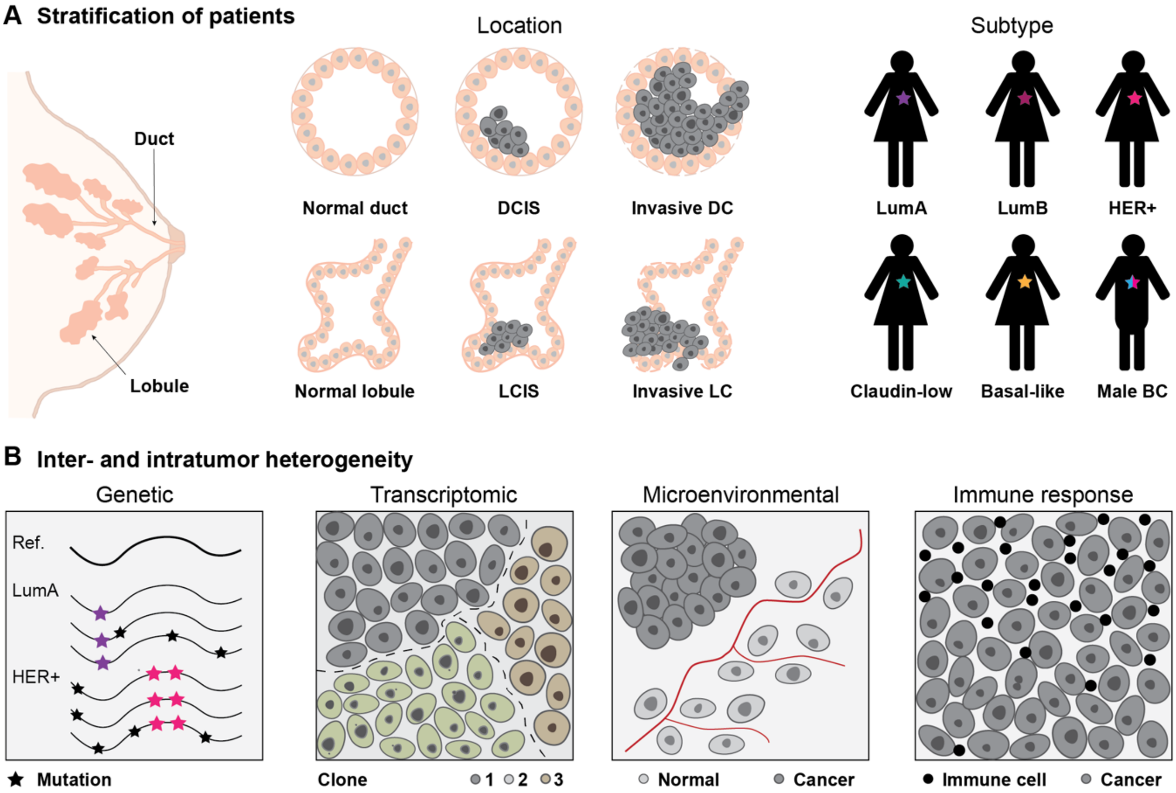


Figure 1 Different types of heterogeneity in breast cancer.

Tumor heterogeneity manifests as the presence of subpopulations of cancer cells with distinct genotypes and/or phenotypes. (A) Patients diagnosed with breast cancer are stratified based on the location of the tumor or stages, and subtypes (subtypes explained in detail in the main text). The presence of abnormal cells can be restricted only to a milk duct or lobules, causing ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS). If the cancerous cells start infiltrating nearby tissues or distant organs, invasive ductal carcinoma (DC) or invasive lobular carcinoma (LC) is diagnosed.
(B) Intertumor and intratumor heterogeneities in breast cancer involve different cell populations that may differ with respect to a high number of capabilities, including genetic changes, differentiation or transcriptomic state, as well as extrinsic factors such as the interaction with the microenvironment or the infiltration of the immune system.

Intratumoral heterogeneity encompasses the diversity of cells within a single disease site[9]. All aspects of tumor heterogeneity influence tumor progression and impact treatment choices (Figure 1B). Considering the significance of interpatient heterogeneity, each patient would ideally receive a targeted therapy tailored to their specific cancer case. Such a scenario is however impossible due to the economic limitations and a finite number of available drugs. Because of that, the patients need to be stratified to direct care.

1.1.2 The molecular subtypes of breast cancer

In the context of breast cancer, the grouping of patients can refer to the location of the tumor, or, more importantly, its histopathological status of estrogen receptor (ER), progesterone receptor (PR), and epidermal growth factor receptor 2 (HER2)[10] (Figure 1A). Almost all instances of breast cancer originate in the lobules or ducts, leading to lobular or ductal carcinoma, respectively. In the earlier stages of cancer, uncontrolled growth of cells is limited only to the lumen of mammary glands without spreading outside of lobules or ducts (*in situ*). In patients with invasive cases or metastasis, cancer cells are also detected in the surrounding breast tissue or other organs. To provide the most optimal treatment, breast cancer was initially divided into three main groups (ER/PR-positive, HER2-positive, and triple-negative). Although such an approach is straightforward and still commonly used in clinics, it is rather outdated and does not fully represent the spectrum of genetic mutations and gene expression patterns in primary breast tumor samples. Therefore, doctors are increasingly using genetic information about breast cancers to better guide decisions about treatment and infer survival chances. Based on molecular profiling at least five major breast cancer subtypes can be classified: luminal A, luminal B, HER2+, basal-like, and claudin-low[11]. Luminal tumors are most frequent with subtype A being the most commonly diagnosed one (approximately half of the cases)[12]. Both subtypes A and B are hormone receptor-positive but the levels of ER and PR are lower in subtype B[13]. Luminal subtype B shows higher expression of proliferative marker Ki67 and may have amplified HER2, which leads to a much poorer prognosis compared to subtype A[13]. The HER2-positive group of tumors has *ERBB2* locus amplification (encoding HER2) with varied hormone-receptor status (generally negative). Amplification of the *ERBB2* leads to a higher number of HER2 receptors on the cell surface. HER receptors contain three important domains: an extracellular ligand-binding domain, a transmembrane domain, and an intracellular tyrosine kinase domain. Upon ligand binding, homodimerization or heterodimerization is induced leading to the activation of pro-survival downstream signaling pathways. HER2 amplification and overexpression drive hyperactivation and abnormal signaling, also in the absence of the ligand[14]. HER2-positive subtype accounts for ~20% of breast cancer cases and is also associated with low survival chances[15,16]. Basal-like and claudin-low cancers represent the most aggressive subtypes[17], predominantly triple-negative (ER-, PR-, HER2-). Basal-like tumors are characterized by the expression of basal epithelial cytokeratins (5, 14, 17), while claudin-low tumors have decreased levels of tight-junction and adhesion proteins (like E-cadherin or claudins)[13]. 90% of male breast cancers are hormone receptors (HR)-positive and HER2-negative[4].

In the context of this thesis, it is important to remember that cancer is not simply an agglomerate of identical mutated cells but rather a complex structure with spatial and temporal changes, also affected by the surrounding microenvironment[7]. Normal tissue functions in an environment where access to oxygen and nutrients through vasculature is optimized.

However, a cell growing in a solid tumor, especially with disorganized blood vessels, is exposed to abnormal levels of oxygenation, growth factors, and pH which may promote dissemination and start the metastatic cascade[9,18]. Breast cancer subtypes differ in the metastasis-free period and site of relapse. ER-negative cancer tends to spread more to visceral organs (lung, liver, and brain) and bones, while ER-positive metastases show tropism mainly toward bones[19]. Microenvironment heterogeneity includes also heterogeneity in immune cell infiltration, in particular, tumor-infiltrating lymphocytes are clinically relevant biomarkers of breast cancer prognosis. Different breast cancers show different levels of immune infiltration with more aggressive and genomic unstable subtypes being more infiltrated (probably due to fostered generation of neoantigens) than luminal subtypes. The infiltration also decreases during disease progression which additionally impacts the potential response to immunotherapy[20].

1.1.3 Genetic heterogeneity in breast cancer and the role of SVs

Breast cancer subtypes are characterized by overall different mutational burdens and different frequently mutated genes. Describing genetic heterogeneity between subtypes seems to be the most studied type of tumor heterogeneity as the advancements in bulk DNA sequencing technologies enable better and faster annotation of tumor-associated mutations and their functional impact. For example, mutations in GATA3 are found almost exclusively in luminal A subtype[21] which is also considered to have the lowest mutation rate and usually almost completely diploid genome[22]. This is in stark contrast to basal-like tumors that are genetically unstable and show a high frequency of chromosome number alterations[23]. Interestingly, the basal-like subtype is frequently developed by patients with a germline *BRCA1* mutation[23]. *BRCA1* similar to some other known breast cancer susceptibility genes (like *ATM*, *BARD1*, *BRCA2*, *CHEK2*, *PALB2*, *RAD51C*, *RAD51D*, and *TP53*)[24], is involved in the DNA damage repair pathway which emphasizes the crucial role of DNA integrity in breast epithelium. Importantly, the majority of studies on mutations in breast cancer have focused on identifying single nucleotide variants (SNVs), relatively easy with next-generation sequencing data. However, breast cancer has on average one of the lowest frequencies of SNVs in solid tumors[25] suggesting that structural variants (SVs) may play an important role in tumorigenesis.

SVs are a class of mutations, next to point mutations (including SNVs) and small insertions or deletions (indels). From these three groups, SVs are comprised of larger genomic changes as they are bigger than 50 base pairs (bp) and can scale up to entire chromosomes[26]. SVs include deletions, insertions, duplications, amplifications, inversions, translocations, chromosomal losses or gains (aneuploidy), and complex DNA rearrangements patterns like chromothripsis, chromoplexy, and breakage-fusion-bridge cycles[27] (methods to detect them are explored more in detail in chapter 1.4.1). Various processes of DNA repair and/or DNA replication contribute to the formation of such a variety of rearrangements, and only for a fraction of them there is a clear mechanistic explanation of how they occur post-DNA break or during a faulty cell division[28]. Viral integration (for example, of human papillomavirus (HPV)) or activation of mobile elements can also result in the formation of SVs[27]. The functional consequences of SVs preserved in the genome can be dreadful leading to congenital disorders[26] or tumorigenesis. For example, SVs may disrupt genes or their regulatory elements, affecting the three-dimensional organization of the genome or creating novel protein-coding gene fusions (change of DNA order)[27,28]. DNA copy number changes may result in the deletion of tumor suppressor genes or amplification of oncogenes (change of DNA dosage). In fact, in the case of HER2-positive tumors, the overexpression of HER2 is a result of gene amplification, for example as a consequence of repeated breakage-fusion-bridge cycles[29].

Early studies with copy number arrays allowed to characterize chromosome gains and losses in breast cancer subtypes, beyond HER2 amplification,[30] but annotation of complex rearrangements has been challenging. Only recently, thanks to the efforts such as the International Cancer Genome Consortium (ICGC) and the Pancancer Analysis of Whole Genomes (PCAWG) consortium, a more detailed description of the SV landscape in breast cancer has been reported identifying breast cancer as one of the types most affected by chromothripsis and reporting an increased burden of small (less than 10 kilo base pairs (kb)) somatic SV deletions and tandem duplications associated with tumors with germline *BRCA1* and *BRCA2* mutation[31,32].

Within a tumor, newly emerging mutations may affect gene expression patterns leading to transcriptional heterogeneity. However, such phenotypic heterogeneity may be also a consequence of non-genetic factors such as stochastic changes in gene regulation[9]. It is worth noting that before the development of single-cell techniques, all forms of intertumor and intratumor heterogeneity have been predominantly described based on a regional sampling of tumors, and even a few biopsies taken from a single tumor may not fully represent the extent and complexity of different clones[33]. All the factors: genetic, epigenetic, phenotypic or microenvironment heterogeneity, both intertumor and intratumor, may influence how cancer cells respond to treatment.

1.2 The opportunities and challenges to defeat cancer

1.2.1 Treatment options for breast cancer

Breast cancer was possibly the earliest ever written description of the disease. According to this first report, found in the Edwin Smith Papyrus that dates back to approximately 3000 BC, there was no treatment for breast cancer. As medicine developed, in particular in Asian countries, more successful attempts at treating breast cancer were introduced by applying herbal remedies but the goal was rather to minimize the pain than cure the patient. The physicians in the oldest known human civilizations recommended early and aggressive surgical therapy to remove the tumor[34], and until now surgery has been usually the first type of treatment for primary breast cancer[35,36]. Radical mastectomy introduced by Halsted at the end of the 19th century included the removal of the whole breast, all the lymph nodes under the arm, and the chest wall muscles, regardless of the type and size of the tumor, or the patient's age. This approach is very different from surgeries performed now with the goal of breast conservation (without compromise to the control of the disease) or immediate reconstruction. Depending on the specific case, patients will receive additional therapy before or after the surgery, including radiotherapy, hormone (endocrine), targeted therapy, or chemotherapy[35,36]. HR-positive breast cancer cells depend on the estrogen and/or progesterone signaling pathways, so the goal of hormonal therapy is to lower the amounts of these hormones produced in ovaries (in premenopausal women) or body fat and muscle (post-menopause), or to block their effects[37,38]. Patients with HR-positive disease (particularly with luminal A subtype) are recommended endocrine therapy with tamoxifen or in the case of postmenopausal women, aromatase inhibitors (anastrozole, letrozole, and exemestane)[36]. Targeted therapy is offered in combination with chemotherapy for HER2-positive cases. HER2 targeting agents are available as monoclonal antibody (Trastuzumab) or small molecule inhibitors (lapatinib, neratinib)[16,39]. From 2018, metastatic HER2-negative breast cancer patients with germline, cancer-associated *BRCA1* or *BRCA2* mutations are eligible for treatment with the poly-(ADP-ribose) polymerase (PARP) inhibitors (olaparib and talazoparib). PARP inhibitors induce synthetic lethality in BRCA mutated cells that rely on PARP as an alternative DNA damage repair mechanism[40,41]. Systemic chemotherapy is generally recommended for all patients. For patients with subtypes that are not adequate for hormonal or targeted therapy, chemotherapy is the only option for drug treatment[36]. Depending on the country and its healthcare system, and breast cancer subtype, there are several standard chemotherapy variants, usually with an anthracycline (typically doxorubicin) and/or a taxane (most commonly paclitaxel)[36]. Other regimens may include platinum agents (carboplatin, cisplatin), alkylating agents (cyclophosphamide), or an antimetabolite (5-fluorouracil). Although different groups of chemotherapeutics have different mechanisms of action, they are administered to, ideally, kill cancer cells, or stop them from dividing. However, chemotherapy is associated with significant side effects due to the harm to normal healthy cells[42].

1.2.2 Chemotherapy-associated mutational signatures

Direct or indirect DNA damage is a mechanism through which the most common chemotherapeutics, as well as radiotherapy, are designed to induce cell death. This may seem counterintuitive as DNA damage and genomic instability are both causes and consequences of cancer. DNA breaks impair DNA transcription and RNA polymerases leading to limited gene expression or DNA replication, stalling the cell cycle progression and proliferation. If the DNA damage load is too high or DNA repair is unsuccessful, apoptosis will be induced[43]. However, cancer cells can show abnormal DNA repair activity or escape apoptosis despite persistent DNA damage[44]. Such a response comes with a trade-off, as even cancer cells cannot sustain all tasks like proliferation, evading cell death, and immune suppression at the same time as they all require energy resources and a favorable microenvironment[45]. Because of that, cancer cells exposed to a drug may either die if not resistant, enter the state of drug tolerance manifested by being able to survive but not proliferate, or inhibit further growth for some time until DNA is repaired (usually for hours or days)[44]. The repair of DNA breaks may be successful and leave no trace, or if erroneous, result in the formation of mutations[43]. It has been speculated that chemotherapies may leave specific imprints, called mutational signatures[46,47], in the genomes of exposed cells and their progeny. Mutational signatures are patterns of modifications of single or a few consecutive nucleotide bases (also sequence-context-dependent), or indels, shared by individuals with diseases of similar etiology or exposed to the same mutagens[47]. Mutational signatures have been identified for platinum-based drugs[48–51] and 5-fluorouracil[51,52]. These studies rely on the statistical analysis of cancer genomes in exposed patients, or experimental model systems. The research in the area of mutational signatures has been focusing on single or doublet base substitutions and indels but virtually nothing is known if chemotherapies promote the formation of SVs. The correlation between the treatment with a chemotherapeutic drug and the emergence of SVs was explored in detail in my PhD project.

1.2.3 Doxorubicin: mechanism of action and metabolism

Of all the chemotherapeutics, doxorubicin is the main focus of this thesis. Doxorubicin is an anthracycline antibiotic isolated first from one of the *Streptomyces* strains. Other clinically important anthracyclines include daunorubicin, idarubicin[53] and epirubicin[54]. Doxorubicin has a broad spectrum of use, treating both adult and childhood cancers, such as acute leukemia, non-Hodgkin lymphomas, soft tissue sarcomas, and solid tumors[55]. Doxorubicin can be included in the treatment regimens for most types of invasive and metastatic breast cancers, also together with targeted therapy for HER2-positive cases. There are several proposed mechanisms for how it kills cancer cells, but the exact mode of action seems to be also influenced by the dose of the drug (which depends on the cancer type, duration of treatment, gender, body mass, and/or age)[55]. Doxorubicin directly targets DNA by intercalating into the helix causing topological and torsional stress that leads to stalling of DNA and RNA polymerases. It inhibits topoisomerase IIα (TOP2A) and IIβ (TOP2B), enzymes that regulate DNA superhelical states and relax positive supercoils. TOP2A and TOP2B share catalytic and most structural properties but they are not functionally redundant. TOP2A is the major form of the enzyme in dividing cells, while TOP2B is expressed in non-dividing cells, for example, cardiomyocytes[56]. TOP2 enzymes act as homodimers that catalyze the cleavage and re-ligation of DNA. During that process, both strands of DNA are cut and TOP2 temporarily attaches to the 5' ends of cleaved DNA via covalent phosphotyrosyl bonds. As a result, the DNA ends are protected and are not recognized by cellular DNA repair pathways[57]. Poisoning of TOP2A and TOP2B results in the enzymes being trapped and stabilized in a state that they cannot reseal DNA double-strand breaks[55,58] (DSBs). Trapped TOP2 complex is primarily removed by ubiquitin-dependent proteolytic degradation or SUMOyaltion-induced direct hydrolysis of the bonds between TOP2 and DNA. After the removal of the trapped TOP2 from DNA, DSBs on naked DNA are recognized by the repair pathway machinery[57]. Doxorubicin interferes also with metabolic processes including the mitochondria's respiratory functions (by contributing to the production of reactive oxygen species (ROS)), calcium and iron homeostasis, or ceramide production[56]. Very high doses of doxorubicin (9 μM) used in one study caused histone eviction from open chromosomal areas in a human melanoma cell line (MelJuSo) and AML patients post-treatment[59]. Lower concentrations of doxorubicin have been shown to cause chromatin damage and induce nucleosome turnover around promoters in murine squamous cell carcinomas cell lines [60]. In a more recent study, the authors showed that anthracycline variants that induce DNA damage but not DSBs have similar anticancer activity in cell lines, mice, and human acute myeloid leukemia patients indicating a significant role of chromatin damage as a cytotoxic mechanism of these drugs contributing to the cardiac dysfunction[61]. Interestingly, it was demonstrated in a different study that breast cancer patients show a naturally occurring variation in the expression level of certain chromatin regulatory genes (including the members of Polycomb and Trithorax complexes) that dictates chromatin accessibility and sensitivity to anthracycline therapy[62]. This has important implications to better stratify patients and distinguish those who will benefit from therapy based on the signature pattern of these chromatin regulatory genes.

So far, a mutational fingerprint of doxorubicin has not been specified. According to one study that aimed to assay the mutagenic impact of common chemotherapeutics in an established DNA repair model system (chicken DT40 lymphoblast cell line), doxorubicin had no detectable mutagenic activity as measured by single or doublet base substitutions and small indels[63]. A more recent analysis of 570 advanced and metastatic cancer patients (including breast cancer patients) associated doxorubicin with SBS17b mutational signature (single-base substitution, T>G with nonrandom sequence context at -2 bp and +2 bp[46]) but it was more frequent in chemotherapy regimens combining both doxorubicin and platinum-based compounds[64]. In one of the studies from the Korbel group, doxorubicin was used as a perturbation agent that promotes DNA alterations in an untransformed cell line (human hTERT RPE-1 retinal pigment epithelial cell line) leading to complex genomic rearrangements giving growth advantage[65]. These examples suggest that treatment with doxorubicin may be associated with the formation of SVs, rather than base substitutions or indels.

Doxorubicin appears on the List of Essential Medicines created by the World Health Organization but it is not a perfect drug. In fact, it is associated with severe side effects including heart muscle damage (both acute or chronic, appearing later in life)[66]. There is a correlation between the dose of anthracycline (cumulative exposure) and the risk of cardiomyopathy, and there is also growing evidence that genetic variation (in form of single nucleotide polymorphisms (SNPs) in, for example, genes associated with drug transformation or DNA repair) contributes to the chemotherapy-related cardiac dysfunction[67]. Due to its characteristic red color (highlighted also by 'ruby' in doxorubicin) and fatal complications, doxorubicin is called 'red devil'.

Doxorubicin is injected intravenously, quickly achieves high concentration in blood, and is rapidly absorbed into tissues but the elimination takes much longer (half-life of 24-36 hours).

Despite the fast distribution of the drug and high penetrance to cells, doxorubicin cannot pass the blood-brain barrier[68].

The metabolism and pharmacokinetics of doxorubicin are fairly well described but different pathways might be involved in its metabolism in cardiac cells (covered in detail in a recent review[67]). Several transporters are involved in influx (SLC22A16[69], and supposedly heart-specific: SLC28A3, SLC10A2, SLC22A17[67]) and efflux (ABCB1, ABCC1, ABCC2[70], ABCG2, RALBP1[71]) of doxorubicin. Doxorubicin also enters the cell via passive diffusion[68]. Once inside the cell, doxorubicin is metabolized through one of three main routes: one-electron reduction, two-electron reduction, and deglycosidation[72]. However, approximately 50% of the drug is eliminated from the body unchanged[68]. The major pathway involves a two-electron reduction of doxorubicin to a secondary alcohol, doxorubicinol. Depending on the cell type, different cytoplasmic NADPH-dependent carbonyl, and aldo-keto reductases can carry out this reaction, for example, CBR1 and CBR3 in liver[73,74], AKR1A in heart[75]. Doxorubicin-semiquinone radical is formed as a result of the one-electron reduction. This reaction can be catalyzed by several oxidoreductases located in different organelles inside the cell: mitochondrial NADH dehydrogenases (in mitochondria and sarcoplasmic reticulum): NDUFS2, NDUFS3, NDUFS7; NADPH dehydrogenase (NQO1, cytosolic), xanthine oxidase (XDH, cytosolic), nitric oxide synthases (NOS1, NOS2, NOS3, cytosolic)[66]. During the re-oxidation of the radical to doxorubicin, ROS and hydrogen peroxide are formed. ROS can be deactivated by glutathione peroxidase (GPX1), catalase (CAT), and superoxide dismutase (SOD1)[55]. The third, minor way, hydrolytic or reductive deglycosidation leading to the formation of 7-hydroxy- or 7-deoxyaglycones, is less characterized. Aglycones do not have cytotoxic activity but they have been suggested to be cardiotoxic[72].

As outlined in this chapter, doxorubicin is a relatively well-studied drug but many aspects of its metabolism and mechanism of action are elusive. Considering the clinical importance of doxorubicin, more research is still needed to fully understand its pharmacodynamics and characterize why some cancer cells stop responding to the treatment with anthracyclines.

1.2.4 Mechanisms of resistance during tumor evolution

Cancer cells vary widely in their susceptibility to chemotherapeutics, including doxorubicin, and very often show resistance to the treatment. Resistance to cancer therapies falls into two categories: primary (intrinsic) and secondary (acquired). Patients with primary resistance do not respond at all to the therapy, those with acquired one experience reoccurrence of the disease after the initial successful treatment. Such acquired resistance is most commonly associated with the effect of the genetic evolution of cancer in response to the therapeutic challenge but the role of non-genetic mechanisms has been more and more recognized. Non-genetic reprogramming allows cells to adapt faster to changing conditions[76], and such phenotypic plasticity has been recently accepted in the field as yet another hallmark of cancer[77]. Regardless of how the resistance has been achieved, the most straightforward method to reduce the toxic effect of the drug is to lower its concentration inside the cell through, for example, increasing its secretion by upregulating the levels of efflux transporters in the plasma membrane[44]. The members of the ABC gene family are among the most commonly reported multifunctional transporters involved in multidrug resistance and a wide range of cancer drugs, including doxorubicin, are their substrates. In the context of breast cancer resistance, the major efflux transporter protein is ABCG2 (also called BCRP, breast cancer-resistant protein). The other strategy exploited by cancer cells is to inactivate the drug with detoxifying enzymes such as aldehyde dehydrogenases or the members of the glutathione-S-transferase (GST) family[78]. Drug treatment may induce dramatic metabolic changes, beyond the detoxification of xenobiotics, and metabolic adaptation is getting more recognition as a therapy resistance mechanism. In a recent study, the authors showed using human breast cancer cell lines and a xenograft model that the resistance to two anthracyclines, doxorubicin, and epirubicin, is mediated through metabolic adaptations but there is a different mechanism for each drug. Doxorubicin-resistant cells used glutamine to drive oxidative phosphorylation and cells resistant to epirubicin significantly upregulated the mitochondrial ATP production[79]. As described in chapter 1.2.1 the treatment of certain breast cancer subtypes includes also different drugs than chemotherapeutics. Resistance to targeted therapies manifests through the activation of alternative signaling pathways or engagement of upstream or downstream effectors[80]. For example, HER2-positive patients treated with HER2-targeting agents like Trastuzumab or lapatinib may stop responding to the drug as tumors may lose HER2 expression due to therapeutic pressure, or promote a constitute activation of the downstream phosphoinositide 3-kinase (PI3K) pathway without reliance on HER2 signaling[80].

In the next paragraph, I will summarize the ultimate solutions that cancer cells can implement to avoid apoptosis due to drug-induced pressure.

1.2.5 Dormancy and its implications for MRD formation

Extreme cases of sustained growth arrest with increased resistance to cell death in response to stress include senescence and dormancy. It is difficult to distinguish these states with clear definitions as there are no 'gold standard' markers for each of them, and especially the words 'dormant' and 'dormancy' are used inconsistently[81]. For clarity, I will refer to senescent cells as the ones with mostly irreversible cell-cycle arrest, resistance to proliferative stimuli, increased activity of lysosomal senescence-associated β-galactosidase (SA-β-gal), and secretion of senescence-associated secretory phenotype (SASP) composed of pro-inflammatory cytokines, growth factors and matrix metalloproteinases (MMPs)[82]. They remain metabolically active but in an altered state[83]. Contrastingly, dormant cells have the potential to exit the cell-cycle arrest. They are characterized by decreased expression of proliferation marker Ki67 and reduced metabolic activity[84]. Such a state is under dynamic control of both cell-intrinsic factors (for example, activation of immune evasion mechanisms) and cell-extrinsic ones provided by the niche in which dormant cells reside[81,84]. It is still debatable if dormant cancer cells are/can be considered cancer stem cells[85]. Cells that survived the initial treatment and entered dormancy may get reactivated to form a relapse but not all dormant cancer cells are metastasis-initiating[81].

Dormant cells contribute to minimal residual disease (MRD) formed by the remaining cancer cells that are present in the patient after seemingly complete radiographical and pathological remission. MRD can appear in three different forms: locally found residual disease (not successfully removed during surgery), circulating tumor cells (found in the bloodstream), or disseminated tumor cells (that invaded different organs). The exact mechanisms of how cancer cells enter and leave dormancy are still unclear but there is a growing interest in treating MRD. Cells that contribute to MRD have already adapted to stress- and drug-induced selective pressure, therefore MRD has most likely less clonal complexity than the primary tumor. As a result, targeting MRD, presumably less heterogenous, could be more successful once a proper drug to which MRD cells are sensitive, is identified. Patients at this stage are also in better physical condition to undergo additional therapy but there is a risk of overtreatment and unjustified costs. It is better to prevent metastasis than to treat them, and targeting MRD may be one solution for blocking the development of metastasis[86,87]. However, that requires frequent use of sensitive MRD-assessment techniques (based on flow cytometry or sequencing) which are not routinely used in the clinics[86]. In an ideal scenario, relapses would not occur if the initial treatment has maximum effectiveness. Therefore, there is a constant need to better understand the resistance of tumors to currently used chemotherapies and to enhance or develop new therapeutic strategies with minimum side effects. Further in the thesis, I investigate whether treatment with doxorubicin promotes MRD phenotype. To do that I take advantage of murine mammary gland organoids. In the following section, I describe the benefits and challenges of applying murine mammary glands and organoids derived from them, to study resistance mechanisms and dormancy in breast cancer.

1.3 Murine mammary gland-derived organoids as a proxy for the human breast

1.3.1 Cellular composition of the human and mouse mammary gland

One of the toughest questions faced by any scientist is the choice of the right model to study and answer their biological question. The proper system should resemble the physiology and pathology of the relevant organism, organ, or tissue. The cellular composition of human and murine mammary gland epithelium is similar and mouse mammary glands are suitable models to study human mammary biology and associated diseases including breast cancer[88]. In both species, the mammary gland starts forming during mid-embryonic stages but shows more striking dynamic changes within its epithelium during puberty and reproductive times as a response to fluctuating hormones during estrous cycles (menstrual cycles in humans) and pregnancy. The mammary gland is formed by a complex epithelial ductal tree surrounded by a stromal matrix containing fibroblasts, adipocytes, endothelial cells, and immune cells. At a cellular level, the mammary gland is a bi-layered structure composed of two major lineages: luminal and basal. Luminal cells, found in the inner layer towards the central lumen, generate milk protein and secrete them during lactation. Basal cells adjacent to the basement membrane are contractile, which is crucial for milk ejection and its transport along the duct to the nipple[89,90]. There are certain anatomical differences between the mammary glands of humans and rodents, for example, in the human breast, the ductal tree is much more complex than the mice one, and several individual branched ductal networks lead to the nipple. In the case of mice, a network of ducts leads to a single primary duct that ends in the nipple[88]. Also, the mice stroma is adipocyte-rich in contrast to the more fibrous stroma present in humans[91]. Despite the essential role of the mammary gland and extensive research in this field, the mammary epithelial cell differentiation process is still inconclusive. Experiments with two important techniques, the mammary transplantation assay, and lineage tracing, confirmed that mammary epithelial differentiation is rather a hierarchical process with a multipotent, mammary stem cell (MaSC) giving rise to increasingly more lineage-restricted progenitors[89]. The development of novel single-cell technologies like single-cell RNA sequencing (scRNA-seq) and single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) allows to better understand how mammary cells differ during consecutive stages of development, and whether

MaSCs or only lineage-restricted stem cells exist postnatally. Surprisingly, different studies come to opposite conclusions, and the existence of MaSC in adults is still controversial[92]. Such confusion in the field highlights the urgent need to unify the findings from mammary gland development studies to describe which mammary cell populations universally exist, what their commonly used names are, and what their markers are for single-cell transcriptomic analysis or fluorescence-activated cell sorting (FACS). Proper annotation of cell types, in both human and murine mammary glands, is particularly important in the context of drug resistance in breast cancer as it remains debatable whether the resistance emerges from pre-existing clones of stem-cell-like properties or more differentiated cell states[93].

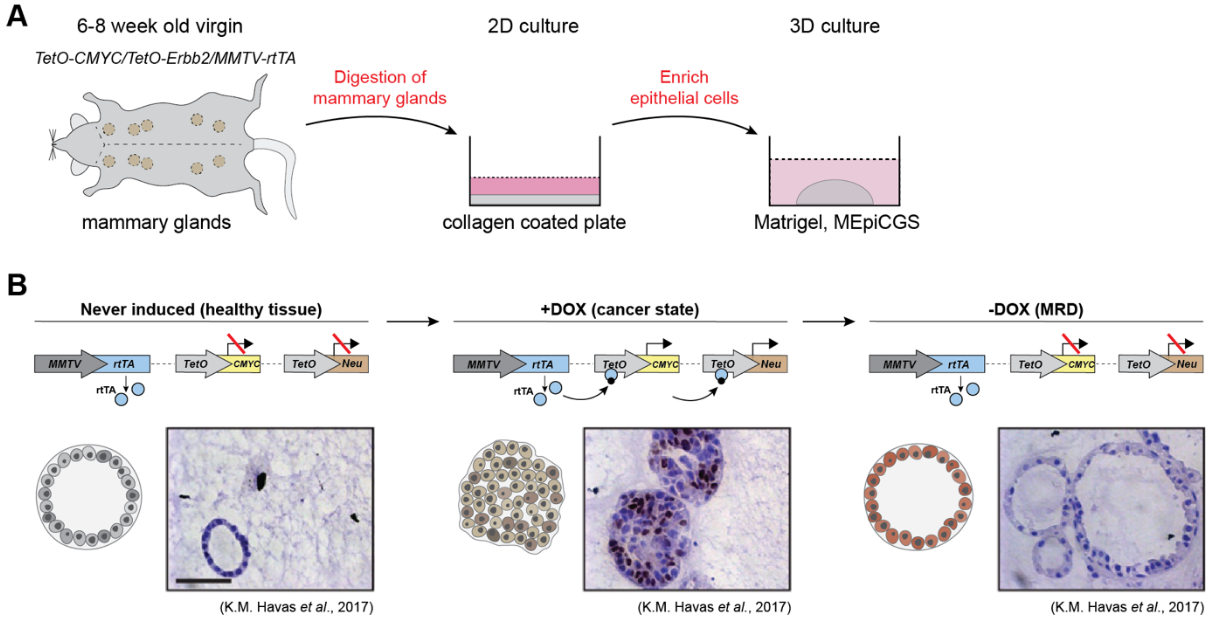
1.3.2 Organoid models for mammary gland studies and breast cancer

Progress to understand mammary biology and breast cancer formation would not be possible without using mice as an *in vivo* model for experimental manipulations. However, research including animals is slow, expensive, and the results are often difficult to interpret due to the study design[94]. Novel cell culture methods, beyond traditional two-dimensional monolayers, allow to better represent complex human diseases while minimizing the drawbacks of animal testing. Applying organoid technology to model different stages of tumorigenesis, especially the response to drugs is a very promising approach. An organoid is a self-organizing three-dimensional structure generated from pluripotent stem cells, adult stem cells, or somatic cells from normal or malignant primary tissue (either human or mice, but more exotic organoids have been also reported derived from, for example, crypts isolated from the intestines of bats[95], or snake venom glands[96]). Organoids retain the near-physiological cellular composition and if grown in a medium containing specific growth and stem-cell renewal factors, they can be expanded extensively. They recapitulate histological and genetic features of original tissue making them suitable for translational studies and drug screenings[97]. For breast cancer research and basic studies to understand mammary gland development, two types of organoids are especially relevant: derived from biopsies from breast cancer patients, or propagated from normal tissue (either from murine mammary gland organoids, or from samples collected during reduction mammoplasties (performed to reduce breast size), or from prophylactic mastectomies (for breast cancer prevention))[98]. There are various protocols to culture breast cancer or mammary epithelial organoids, which differ mainly regarding matrix type, medium components, and plating strategy. These protocols may or may not include additional purification steps (like FACS, differential centrifugation, or cell straining) to enrich certain cell types[99]. One attempt to increase reproducibility after following a universal protocol is to develop biobanks. For example, authors of two independent breast cancer organoid biobanks created in total almost 200 primary and metastatic breast cancer organoid lines that recapitulate histological and genetic features of original tumors, and represent different breast cancer subtypes[100,101]. The biobanks derived from patients' samples provide a great representation of interpatient heterogeneity crucial for high-throughput drug testing.

1.3.3 An inducible mouse model for studying tumor progression of HER2-positive cancer

In certain cases, such genetic, transcriptomic, or metabolic variation provided by biobanks to progress personalized medicine, may introduce too much complexity due to practically unidentifiable parameters. Because of that, in my project, I am using organoids derived from mammary glands from a well-defined, inbred, transgenic mouse strain (Figure 2A). Applying the TetO-cMYC/TetO-Neu/MMTV-rtTA tetracycline-inducible model of breast cancer allows studying the mechanisms of HER2-positive breast disease together with cMYC activation[102,103]. Overexpression of both HER2 and cMYC is a common case in breast cancer and correlates with aggressive phenotype and poor prognosis[104,105].

With the model used in this thesis, it is possible to temporally (with Tet-On[106]) and spatially (in the mammary gland, owing to mouse mammary tumor virus long terminal repeat (MMTV-LTR) sequence) control tumorigenesis. Upon addition of doxycycline (tetracycline-derivative) to the medium (in 3D cultures, or to the animal diet), the system allows for overexpression of two potent oncogenes (activated rat *Neu/Erbb2* (homolog of human HER2) and truncated human *CMYC* gene with exon 2 and 3) that are also dysregulated in Her2-positive breast cancer patients (Figure 2B). Tumor formation and maintenance are then dependent on the action of these two oncogenes, mimicking oncogene addiction. In the case of mammary gland organoids derived from this strain, the induction of oncogenes causes the transformation of the hollow acini into highly proliferative solid spheres that represent ductal carcinoma *in situ*. If the expression of oncogenes is silenced by the removal of doxycycline from the medium, the filled-in spheres regress to the re-polarized monolayer of viable cells that escape oncogene withdrawal and appear normal in morphology, corresponding to the dormant state of MRD seen in the clinic. As in breast cancer patients, after complete regression mice develop spontaneous relapse without doxycycline induction. Similar reactivation is observed in the organoids that increase proliferation rate without repeated induction making again solid structures[102] (not included in panel B of Figure 2). Previous work using mammary gland organoids derived from TetO-CMYC/TetO-Neu/MMTV-rtTA mice, and verified in the mouse model *in vivo*, as well as correlated with patient-derived samples, has characterized metabolic changes that drive tumor reoccurrence upon perfect targeted therapy at the driver oncogenes[103]. However, silencing of both oncogenes is an idealized scenario, as CMYC is still considered undruggable[107], and the experiments (RNA-seq, lipidomic profiling, untargeted and targeted metabolomic analysis, DNA methylation analysis) were performed in bulk not considering the impact of cell-type differences and single-cell heterogeneity. In chapter 1.4, I provide background information on the single-cell methods crucial to characterize responses to drug treatment with a particular focus on single-cell DNA template strand sequencing (Strand-seq) and computational pipelines created around it.



(K.M. Havas *et al.*, 2017) (K.M. Havas *et al.*, 2017) (K.M. Havas *et al.*, 2017)

Figure 2 Organoid culture of mammary glands from tritransgenic mice.

(A) For the project described in this thesis, I was culturing organoids from the mammary glands of tritransgenic mice (*TetO-CMYC/TetO-Neu/MMTV-rtTA*). The mammary glands of adult virgin mice are first digested overnight to remove fat and muscle tissue, and cultured for one night on collagen-coated plates to enrich the population of epithelial cells. A suspension of single cells is then mixed with Matrigel and cultured 3D in a well-defined medium containing mammary epithelial cell growth supplement (MEpiCGS). (B) With this model, depending on the status of the Tet-ON system for doxycycline-inducible gene expression, it is possible to mimic different stages of tumorigenesis, from healthy tissue, cancer state, and MRD. MMTV-LTR sequence restricts the expression of the reverse tetracycline-dependent transcriptional activator (rtTA) to mammary glands. If doxycycline is added to the medium, the doxycycline-rtTA complex binds to the tetracycline operator (TetO), driving the transcription of the transgene oncogenes and inducing tumorigenesis. Doxycycline removal from the system results in the silencing of the expression of the oncogenes. Induction and deinduction of oncogenes affect the morphology of the organoids represented by schematic graphics and visible on histological stainings of the organoids. Representative histological stainings of the organoids used in panel B are adapted from Havas *et al.*[102] (scale bar 50 μM).

1.4 Applying single-cell multi-omic approaches to study tumorigenesis

In the previous chapters of the introduction, I already indicated the importance of applying single-cell technologies to better understand cancer progression, resistance to treatment, and cell type hierarchy of the mammary gland. The last couple of years have witnessed an explosion in the development of single-cell methods, both regarding experimental and statistical analysis. These technologies allow to molecularly characterize single cells on a genomic, transcriptomic, epigenomic, metabolomic, and protein level but none of the techniques is omni-comprehensive. Therefore, with even more sophisticated approaches, a few functional readouts can be collected from the same single cells, and different assays can be analytically linked to create a truly muti-omic picture of a cell reducing the impact of technical and batch effects[108]. However, with collecting more information comes the computational and statistical challenge of how to integrate big data and how to compensate for the fact that different techniques capture only a fraction of molecules that are present in the cell. For example, single-cell transcriptomics is the most frequently used single-cell technology and scRNA-seq protocols seem to be the most optimized and user-friendly, yet high-throughput protocols capture only 5% to 20% of RNA present in the cell (also not accounting for cell-type specific differences in RNA content)[109]. The sparsity of starting material is yet more pronounced when it comes to single-cell DNA sequencing (scDNA-seq) and scATAC-seq as normal cells have only two copies of their nuclear genes and only part of chromatin-accessible sites can be extracted. Until recently, even bigger limitations of the scDNA-seq techniques were that they do not provide any information on cell type, function, or state of analyzed cells[110], and were limited in their resolution (to copy-number profile analyses) and characterization of functional consequences of the genomic rearrangements[27]. In my project I take advantage of the unique features of Strand-seq[111,112] as well as of two computational methods, single-cell tri-channel processing (scTRIP)[113] and single-cell Nucleosome Occupancy and Genetic Variation Analysis (scNOVA)[114], to identify and characterize diverse SV classes in a heterogenous population of single cells preserving the information about their cell types.

1.4.1 Detection of SVs

Compared to well-studied SNVs and indels, SVs, introduced in chapter 1.1.3, are the most common, yet understudied, class of driver mutations in cancer. The identification of SVs, regardless of the technology, is difficult[115,116].

SV discovery methods can be divided into four main groups: cytogenetics, short-read-based methods, long-fragment-based methods, and scDNA-seq[27]. Cytogenetics approaches such as fluorescence *in situ* hybridization (FISH), spectral karyotyping, or microarray-based comparative genomic hybridization are commonly used in the clinics as diagnostic tools for genetic disorders but they are unsuitable to detect complex, overlapping, or nested SVs, labor-intensive and low-throughput[116]. With both cytogenetics approaches and sequencing based-strategies, it is possible to confirm the presence of common germline variants or highly frequent pathogenic SVs[115] but the sequencing approaches provide a better resolution. In theory with high coverage short-read whole genome sequencing data (and sometimes whole exome sequencing data) all SVs may be detected and breakpoints annotated. However, in cases where SVs cover a large part of the read or are even larger than the read, the mapping to the reference genome becomes problematic. Detection of SVs in segmental duplications and repetitive regions of the genome is also complicated, if not impossible with short reads. These problems are circumvented by using third-generation sequencing technologies that feature long reads without amplifying the templates. As the reads can span sequences as long as a megabase (Mb), entire SVs can be captured. The main drawbacks of long-read sequencing are the high error rates, and the need to provide high-quality high molecular weight DNA as a starting material[115,116]. The common disadvantage of all strategies described so far is that they are limited in detecting SVs of subclones with lower frequency. One solution is to perform multiregional sampling and profiling to get a better overview of clonal evolution[108,117] but precise clonal dynamics can be assessed only with single-cell resolution. Because of that, there is an increasing interest in applying single-cell genomic platforms, especially to create phylogenetic relationships between subclones[108]. Considering the limited amount of DNA in single cells, a common approach is to include a whole-genome amplification step during library preparation to increase the amount of DNA available for sequencing[118]. It is not a perfect solution as PCR amplification can be biased and may introduce artifacts that impact variant calling, especially for SVs.

1.4.2 SV discovery and functional annotation with Strand-seq

Strand-seq is a unique technology as it preserves the information about the homologs in the individual cells by sequencing exclusively the template strands (Figure 3A). It allows the discovery of balanced, imbalanced, and complex DNA rearrangements down to 200 kb resolution[111–113]. Strand-seq protocol relies on the incorporation of BrdU, a thymidine analog, into nascent DNA strands while the cell completes exactly one cell division. BrdU-labelled single nuclei are isolated by FACS. The presence of BrdU in a DNA strand changes the binding pattern of Hoechst 33258 leading to a decreased fluorescence signal. Therefore, nuclei with hemi-substituted DNA have half the fluorescence compared to nuclei without BrdU incorporation and after identification, they can be sorted into 96-well plates. BrdU-labelled single nuclei are subjected to library preparation without any pre-amplification steps. First DNA is fragmented with micrococcal nuclease (MNase). Then the nascent DNA strands are nicked at the sites of BrdU incorporation during UV irradiation. In the next step, only the original DNA template strand is amplified during PCR as single-stranded nicks on the nascent strand inhibit proper elongation by the polymerase. PCR also allows for the introduction of unique barcodes for each analyzed cell, so that the barcoded directional libraries can be pooled and sequenced on an Illumina platform[112].

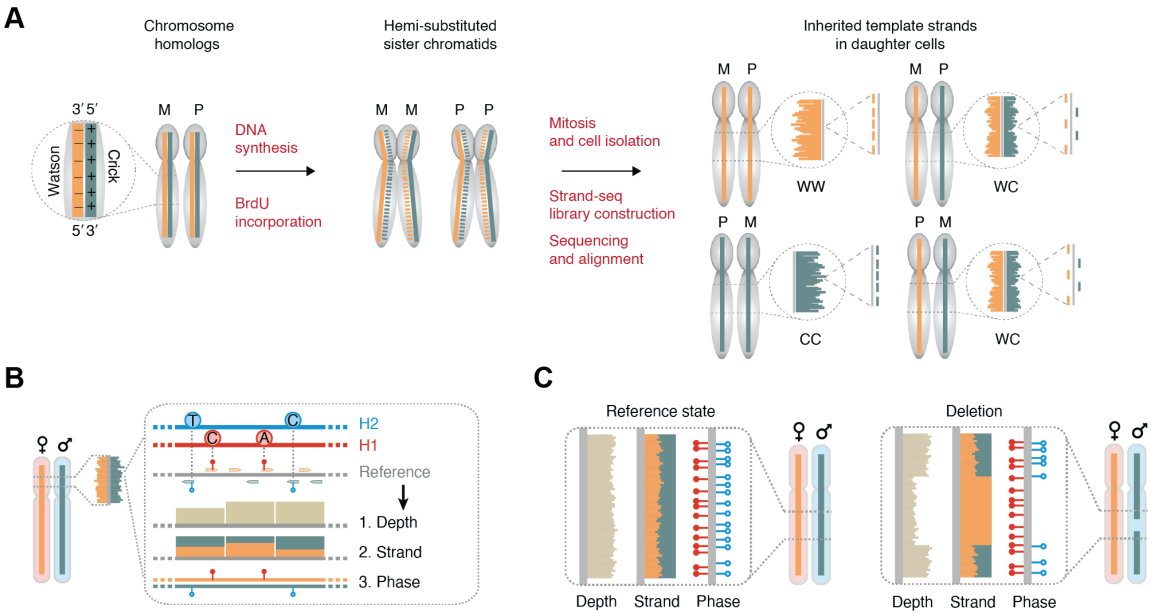


Figure 3 Overview of Strand-seq and scTRIP.

(A) Strand-seq, a single-cell sequencing technique, takes advantage of the directionality of single-stranded DNA molecules, distinguishable as Crick ('C', forward strand, green) and Watson ('W', reverse strand, orange). During DNA replication BrdU is incorporated into nascent DNA strands (dashed lines). Daughter cells inherit CC, WW, or WC template strands of each parental chromosome. Nuclei from hemi-substituted cells are sorted and the isolated DNA is used to prepare Strand-seq libraries. (B) After sequencing and alignment, the generated data can be used to call copy number alterations and balanced DNA rearrangements (including translocations and inversions) by single-cell tri-channel processing (scTRIP), which integrates read depth, template strand, and whole-chromosome haplotype assignments (H: haplotype). (C) Each of these DNA rearrangements can be recognized based on 'diagnostic footprints'. In the presented example, a deletion is detected as a loss in read depth affecting a single haplotype without a change in the read orientation. Panel A adapted from Sanders *et al.*[112], and panels B and C are adapted from Sanders *et al.*[113].

Results of Strand-seq experiments have been used to study chromosome segregation[119], sister chromatid exchange (SCE) events[111,120–122], perform whole-chromosome haplotyping and *de novo* genome assembly[123–126], and characterize SVs in single cells[111,113,114] with a particular focus on inversions[127,128]. SV detection in Strand-seq data is even more precise than manual annotation thanks to scTRIP[113]. scTRIP is a computational framework built for Strand-seq data to facilitate and automate SV discovery. It incorporates three readouts from Strand-seq libraries: depth coverage, read orientation, and haplotype phase (Figure 3B). Several different types of SVs (including deletions, duplications, balanced inversions, inverted duplications, balanced translocations, aneuploidies, chromothripsis, and breakage-fusion-bridge-cycle events) can be identified via a specific 'diagnostic footprint' that combines the information from all three channels. For example, deletion is characterized by a read-depth loss affecting a single strand and haplotype (Figure 3C). The great advantage of scTRIP is that it can identify these SVs in a heterogeneous population at very low clonal frequency levels, including in individual cells. Another computational pipeline very recently developed in our group, scNOVA[114], combines the SV discovery of scTRIP with molecular phenotyping to infer gene expression as a readout. Since MNase is used for Strand-seq library preparation, an enzyme that cuts nucleosome-free regions of DNA, we additionally gain the ability to analyze nucleosome occupancy (NO) profiles providing a complementary epigenetic readout in the same single cells. scNOVA relies on deep convolutional neural networks and negative binomial generalized linear models to infer gene activities from this epigenetic readout. As shown with Strand-seq data from cell lines and patient-derived leukemic samples, it is now possible to infer cell types based on the NO of lineage-specific genes (as long as reference data is available) and predict gene expression differences between defined cell populations (for example, subclones distinguishable on the presence of an SV). In summary, scTRIP, and scNOVA, a truly multi-omic technique, offer a unique chance to explore the consequences of structural variation in heterogeneous cell populations.

Taken together, with the great progress of single-cell technologies and the development of organoid models, it is now possible to study the consequences of multi-layered heterogeneity of breast cancer cells, especially in the context of treatment resistance. What is still missing though, is an in-depth analysis of chemotherapy-related mutational signatures focusing on SVs.

Chapter 2 Objectives and thesis outline

The majority of chemotherapeutics, including doxorubicin, damage DNA and induce cell death due to the stress-related burden. However, some cells survive the treatment and their presence may result in the development of therapy resistance. These residual cells have managed to successfully repair DNA breaks or converted the original damage into mutations so that the DNA integrity is protected. For some chemotherapies, the mutational signatures, including base substitutions or small indels, have been identified but the contribution of SVs has been underexplored. Characterizing the genetic and phenotypic composition of the tumor pre- and post-treatment will provide a better understanding of how different cell types are affected by the same drug, and whether the surviving populations show a cell-type bias and dormancy phenotype that could contribute to the development of minimal residual disease. Such results would be particularly important to identify novel vulnerabilities that could be translated into clinical use for cancer patients (in the context of this thesis, for breast cancer patients).

The overall goal of this dissertation is to identify SVs induced by doxorubicin in different cell types of murine mammary gland organoids from TetO-CMYC/TetO-Neu/MMTV-rtTA strain, and classify cell-type specific transcriptomic changes following doxorubicin treatment in the same breast cancer model. As indicated in the introduction, this study relies on the reductionist model of breast cancer and takes advantage of new technologies that provide a detailed single-cell resolution, necessary to profile heterogeneous samples. I hypothesize that certain SV types will be more common after doxorubicin treatment, depending also on the cell type present in murine mammary gland organoids.

The thesis is structured as follows:

In Chapter 1 I review relevant literature to provide background information for the project.

Chapter 2 and Chapter 3 include the objectives and a statement of contribution.

In Chapter 4 I present the results of a drug screen on murine mammary gland organoids. The aim of the screen was to find drugs and their concentrations that could be used to mimic the treatment received by the patients. Based on the obtained data, I decided to focus on doxorubicin for the follow-up experiments.

In Chapter 5 I outline the attempt to characterize cell-type-specific transcriptomic changes induced by doxorubicin in murine mammary gland organoids. Based on the expression of canonical markers, I identified five different cell types present in the organoids. I also showed that doxorubicin treatment induced G1-cell cycle arrest and was particularly cytotoxic for the population of basal cells. Doxorubicin had a strong negative effect on the fundamental cellular processes.

In Chapter 6 I describe experimental and computational challenges that needed to be overcome to adapt Strand-seq, scTRIP and scNOVA for the purposes of this project. I established Strand-seq protocol in organoids and thus a solid tumor model for the first time. Altogether, these technologies allow identifying SVs in cells of murine mammary gland organoids in a cell-type-specific manner.

In Chapter 7 I summarize the SVs induced by doxorubicin in murine mammary gland organoids. I detected a higher SV burden, as well as an increased frequency of SCEs in all cell types following doxorubicin treatment compared to controls. Complex events occurred exclusively in drug-treated cells.

In Chapter 8 I discuss the findings considering the strengths and weaknesses of the approaches used in this dissertation.

Chapter 9 contains the details of the experimental and computational methods applied in this project, while in Chapter 10 I provided supplementary data.

Chapter 3 Contributions

Unless stated otherwise, I performed all the experiments, analyzed the data, and interpreted the results, with support from my supervisor Prof. Dr Jan Korbel in collaboration with Dr. Martin Jechlinger.

I cultured mammary gland organoids and cell lines for all the experiments, and I isolated the mouse embryonic fibroblasts from embryos. I confirmed the correct genotype of the mice used in this study. I applied molecular biology methods and techniques, such as immunofluorescence (including imaging), RNA and DNA extraction, real-time qPCR, cytotoxicity assays, and flow cytometry. I prepared the samples for Strand-seq and created the libraries for scRNA-seq and scATAC-seq. I analyzed the data from single-cell sequencing technologies and whole-genome sequencing of murine mammary gland organoids. I created all the schemes and figures presented in this thesis apart from figures 3, 12, 14, and 15 (indicated also in the figure legends). I wrote the entire text of this dissertation.

The following people contributed to the work presented in this doctoral thesis:

Dr Hyobin Jeong (Korbel group, EMBL Heidelberg) adapted scTRIP for the murine genome. She performed the computational analysis related to establishing the scNOVA cell-type classifier and inference of gene expression changes.

Marta Garcia Montero and Dr. Martin Jechlinger (Jechlinger group, EMBL Heidelberg) maintained the mouse colony, euthanized the mice used in this study, and provided technical support to isolate mammary glands.

Dr. Sylwia Gawrzak and Marta Garcia Montero (Jechlinger group, EMBL Heidelberg) were involved in the design of the drug screen and provided technical support with organoid seeding, medium change, and endpoint assays related to the screen.

The Strand-seq libraries were generated by Dr. Eva Benito Garagorri, Catherine Stober Brasseur, Patrick Hasenfeld, Dr. Maise Gomes Queiroz, and Benjamin Raeder (Korbel group, EMBL Heidelberg).

Laura Villacorta (Genomics Core Facility, EMBL Heidelberg) provided technical assistance during the generation of scRNA-seq libraries.

I received technical support with FACS from the members of the Flow Cytometry Core Facility at EMBL Heidelberg: Dr. Diana Ordonez, Dr. Beata Ramasz, Dr Daniel Gimenes, Michael Bonadonna, and Dr. Malte Paulsen.

All sequencing was performed at the Genomic Core Facility at EMBL Heidelberg.

Chapter 4 Results: identification of doxorubicin as a drug of interest through a drug screen on murine mammary gland organoids.

The 3D culture system of murine mammary glands from TetO-CMYC/TetO-Neu/MMTV-rtTA strain faithfully recapitulates the dynamics of tumorigenesis of HER2 positive breast disease together with cMYC activation[102,103]. Doxycycline-controlled oncogene induction offers a unique opportunity to study the impact of the drugs on both healthy and cancer cells coming from the same organism. I intended to adapt a drug screening platform using murine mammary gland organoids derived from TetO-CMYC/TetO-Neu/MMTV-rtTA strain, so that my experimental setup would mimic the treatment received by the patients. I decided to focus on the drugs that are clinically relevant for HER2-positive breast cancer patients, such as doxorubicin, lapatinib, and paclitaxel, and for each of them to find the IC50 value, which corresponds to the concentration at which 50% of cells in a population die after being exposed to the substance. Knowing the concentration range in which the drugs were active was the first step to planning single-cell transcriptomic and genomic experiments that aimed at characterizing the impact of these drugs on cancer cells.

4.1 Experimental design to test the cytotoxicity of common cancer drugs

Reported IC50 values for 3D cultures are known to be usually higher than for 2D cultures[129], and there are no literature reports with information about concentrations that could be used specifically for murine mammary gland organoids. Because of that, the choice of the initial concentration range for experiments was based on the data from breast cancer cell lines included in the 'Genomics of Drug Sensitivity in Cancer' project (the Wellcome Sanger Institute). The screen was performed using mini 3D gels (seeding density of 600 cells in 10 μl) growing in 96-well plates. Such an approach reduces variability as all concentrations of a drug (including necessary controls) can be tested on the same plate, and there are enough technical replicates per concentration to perform statistical testing. Seven days after seeding, when the structures reached their final size, doxycycline was added to the medium to induce oncogene overexpression. The organoids, both those never induced and those treated with doxycycline for seven days were then treated with drugs for 72 hours (each drug was tested at five different concentrations with four to five technical replicates). Treatment with dimethyl sulfoxide

(DMSO), a solvent for all the drugs, was used as a control. The growth of the organoids was monitored regularly using a high-throughput brightfield microscope.

Following the incubation with the drugs, I performed commercially available cytotoxicity and cell viability assays (that can be multiplexed) and then calculated the IC50 value for each of the drugs. After optimizing the culture and handling conditions, I successfully performed three (for paclitaxel) and four (for lapatinib and doxorubicin) biological replicates (different mice used for cell isolation and different days of seeding) of the screen with consistent results between the screens. The fluorescence-based cytotoxicity assay detects biomarkers released to the medium during apoptosis, while the viability assay provides a luminescent signal proportional to ATP levels released from lysed metabolically active cells. For all tested drugs, for both never induced organoids and induced with doxycycline, these two inverse measures of cell health resulted in IC50 agreement, and for clarity, only the data from the cell viability assay is shown in this chapter.

4.2 Doxorubicin, but not paclitaxel, has a cytotoxic effect on murine mammary gland organoids

Doxorubicin is a commonly used chemotherapeutics that induces cell death via several mechanisms covered in detail in section 1.2.3 of the introduction. The data from the drug screen indicated that treatment with increasing concentrations of doxorubicin negatively affected the viability of both healthy and cancer cells (Figure 4A). Non-induced cells were more sensitive to drug concentrations higher than 500 nM, reflected by the calculated IC50 value: 800 nM for non-induced cells and 1 μM for cells induced with doxycycline.

Paclitaxel is a cytoskeletal drug that promotes the assembly of microtubules and inhibits tubulin disassembly. The exact mechanism by which paclitaxel induces cell death is still unclear, but one hypothesis is that it affects mitotic spindle formation and therefore affects proper chromosome segregation fidelity during cell division[130]. Although paclitaxel is a widely used microtubule toxin used to treat a number of types of cancer, it did not have a strong cytotoxic effect on either healthy or cancerous cells from murine mammary gland organoids, even if they were exposed to very high concentrations of the drug (Figure 4B). In fact, some of the tested concentrations were so high that paclitaxel precipitated in the medium and formed crystals visible in brightfield (this effect was observed with micromolar concentrations, whereas the typical IC50 value for paclitaxel, reported for human cell lines, is in the nanomolar range) (Sup. Figure 1).

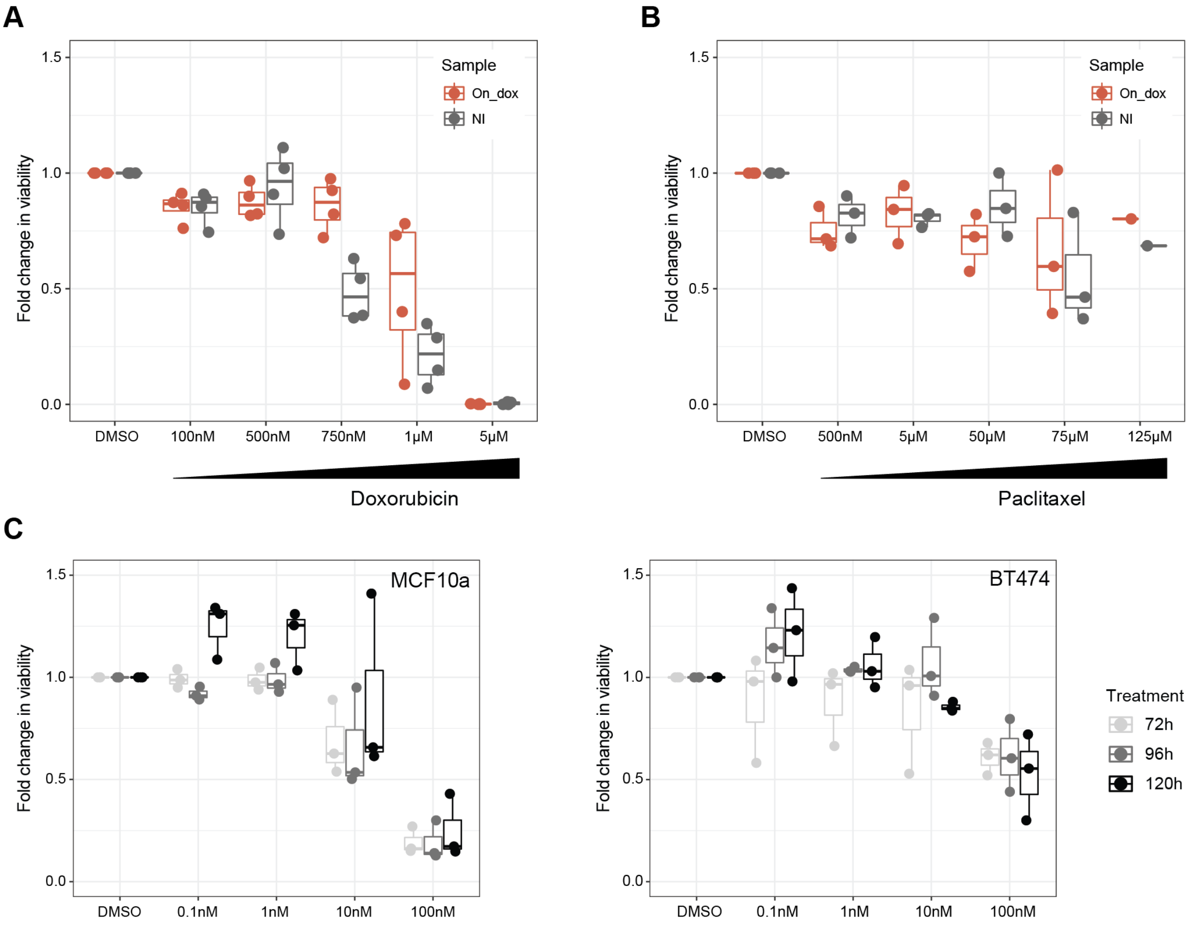


Figure 4 Summary of cytotoxic effect of chemotherapeutics (paclitaxel, doxorubicin) on murine mammary gland organoids and human cell lines.

Murine mammary gland organoids, both never-induced (NI) and induced with doxycycline (On_dox) to overexpress CMYC and HER2 (and mimic cancer phenotype) were treated with increasing concentrations of drugs for 72 hours before measuring the cell viability. Treatment with DMSO (a solvent for both drugs) was used as a control. The results of the luminescence-based assay are presented from 3 or 4 biological replicates. (A) The viability of both NI and On_dox cells decreases with increasing concentration of doxorubicin. (B) Very high concentrations (higher than 100 nM) of paclitaxel reduce slightly the viability of cells forming organoids but the concentration of the drug is much beyond the reported concentration range of paclitaxel cytotoxicity. (C) To exclude the possibility that the drug is not active or the exposure time is too short, I measured the cytotoxic effect of paclitaxel on human mammary gland (MCF10a) and breast cancer (BT474) cell lines. The cells were incubated with increasing concentrations of the drug for up to 120 h. The cytotoxic activity of paclitaxel is already present in the nanomolar concentration range. Prolonged exposure does not increase cytotoxicity. Data from three experimental replicates.

To exclude the possibility that the drug I acquired for the screen was inactive, or that the cells should have been treated for a longer time (as they might require more time to divide), I performed a similar drug test on two cell lines, MCF10a (a spontaneously immortalized human breast epithelial cell line) and BT474 (a human breast tumor cell line overexpressing HER2) which differ in the doubling time (approximately 16 hours for MCF10a and 60-80 hours for BT474, data from the American Type Culture Collection, ATCC). The cells were exposed to paclitaxel for 72, 96, or 120 hours. With both cell lines, I could observe that the higher the concentration of the drug, the lower the viability of the cells (Figure 4C). The cytotoxic effect was already present after 72 hours and the longer exposure did not increase the cytotoxicity. In addition, using the same cell lines, I compared the activity of the drug that I used for organoid screens to one from a different vendor and did not observe any differences. The presented data confirm that exposure to paclitaxel does not affect the viability of murine mammary gland organoids. Nevertheless, it is still possible that paclitaxel induces genomic rearrangements in these organoids but they are not detectable with cell viability or cytotoxicity assays (used as a proxy for DNA damage as they allow high-throughput readout).

4.3 Lapatinib negatively affects the viability of murine mammary gland organoids regardless of their HER2 status

Contrary to paclitaxel and doxorubicin, treatment with lapatinib (used in a form of lapatinib ditosylate) is an example of targeted therapy. Lapatinib is a dual tyrosine kinase inhibitor blocking HER2 and EGFR signalling pathways[131]. In contrast to trastuzumab (Herceptin), an anti-HER2 antibody used in the clinics that is binding only to human HER2[132], lapatinib has been already demonstrated to bind, in addition to the human HER2, to mouse and rat HER2/ERBB2/NEU[133] (and one of the transgenes overexpressed by the doxycycline induced organoids used in this study is rat *Her2/Erbb2/Neu*).

Like during the treatment with doxorubicin, the viability of organoids exposed to lapatinib was affected in a concentration-dependent manner (Figure 5A). Surprisingly though, there was not much difference in the cytotoxic effect of lapatinib between cells that had been induced with doxycycline (so overexpressing HER2 and CMYC) and those that had not (IC50 values 2 µM and 1.95 µM respectively). Potential explanations for that might be that the non-induced cells already had a high level of expression of *Her2*, or that the system is leaking, which would result in the overexpression of oncogenes in the absence of doxycycline. I excluded these possibilities by performing reverse transcription quantitative real-time PCR (RT-qPCR) to check the mRNA levels of the oncogenes, both mouse-specific and from the transgenes (human *CMYC* and rat *Erbb2/Neu*) at three different time points (before induction, after seven days on doxycycline and after seven days on doxycycline followed by seven days off doxycycline). The transgenes were expressed strongly only if the cells were grown in the medium containing doxycycline (Figure 5B).

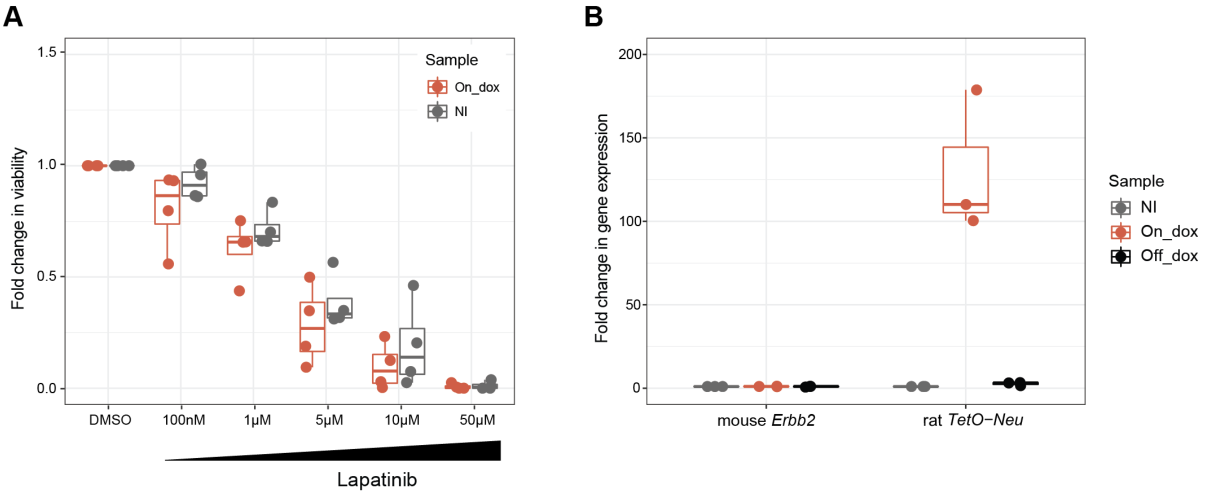


Figure 5 Non-specific cytotoxic effect of lapatinib on murine mammary gland organoids. (A) Murine mammary gland organoids, both never-induced (NI) and induced with doxycycline (On_dox) to overexpress CMYC and HER2 were treated with different concentration of lapatinib for 72 hours before measuring the cell viability. Treatment with DMSO (a solvent for lapatinib) was used as a control. The results of the luminescence-based assay are presented from 4 biological replicates. Even though lapatinib should only target the cells overexpressing HER2, it had a similar cytotoxic effect on organoids without and with induced overexpression of HER2. (B) By performing RT-qPCR on RNA extracted from mammary gland organoids at different stages of growth, I confirmed that the HER2 transgene is overexpressed only if the cells are induced with doxycycline and the baseline expression of murine Erbb2 is low. Off_dox cells were grown for 6 days without doxycycline, then induced with doxycycline for 7 days followed by 7 days off doxycycline. Data from three biological replicates.

According to the supplementary data, provided by Dr. Sylwia Gawrzak, 3D cultures of human breast cancer cell lines overexpressing HER2 (BT474 and SKBR3) are more sensitive to lapatinib than cells without the amplification of the HER2 oncogene (MCF-7, MDA-MB-231) with app. up to a 10-fold difference in the IC50 value (IC50 values were as followed: BT474: 0.94 µM, SKBR3: 2.1 µM, MCF-7: 9.6 µM, MDA-MB-231: 9.1 µM). The reason why the viability of not-induced organoids was negatively affected by lapatinib remains elusive.

All three drugs were interesting potential candidates for follow-up genomic and transcriptomic experiments that would allow me to understand how different drugs affect tumor heterogeneity, and whether they could be associated with certain mutational signatures (especially important in the context of doxorubicin that directly affects DNA and paclitaxel which impacts chromosome segregation potentially contributing to genomic instability). Considering the mechanism of action of anthracyclines and their clinical significance for HER2-positive breast cancer patients, as well as the clear cytotoxic effect on murine mammary gland organoids, I decided to focus my further research on doxorubicin.

Chapter 5 Results: single-cell transcriptomic profiling of murine mammary gland organoids after doxorubicin treatment

In the next step, to get a better assessment of how heterogenous the cell population is after doxorubicin treatment at the transcriptome level, I performed scRNA-seq. With the generated data, I would be able to annotate for the first time the cell types present in murine mammary gland organoids and detect if there are subpopulations or cell types affected differently by doxorubicin. In addition, with these results, I would be able to determine if cells after doxorubicin treatment show early indications of dormancy and MRD phenotype or express candidate genes that could be associated with doxorubicin resistance.

5.1 Experimental workflow and data processing

Based on the results of the screen, for further experiments, I selected the concentration of doxorubicin of 100 nM. Organoids in the cancer state (induced with doxycycline for 6 days) from mammary glands of two independent mice (mouse_498, mouse_000) were treated with doxorubicin or DMSO (solvent for doxorubicin) for 72 hours and allowed to recover for the next three days. Matching control consisted of same-age organoids induced with doxycycline but not treated with DMSO or the drug. For each condition, I sorted cells based on their viability (excluding dead and early apoptotic cells which were present as a consequence of the drug/solvent treatment, or as an effect of the dissociation of organoids from the matrix into single cell solution) and then prepared libraries for scRNA-seq using the 10x Chromium platform. Flow sorting of cells before the 10x Genomics assay is a standard clean-up procedure that allows also to remove the debris that could potentially clog the microfluidic chip. All the libraries were pooled and sequenced together during one run of Illumina NextSeq 2000 to avoid batch effects.

To analyze the sequencing data, I first applied two steps of Cell Ranger (cell ranger mkfastq and cell ranger count), an analysis pipeline provided by 10x Genomics. The Illumina sequencer's base call files were demultiplexed and converted into FASTQ files. Then the sequencing reads were aligned to a mouse reference transcriptome. As next, the generated feature-barcode matrices were processed with the Seurat R package, one of the most common toolkits for quality control and exploration of scRNA-seq data[134]. After filtering high-quality cells, I corrected for technical variability with sctransform, a modeling framework for the normalization implemented in the Seurat workflow[135]. A total of 17,445 cells were used for subsequent analysis (for two biological replicates (mouse_498 and mouse_000 respectively): 2293 or 4005 for untreated control, 3326 or 4770 for DMSO-treated, 255 or 2796 for doxorubicin-treated). Visual inspection of the data following dimensionality reduction and clustering indicated that cells group by cell type or experimental condition, and cells coming from the same experimental condition but a different biological replicate mix well together (Figure 6A).

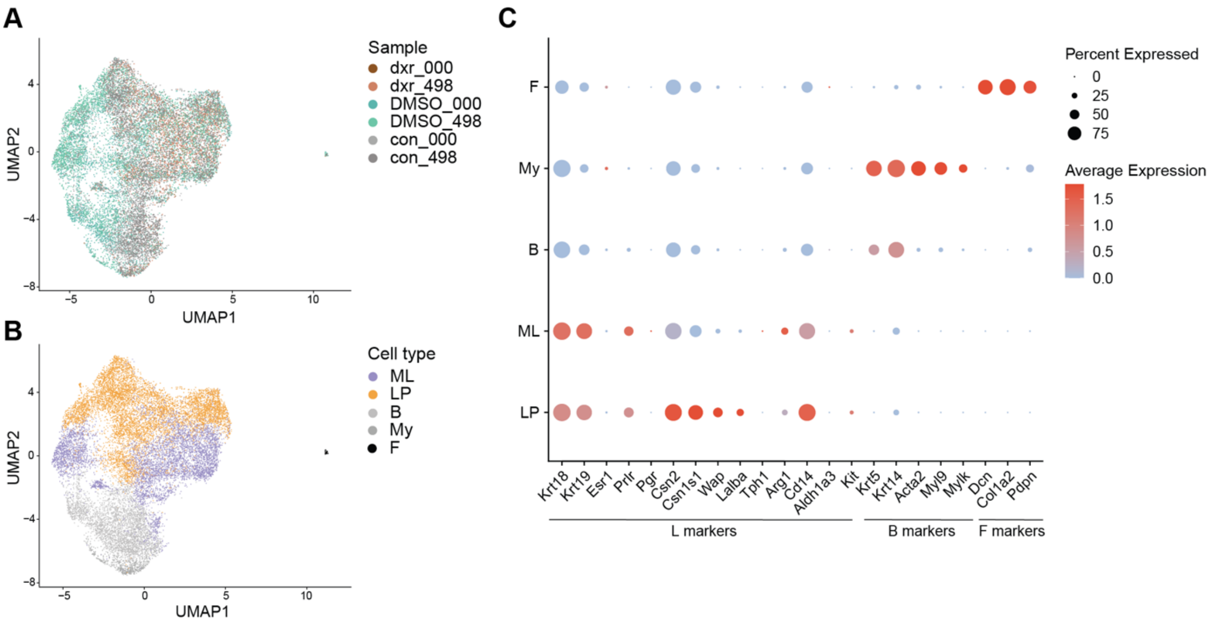


Figure 6 Successful cluster identification and cell-type annotation of cells in murine mammary gland organoids.

(A) Uniform Manifold Approximation and Projection (UMAP) embedding of scRNA-seq of cancer cells forming murine mammary gland organoids, treated with doxorubicin (dxr), DMSO (DMSO), or left untreated (con). Each of the 17,445 cells is represented by a dot and color-coded based on the experimental condition and a biological replicate (organoids were derived from mammary glands of two different mice, 000 and 498 are mice ID numbers). (B) The cell type for each cell was inferred and annotated based on the expression of marker genes (based on a literature review). Five different cell types (labelled with different colors) were present in murine mammary gland organoids: luminal progenitors (LP), mature luminal (ML), basal (B), myoepithelial (My), and fibroblasts (F). (C) Dot plot showing the relative average expression of cell-type specific marker genes between different subpopulations in murine mammary gland organoids. The size of the dot represents the percentage of cells expressing the gene, and the color corresponds to the average expression across the cell type.

5.2 Identification of major mammary cell types from scRNA-seq profiles

To annotate 17 clusters generated through unbiased clustering, I took advantage of several recent papers in which authors performed scRNA-seq of murine mammary glands at different developmental stages and at different phases of the estrous cycle[136–139]. Although these studies provide valuable contributions to understanding the complex composition of the mammary epithelium, their results are not fully coherent, probably due to experimental differences (for example, different markers for cell sorting) and data analysis strategies[2]. It should be noted that the culture conditions and medium composition for murine mammary gland organoids select for epithelium cells. Therefore, in the analyzed dataset, I did not detect any endothelial, neuronal, or immune cells that would be normally present in the mammary gland immediately after extraction. Based on the expression of canonical markers, I identified 11 clusters with luminal cells, five clusters with a basal profile, and one cluster containing fibroblasts (for clarity only annotated clusters are shown, Figure 6B). Particularly the luminal population seemed to display a differentiation continuum rather than clearly separated clusters. The cells from the luminal compartment did not express progesterone or estrogen receptors (*Pgr*, *Esr1*) but they did express prolactin receptor (*Prlr*) (Figure 6C). Within the basal cells, I identified a subgroup that shows signatures of myoepithelial cells like a high expression of *Acta2*, *Myl9*, and *Mylk*. Despite the fact that lactation is normally induced by hormonal changes during pregnancy, luminal cells, in particular luminal progenitors, expressed genes associated with milk production (*Csn2*, *Csn1s1*, *Wap*, *Lalba*). Apart from *Cd14*, the cells included in this analysis did not express other progenitor markers like *Aldh1a3* or *Kit*. Altogether, these results indicate that murine mammary glands organoids are formed by at least five distinct cell types (luminal progenitors (LP), mature luminal (ML), basal (B), myoepithelial, and fibroblasts) and that the majority of cells resemble a more differentiated alveolar state.

5.3 Doxorubicin affects the cell-type composition of organoids and the cell cycle

Once the cell types were annotated, the immediate observation was that after the treatment with doxorubicin, the cell-type composition of murine mammary gland organoids changed. In normal conditions, the majority of epithelial cells in the mammary glands of adult, nulliparous mice are luminal cells, and the frequency of basal cells is between 20-30%[136,138,140]. The distribution of cell types similar to the previously reported ones was observed in the murine mammary gland organoids from untreated or DMSO-treated samples (Figure 7A,

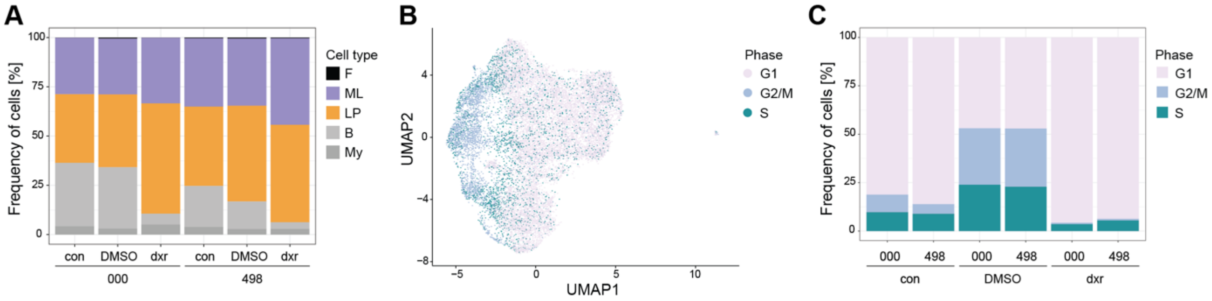


Figure 7 Impact of doxorubicin on the cell-type composition of murine mammary gland organoids and cell cycle phase.

(A) Relative frequency bar chart of cell types present in murine mammary gland organoids treated with doxorubicin (dxr), DMSO (DMSO), or left untreated (con). Data from two biological replicates (000 and 498 are ID numbers of mice from which mammary glands were extracted). The population of basal cells decreases after doxorubicin treatment. (B) UMAP embedding of scRNA-seq data colored by the assigned cell-cycle phase. The phase of the cell cycle is one of the main factors that drive cell separation. (C) Relative frequency bar chart of cell cycle phases associated with different conditions. Both doxorubicin and DMSO affect cell cycle progression compared to the control but only doxorubicin induces G1 arrest.

Sup. Figure 2). After the drug treatment, the population of basal cells decreased as observed in both independent biological replicates indicating that basal cells are more sensitive to the cytotoxic effect of doxorubicin.

Considering that doxorubicin acts predominantly by the induction of DNA damage which may lead to cell cycle arrest or cell death, I checked how the treatment with doxorubicin affected the cell cycle progression. One of the most common practices in scRNA-seq data analysis is to correct for the effect of the cell cycle as in certain cases it may confound a true biological signal[141]. However, in this experiment, the information about the cell cycle state following doxorubicin treatment was particularly crucial, therefore it was not regressed out. Using one of the Seurat functions, I assigned each cell a cell cycle score based on the expression of S and G2/M markers (cells expressing neither of them are classified as G1). Indeed, the cell cycle state was one of the main factors driving the clustering of cells within the same cell type (Figure 7B). After doxorubicin treatment, the vast majority of cells could be associated with the G1 state with only a low fraction of cells entering the S or G2/M phases (Figure 7C). In all three most abundant cell types (B, LP, ML) the effect of doxorubicin on cell-cycle was the same (Sup. Figure 3). Surprisingly, treatment with DMSO also affected cell cycle progression with a higher fraction of cells in S and G2/M phases compared to untreated organoids. Only cells from untreated and DMSO-treated samples contributed to a subgroup of luminal cells considered highly proliferative based on high expression of proliferation-promoting genes (such as *Mki67*, *Birc5,* and *Tyms*). Overall, these data indicate that doxorubicin treatment induces a G1 arrest, and additional recovery time after doxorubicin treatment is required for cells to be able to divide again.

5.4 Comparison of the shared and cell-type-specific transcriptomic responses to doxorubicin

To summarize global changes induced by doxorubicin, I performed differential gene expression analysis (using log-normalized data from both biological replicates and DMSO-treated samples as controls) and gene ontology (GO) and gene set enrichment analysis (GSEA) (with Cluster profiler 4.0[142,143]). Compared to the DMSO-treated control, doxorubicin-treated cells downregulate the expression of 803 genes and upregulate the expression of 427 (avg_log2FC smaller or bigger than 0.2, adjusted *p*-value smaller than 0.05) (Figure 8A). Products of upregulated genes were involved in biological processes connected to cell motility and migration, as well as stress and inflammatory response, while products of downregulated genes were participating in the most crucial cellular functions such as protein synthesis and translation, and cellular respiration (Figure 8B). On a global level, cells after doxorubicin treatment downregulated the expression of *Top2a*, one of the direct targets of doxorubicin, and *Topbp1*, DNA topoisomerase 2-binding protein 1 (TOPBP1), a binding partner of TOP2A, also involved in DNA repair[144]. Downregulation of *Top2a* is observed in doxorubicin-resistant cells[145,146]. Apart from *Top2a*, doxorubicin-exposed cells did not significantly deregulate the expression of some of the most common genes whose products may promote resistance, such as multidrug resistance transporters. Based on the GSEA results, doxorubicin-treated cells did not show the phenotype of senescent cells[147]. Taken together, the treatment with doxorubicin has a strong effect on the transcriptome of cancer cells forming murine mammary gland organoids, and the consequences are still visible after 3 days since the drug was removed from the medium.

In the next step, I wanted to get more insight into whether different cell types of mammary gland organoids have unique responses to doxorubicin. In all samples the fraction of luminal cells, both progenitors and mature, is the highest. Compared to untreated and DMSO-treated controls, only the population of basal cells decreases strongly after doxorubicin treatment indicating that basal cells were more sensitive to the cytotoxic effect of doxorubicin. To look for cell-type specific differentially expressed genes, I performed the analysis with EdgeR[148–150] (based on the design matrix and voom method[151] available in the R package

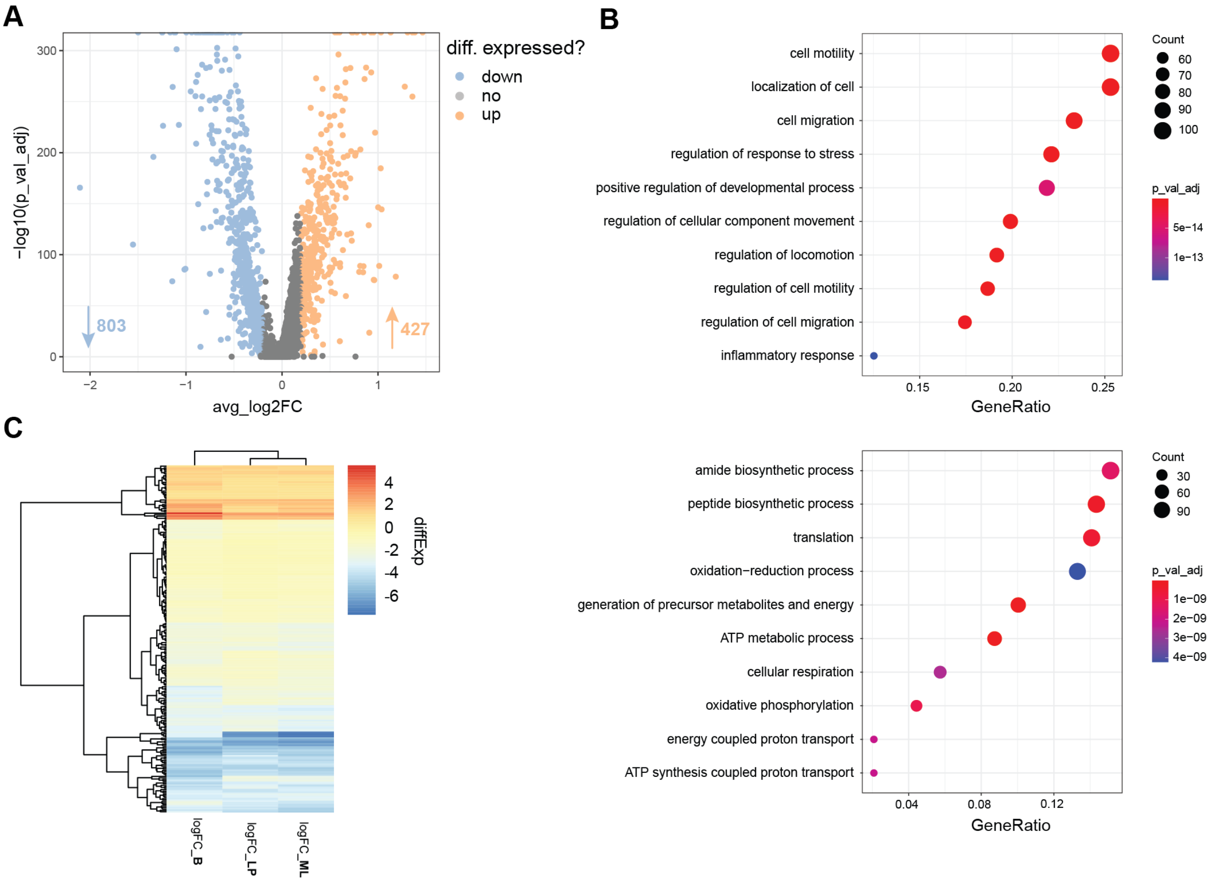


Figure 8 Identification of key genes and biological processes following doxorubicin treatment.

(A) Volcano plot showing differentially expressed genes of doxorubicin-treated cells compared to DMSO-control. Significantly up-regulated and down-regulated genes are shown as orange and blue dots, respectively. The differential expression test was performed based on the Wilcoxon rank sum test, adjusted *p*-value based on Bonferroni correction. (B) Bubble plot showing GO enrichment of differentially expressed genes between doxorubicin vs DMSO-treated cells. The top 10 GO terms of biological processes significantly enriched by up-regulated (top) and down-regulated (bottom) genes (Bonferroni-adjusted *p*-values). (C) A heatmap comparing the expression of 236 deregulated genes shared by basal and luminal cells after differential gene expression analysis using cell type as a confounding factor. For each cell type, the difference in gene expression was calculated between the doxorubicin-treated sample and solvent control.

limma[152], adjusted *p*-value based on Bonferroni correction). Both luminal cell types show a greatly increased number of statistically significant (adjusted *p*-value smaller than 0.05) deregulated genes compared to basal and myoepithelial cells (number of deregulated genes: 956 for luminal progenitor, 1,215 for mature luminal, 470 for basal and 55 for myoepithelial). Two luminal subtypes shared 785 deregulated genes of which 236 were also deregulated in the same direction in basal cells (Figure 8C). Genes whose products control cell cycle progression (like *Mki67*), as well as *Top2a and Topbp1*, were downregulated in all three cell types. Interestingly, luminal cells but not basal cells, strongly downregulate *Cbr3* coding for carbonyl reductase that catalyzes the reduction of doxorubicin to toxic alcohol metabolites. GO and GSEA on luminal-specific deregulated genes were inconclusive and did not provide an explanation why these cells may have a different response to doxorubicin or a survival advantage compared to basal cells.

The analysis described in this chapter revealed the cell-type composition of murine mammary gland organoids and confirmed that all major cell types from epithelial lineage normally present in murine mammary glands are preserved. Treatment with doxorubicin induces strong transcriptomic changes that affect the expression of genes essential in regulating cell cycle progression and metabolism. Although the basal cells seem to be more sensitive to the drug, the findings suggest that the overall transcriptional changes are rather shared between all cell types.

Chapter 6 Results: establishing a single-cell multi-omics approach to study cell-type specific SVs

Even though it is well-established that doxorubicin is a DNA-damaging agent, surprisingly little is known about whether it can be associated with mutational signatures and the formation of SVs. To get better insight into cell-type specific mutation patterns observed after doxorubicin treatment, I took advantage of a single-cell genomics technique Strand-seq, and computational methods built around it, scTRIP and scNOVA. In this chapter, I describe how I adapted both experimental and computational workflows for the needs of this project. In Chapter 7 using scTRIP and scNOVA modified for mice genome, I characterize the types and frequencies of SVs detected in different cell types present in murine mammary gland organoids, at different stages of tumorigenesis (from normal cells to cancer cells before and after doxorubicin treatment).

6.1 Successful generation of Strand-seq libraries from cells of murine mammary gland organoids

The critical step in the Strand-seq protocol is BrdU incorporation into nascent DNA strands followed by the isolation of single nuclei. Although Strand-seq libraries have been previously generated from human[112–114], primate[127,153], and mouse[111] cell lines, human primary cells[113,114], and yeast[120], there are no published reports of applying Strand-seq libraries on cells forming organoids.

I first optimized the protocol to isolate the nuclei from 3D structures (Figure 9A) and for clarity, I present the most optimal method relying on enzymatic digestion. Such an approach was more efficient in my hands compared to non-enzymatic methods (including dissolving with ice-cold phosphate-buffered saline (PBS) or a commercially available non-enzymatic dissociation reagent), and its first steps were also applied to isolate single cells before scRNA-seq (described in the previous chapter). Briefly, Matrigel, the matrix in which organoids grow, was enzymatically digested and the loose gel was mechanically disrupted by pipetting up and down. A high concentration of trypsin was used to isolate single cells from the organoids, and the single nuclei were extracted with previously reported ‘nuclei staining buffer A’ (a high salt buffer with a detergent)[113]. As high concentrations of BrdU might be toxic to primary cells (and supposedly to organoids derived from them), I then tested if it was also the case for murine mammary gland organoids. These organoids are most sensitive to perturbed growth conditions at the very beginning of culture, therefore I treated them for 72 hours with different concentrations of BrdU already two days after seeding. BrdU concentration of 40 µM is typically used in Strand-seq experiments on human cell lines and was also used in one study including a murine cell line. Although even the lowest tested BrdU concentration (5 µM) negatively affected cell viability, the effect was much stronger with concentrations higher than 20 µM (Figure 9B). Based on the results, a BrdU concentration of 20 µM was selected for all Strand-seq experiments performed on murine mammary gland organoids.

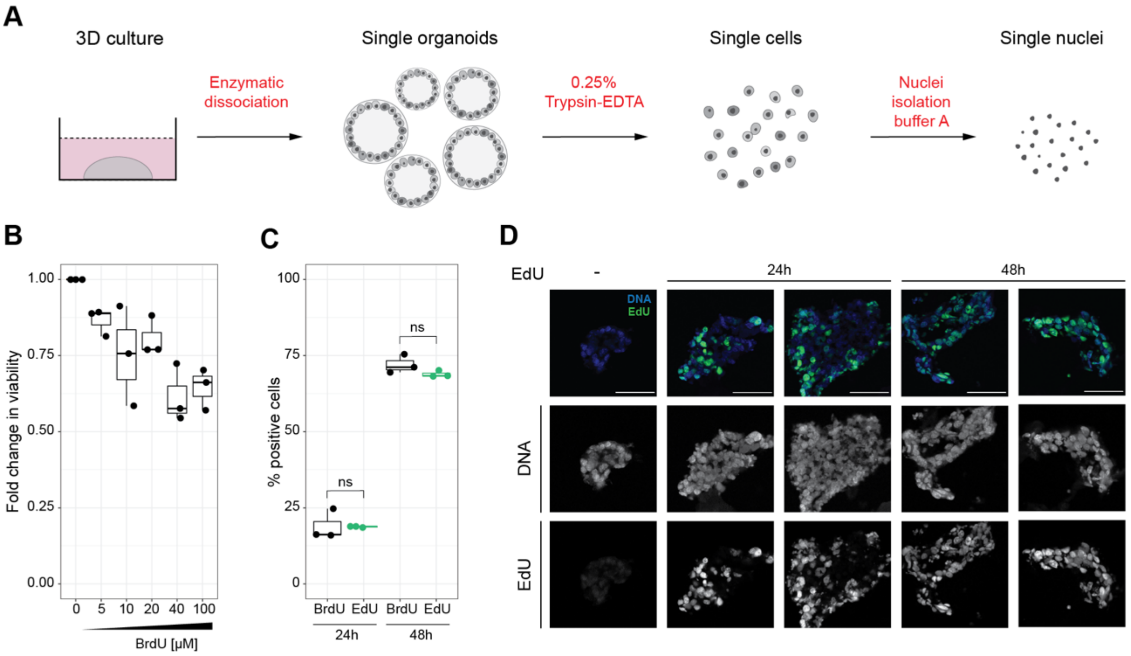


Figure 9 Establishing an experimental workflow for detection and isolation of BrdU-positive nuclei from murine mammary gland organoids.

(A) Organoids were released from Matrigel through enzymatic dissociation and mechanical disruption of the matrix. Single cells were isolated following trypsinization, and single nuclei were extracted with an established gentle cell lysis buffer. (B) Never-induced murine mammary gland organoids were cultured in the presence of different concentrations of BrdU for 72 hours before measuring cell viability with a luminescence-based assay. The plot shows the results of three independent experiments. (C) Murine mammary gland organoids induced with doxycycline were incubated with either 20 µM BrdU or 20 µM EdU for 24 and 48 hours. The frequency of BrdU-positive and EdU-positive cells was analyzed with flow cytometry based on Hoechst quenching or Click-iT staining. BrdU and EdU are incorporated into the DNA of cells forming murine mammary gland organoids with the same efficiency (not significant difference, Student's *t*-test). The plot shows the results of three independent experiments (with at least 10,000 single nuclei or cells recorded). (D) Doxycycline-induced murine mammary gland organoids were cultured in the absence or presence of 20 µM EdU for 24 or 48 hours. The gels containing the organoids were then cut into 8 µm cryosections. The cryosections were fixed and labeled for EdU with Click-iT chemistry (green) and DNA was counterstained with 4',6-diamidino-2-phenylindole (DAPI) (blue). Example images of single organoids are presented showing that EdU is incorporated into cells growing both inside and on the rim of structures. Scale bar 50 µm.

To confirm that cells inside the organoid are incorporating BrdU as well as the cells in the outer rim of the structure, I intended to analyze them by immunofluorescence with an anti-BrdU antibody. However, assays based on anti-BrdU antibodies require that DNA is first denatured with acid or heat to expose BrdU (Sup. Figure 4). In my hands, despite testing different protocols on both whole gels or cryosections, such harsh processing destroyed the structural complexity of the samples. Because of that, I took advantage of Click-iT EdU labeling. EdU is a nucleoside analog to thymidine, and like BrdU, it is incorporated into DNA during active DNA synthesis. EdU can be detected based on a click reaction between one of its alkyne moieties and the azide coupled to an Alexa Flour® dye. Contrary to anti-BrdU antibody staining, the click reaction is performed under mild conditions. To verify that BrdU and EdU are incorporated with the same frequency, I first incubated the doxycycline-induced organoids with either 20 µM BrdU or 20 µM EdU for 24 hours or 48 hours. BrdU- and EdU-negative organoids were used as control. BrdU-positive nuclei or EdU-positive cells were detected with flow cytometry following respective stainings. BrdU incorporation can be identified by quenching of Hoechst fluorescence, and the presence of EdU in DNA is proportional to the fluorescent signal of a dye after the click reaction. Based on flow cytometry data, there was no statistically significant difference between the frequency of BrdU and EdU incorporation into murine mammary gland organoids at two different time points which suggests that EdU labeling would resemble well anti-BrdU staining (Figure 9C). Organoids that were cultured longer in the presence of BrdU and EdU have a higher number of positive cells. By performing Click-iT EdU reaction on 8 µm cryosections of gels containing doxycycline-induced organoids and visualizing them with the confocal microscope, I confirmed that EdU was incorporated by cells growing both on the outer rim of the structures as well as cells inside the organoids (Figure 9D).

Encouraged by the results, I proceeded to generate the first Strand-seq libraries in an organoid system, by using the murine mammary gland organoids at different time points during the culture: never induced with doxycycline (representing healthy tissue) and induced with doxycycline for 5 and 12 days (corresponding to cancer phenotype). The optimal duration of BrdU exposure was tested for each sample and ranged between 20 hours to 48 hours. Single nuclei were isolated as described above and profiled with flow cytometry (Figure 10A) to determine what proportion of cells had divided (Figure 10B). Single, hemi-substituted nuclei were then sorted into 96-well plates and processed on a Biomek FXP liquid-handling robotic system to prepare libraries from each well. Pooled single libraries were sequenced by paired-end sequencing on the NextSeq Mid sequencer.

Taken together, these data indicate that I succeeded in optimizing a wet-lab workflow for producing Strand-seq libraries from murine mammary gland organoids.

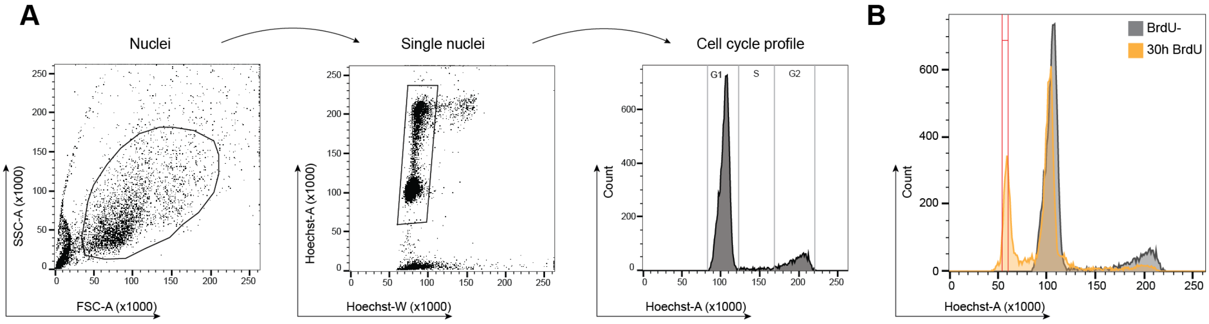


Figure 10 BrdU-positive nuclei are detected using flow cytometry based on Hoechst quenching. (A) Gating strategy: nuclei are selected from debris based on the forward (FSC) and side (SCC) scatter. Doublets are excluded by size. The cell-cycle kinetic and BrdU incorporation status is then analyzed with Hoechst staining. (B) Nuclei from murine mammary gland organoids incubated with (orange) and without (grey) BrdU for 30 hours were isolated and stained with Hoechst before analyzing on a BD FACS Melody. BrdU-positive and BrdU-negative populations are distinguished due to the difference in the fluorescent quenching in the Hoechst channel. The sorting gate (in red) is set on the peak that has half of the fluorescence compared to the G1 peak of the BrdU-negative sample.

6.2 Updating the scTRIP pipeline for the mouse genome

As the next step, I wanted to combine the experimental setup with scTRIP computational analysis[113], so that we would have a tool that will allow us to detect SVs in individual cells of murine mammary gland organoids. First, I wanted to test the pipeline on samples without drug treatment to check if it can be applied to correctly detect germline SVs and to get the initial estimate about the frequency of somatic SVs without any perturbations.

scTRIP has been previously successfully applied to the human genome to analyze the frequency of different SV classes in transformed epithelial cell lines and patient-derived bone marrow and leukemic samples[113]. We reasoned that, after certain modifications, scTRIP could be also used to discover somatic SVs in mice genomes. Therefore, we adapted the pipeline, developed by a past lab member Dr. Sascha Meiers in collaboration with the group of Prof. Dr. Tobias Marschall (Universitätsklinikum Düsseldorf), for the project. After sequencing, the raw data (as fastq files) from every single library were aligned to the mouse reference genome (mm10), marked for duplicates, sorted into bam files, and indexed. In the next step, strand-specific reads (Watson or Crick) were counted into bins of various sizes (20 kb, 50 kb, 100 kb, 200 kb, 500 kb) and plotted as histograms on a cell-by-cell basis (Figure 11). Each overview plot (representing a single cell) was manually curated so that further analysis would be performed without low-quality cells (e.g. cells with incomplete BrdU incorporation, cells that divided twice in the presence of BrdU, cells with too few sequencing reads, or far too many reads compared to the other single cells) (Sup. Figure 5). 47 never-induced cells, and 49 and 47 doxycycline-induced (for 5 and 12 days, respectively) cells state passed the quality control step and were included in the downstream analysis.

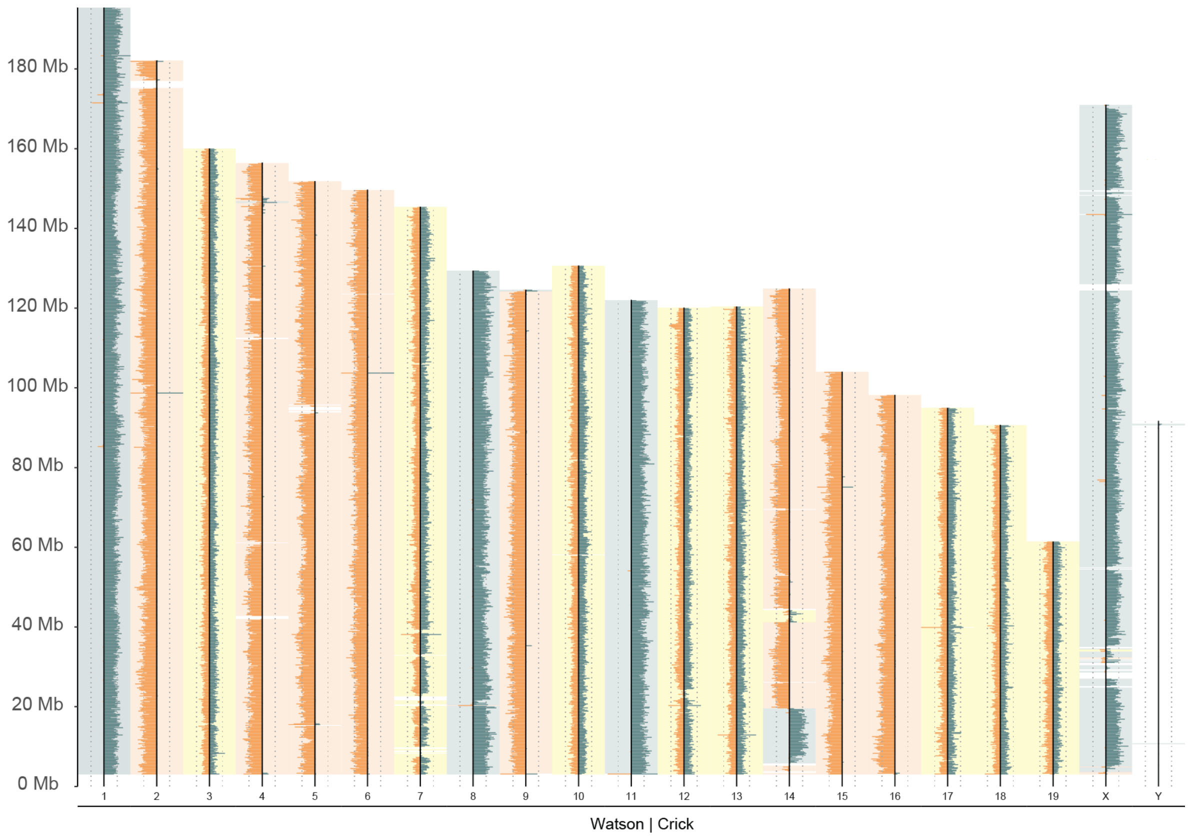


Figure 11 Example of high-quality Strand-seq library from a cell of never-induced murine mammary gland organoid.

As Strand-seq libraries show variability in coverage and background, a quality control (QC) step is performed at the beginning of the analysis to exclude low-quality cells. An ideogram is generated for each library (using the Strand-seq plotting pipeline). After the alignment to the murine reference genome (mm10), chromosomes are represented with Watson reads in orange and Crick reads in blue. For any given chromosome, the cell can inherit the maternal and paternal template strands as either WW and CC, or WC and CW. A high-quality library shows consistent coverage on all chromosomes and directionality, with at least 200,000 reads (ideally more than 300,000). The QC is done manually on each library in an experiment, and the maximum acceptable values for background levels are dependent on the analyzer (but remain constant between experiments).

The scTRIP pipeline is composed of segmentation, haplotype phasing, and Bayesian calculation steps to accurately identify SVs. We first ran these steps of the scTRIP pipeline without blacklisting any genomic regions, without normalization, and a ploidy estimate to verify which parts of the workflow would have to be corrected. The obtained results were clearly incorrect as the genome was not segmented properly, and the SVs were called even though there were no indications or diagnostic footprints of them (Sup. Figure 6). These issues were solved by including blacklisting (the list of regions of low mappability in the mm10 reference genome was created by Dr. Hyobin Jeong) and correcting a segmentation bug that was additionally identified. Once the pipeline was optimized, I was able to annotate all germline and somatic SVs present in cells forming mammary gland organoids. These results are covered in detail in Chapter 7

6.3 scNOVA as a tool for cell-type prediction and inference of gene activity changes

scNOVA, a recent development in the Korbel lab, allows us to integrate the discovery of somatic SVs and NO measurements in the same cell[114]. As scNOVA has been applied so far only to the human genome, we decided to expand its utility also to the mice genome and mice tissues, with a particular focus on murine mammary gland organoids. In this project, we applied one of scNOVA functionalities to annotate a cell type (basal, luminal progenitor, or mature luminal) to each of Strand-seq libraries derived from murine mammary gland organoids, and then characterized drug-induced SVs in those organoids in a cell-type specific manner.

6.3.1 Creating and validating a cell type classifier for mammary cells

While applying scNOVA, we take advantage of one of the Strand-seq library preparation steps during which DNA is digested with MNase. MNase digests protein-unbound DNA, so that DNA wrapped around histones remains intact and contributes to sequence read counts[154]. scNOVA has two main functions: supervised cell-type classification and inference of altered gene activities between cell populations or conditions. The first functionality, which will be covered in detail now, is based on the fact that transcribed genes exhibit reduced NO in their transcription start sites and gene bodies[114]. Chromatin accessibility patterns are tightly regulated by a network of transcription factors and the motifs, and their combinatorial activity drives cell-type-specific gene expression programs. Using one type of single-cell epigenomic data as a reference, we can build a supervised classifier that classifies single cells into one of the cell-type categories based on the activity of the transcription factor motifs as a feature set. And then we can use that information to classify the cell type of each single-cell library newly generated with the same or different kind of single-cell epigenomics method.

To create a cell-type classifier, it is required to train the model using a reference single-cell epigenome dataset in which a correct cell type was labelled for each single-cell library. In the currently ongoing study in the Korbel group, in which the scNOVA pipeline was applied to the human hematopoietic system, the authors created a single-cell micrococcal nuclease sequencing (scMNase-seq) reference for human bone marrow and umbilical cord blood hematopoietic stem and progenitor cells. Such an approach is appropriate as NO profiles generated during the Strand-seq protocol are highly similar to the ones obtained after scMNase-seq[114]. Performing scMNase-seq relies on index sorting: cells are first stained based on the expression of cell-surface markers and then sorted with flow cytometry into a 96 or 384-well cell culture plate in a way that it is possible to identify which cell was sorted to which well. Index sorting is a common practice to analyze, for example, hematopoietic cells with well-established cell-surface markers, but it is difficult to do on cells without specific cell-surface markers like mammary cells. For this project instead, I decided to apply scATAC-seq data from mammary glands for model training as this technique measures chromatin accessibility in the motifs, does not rely on index sorting, and is broadly used in the field allowing us to take advantage of public datasets. In addition, fewer cells are required as starting material for scATAC-seq compared to scMNase-seq, and commercial protocols for scATAC-seq are more and more available. Opposite to scMNase-seq, scATAC-seq data contains information about nucleosome-depleted chromatin in a cell rather than NO[155]. Therefore, during the cell type annotation, we need to invert the Strand-seq-derived-NO Z-score to obtain chromatin accessibility (more details described below). We used published data to create a cell-type classifier, and as an additional sanity check, I performed scATAC-seq on murine mammary gland organoids to confirm the correlation between the reference and our data.

A study performed by Chung *et al.* contains single-nucleus (sn)ATAC-seq data of fetal and adult mammary cells[140]. The authors profiled 7,846 high-quality single nuclei derived from 2,577 fetal and 5,269 adult cells, and only the data from the adult cells were used for the model training. During the sample preparation, they removed non-epithelial stromal and blood cells, so that snATAC-seq was performed only on epithelial cells. The results reveal chromatin changes that correlate with basal and luminal (progenitor and mature) cell states, which perfectly corresponds to cell types that we expect in murine mammary gland organoids (that are enriched for epithelial cells during culture, check Methods). For our aim, the snATAC-seq

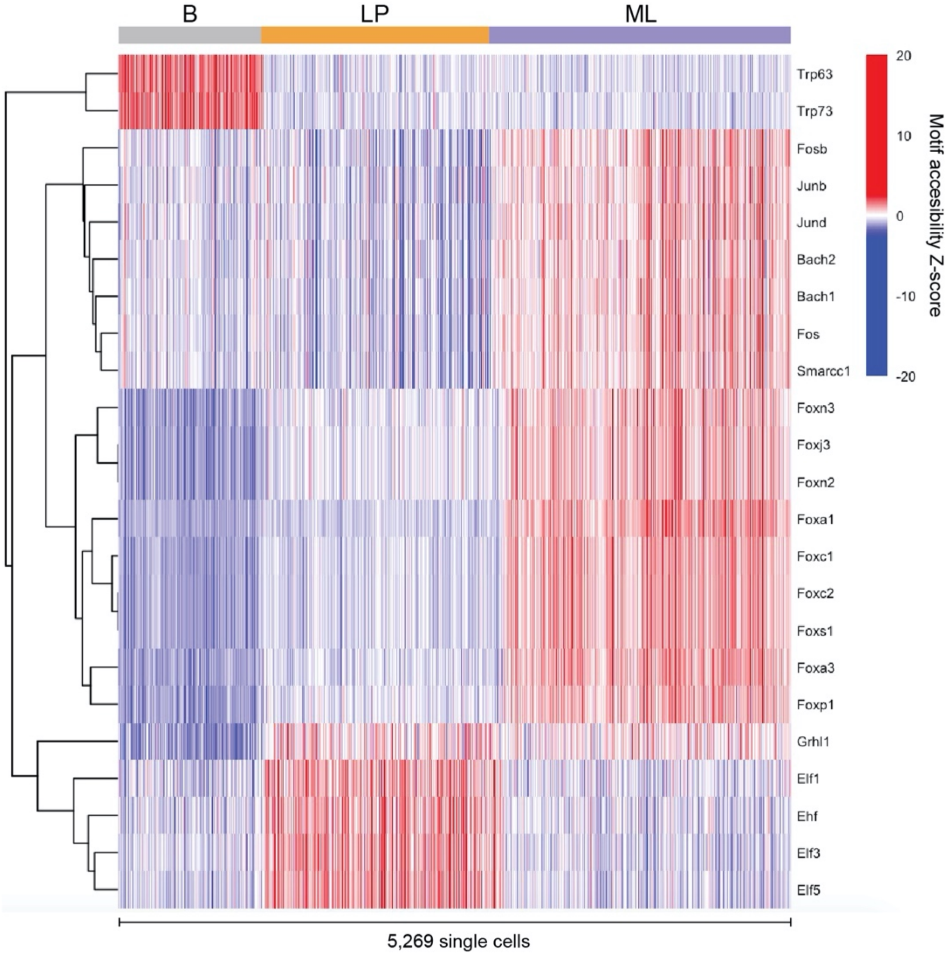


Figure 12 Establishing a mammary gland cell-type classifier for scNOVA from snATAC-seq.
To train the cell-type classifier of murine mammary gland organoids, we used publicly available snATAC-seq reference data of murine mammary glands. The motif accessibility on 23 motifs differs between three main cell types (B, LP, ML) allowing to distinguish these cell types based on NO. Transcription factors that are well-known to regulate different mammary cell states are enriched in their corresponding groups. Heatmap prepared by Dr. Hyobin Jeong.

count matrix (peak by cells) from Chung *et al*. was first converted into a motif accessibility matrix (motifs by cells) using the chromVAR package[156]. This motif accessibility was then used as a feature to build a classifier using Partial least squares discrimination analysis (PLS-DA). For feature selection, we calculated Variable Importance in Projection (VIP) values which measure discriminant power for each motif. We took motifs with significant VIP values compared to null distribution (FDR 10%) to finalize the model and evaluated the performance using leave-one-out cross-validation. Our scATAC-seq-based classifier relies on 23 (short version based on the stringent feature selection criteria using FDR of VIP <10%) (Figure 12) or 50 (extended version based on the lenient criteria using VIP values >90% of the null distribution) motifs. For further applications, only the classifier based on 23 motifs was used. To predict cell types of single cells profiled by Strand-seq, their NO count matrix (peak by cells) is converted into motif occupancy matrix (motifs by cells) using chromVAR[156]. Then we inverted this motif occupancy matrix into an accessibility matrix (motifs by cells) using a formula: motif accessibility=(-1)*motif NO Z-score. The resulting matrix is then used as an input to the classifier which provides information about the most likely cell type of each single-cell library.

The reference dataset used to create a cell-type classifier contains mammary tissue freshly isolated from murine late-stage embryos and adults. To confirm that the cell-type-specific patterns of chromatin accessibility are not affected by culture conditions and that they are shared by cells forming mammary gland organoids, I performed scATAC-seq (using the 10x Genomics Chromium Next GEM Single Cell ATAC v1.1 protocol) on two independent biological replicates of murine mammary gland organoids that have been in culture for 7 days. After scATAC-seq library preparation, sequencing, and filtering out low-quality nuclei, I obtained 1,443 cells (882 from biological replicate 1 (mouse_788) and 561 from biological replicate 2 (mouse_839)) for further analysis with Signac[157]. The median reads per nucleus were similar between the two replicates (33,451 and 45,333), like the median reads in peaks (69% and 66%). In both samples app. 86% of peaks were derived from promoter-distal regions. I then created a common set of peaks for these two replicates and merged the samples. After data processing and dimensionality reduction, I performed unbiased clustering on all peaks (Figure 13A). Interpretation of clusters in scATAC-seq data is particularly challenging due to its sparsity, and limited information about the functional role of chromatin accessibility compared to the actual transcription of genes. To overcome these issues and to annotate the groups revealed by the UMAP visualization, I first generated a gene activity matrix for each cell (by summing the fragments intersecting the gene body and promoter region), and then classified them by label transfer based on scRNA-seq data from the same biological system (from the experiment on murine mammary gland organoids described in Chapter 5 With such an approach, I aimed to identify shared correlation patterns between the gene activity matrix and different cell types present in the scRNA-seq dataset, resulting in a classification score for each cell from scATAC-seq data. This allowed me to distinguish clusters of epithelial cells: luminal progenitor, mature luminal, basal, and myoepithelial cells, as well as a minor population of stromal cells (fibroblasts) (Figure 13B). The LP, ML, and basal cells represent 69%, 20%, and 8% of the total population included in the analysis, respectively. The cellular composition resembles more of a fetal-like population described in Chung *et al.*, probably because the mammary gland organoids used for scATAC-seq were in culture only for 7 days.

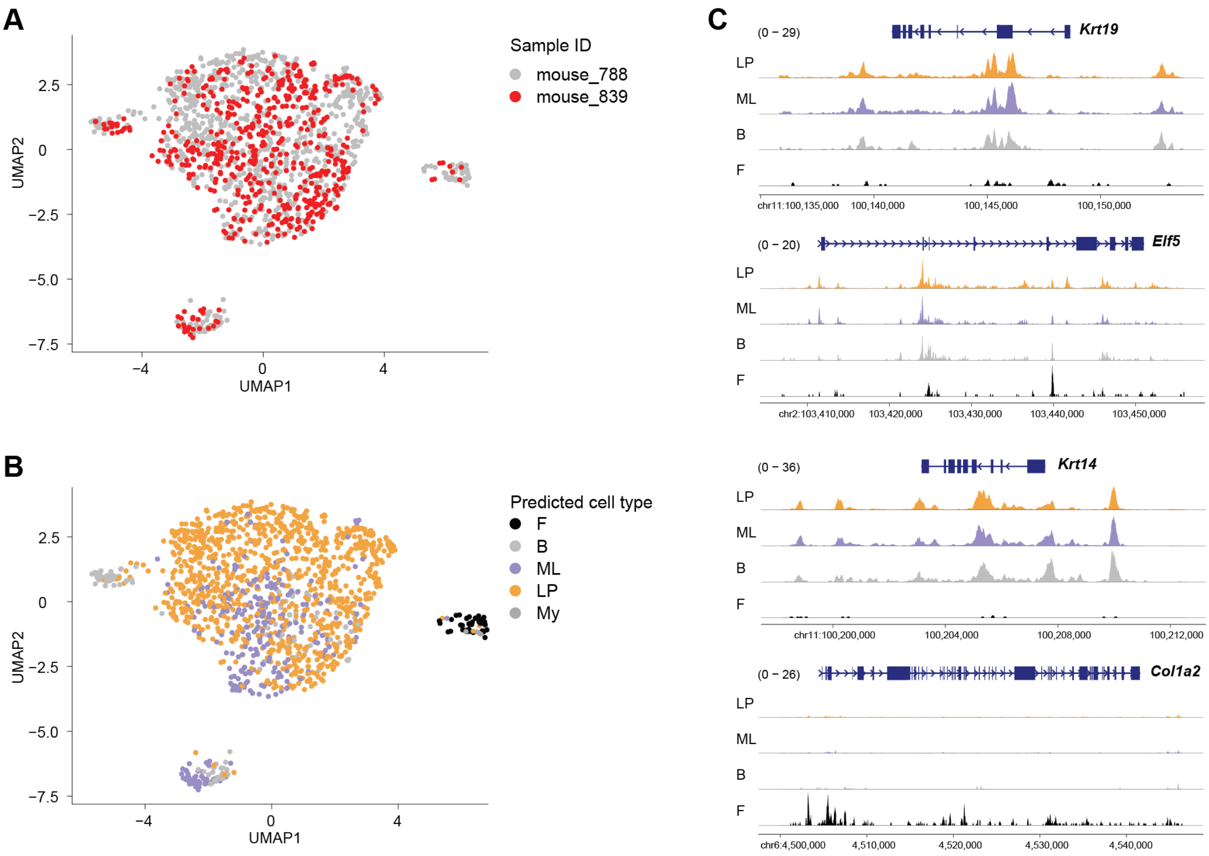


Figure 13 Up to 5 different cell types can be detected in scATAC-seq from murine mammary gland organoids.

scATAC-seq was performed on mammary gland organoids derived from two different mice (mouse_788 and mouse_839) using a commercially available kit. (A) UMAP representation of the scATAC-seq results after merging the data from two biological replicates by creating a common peak set. (B) scATAC-seq cells were annotated via label transfer from scRNA-seq experiments. Major epithelial cell types present in mammary glands were identified (LP, ML, B, My), as well as a small population of fibroblasts. (C) Aggregate scATAC-seq profiles of different cell types present in mammary gland organoids. Tracks are normalized to correct for the number of cells and potential differences in sequencing depth. Signal ranges are shown in parentheses.

As an additional control, I confirmed that LP, ML, and basal cells differ in accessibility and expression of canonical markers (as visualized using pseudobulk chromatin accessibility profiles and pseudoexpression data) (Figure 13C). Luminal cells are more accessible at pan-luminal markers such as Krt19, with LP cells showing higher accessibility at progenitor marker Elf5. Basal cells are characterized by increased accessibility at basal marker Krt14, and fibroblasts- at fibroblast-specific Col1a2. Altogether, these initial results showed that the scATAC-seq data from murine mammary gland organoids is of high quality and that based on chromatin accessibility we can detect all cell types that have been reported in the reference dataset from murine mammary glands used for the scNOVA-based cell-type classifier.

To further investigate the similarity between scATAC-seq from murine mammary gland organoids and the reference dataset, we focused on motif-based analysis. By using chromVAR we annotated scATAC-seq peaks depending on the presence of transcription factor (TF) motifs, including 23 motifs used in the cell-type classifier (Figure 14A). We then compared the cell-type specific motif accessibility between the reference and scATAC-seq datasets on these 23 motifs. There was a positive correlation between corresponding cell types in these two samples (Figure 14B) confirming the utility of the published dataset to create our cell-type classifier.

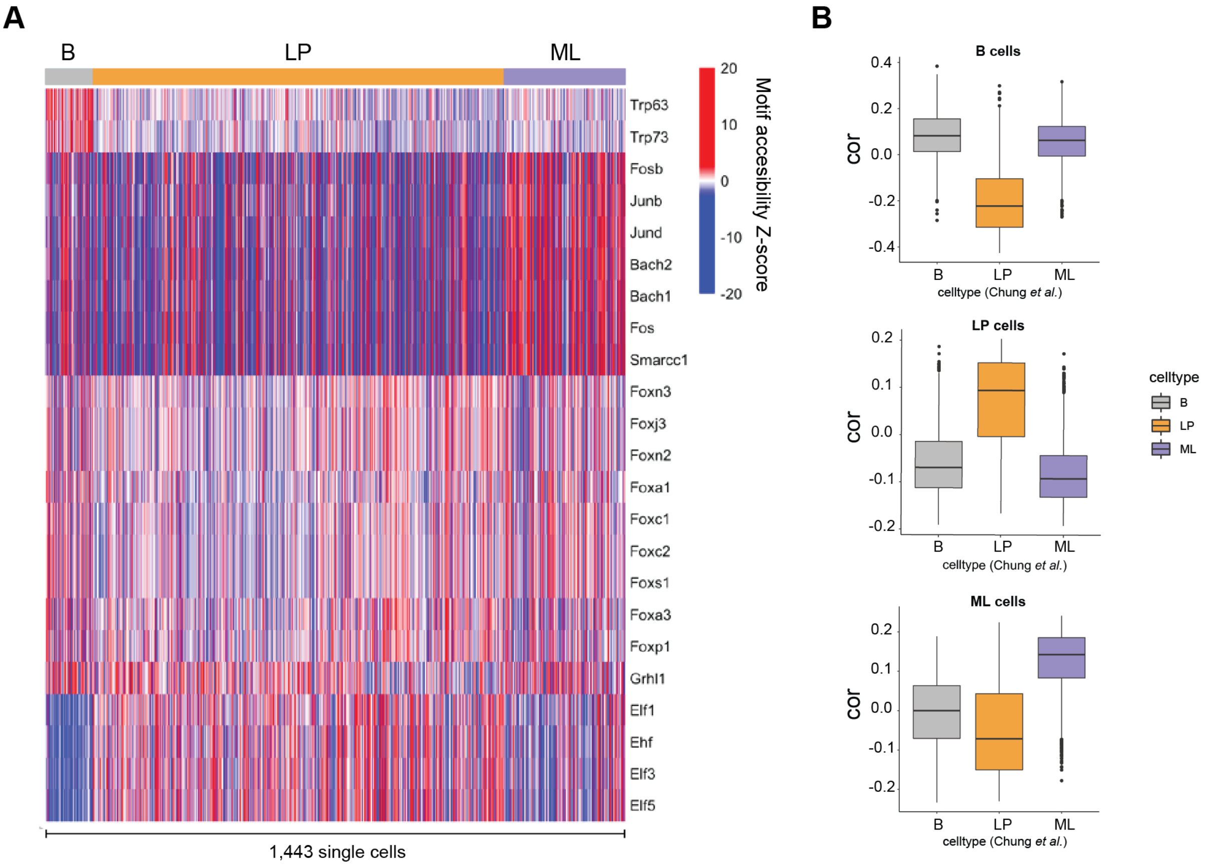


Figure 14 Assessing the correlation in motif accessibility between reference and control scATAC-seq datasets.

(A) scATAC-seq peaks (derived from three different epithelial cell types present in murine mammary gland organoids) were annotated based on the presence of TF motifs. Similarly to the reference dataset, the accessibility of 23 motifs that were used to create a cell-type classifier, differs between B, LP, and ML cells (see Figure 12). (B) Boxplot of correlation scores for B, LP, and ML cells. For each cell type identified in the scATAC-seq dataset from the organoids, we compared how well it correlates with the accessibility of 23 motifs to cell types present in the reference dataset. For each cell type, there was a positive correlation between the two datasets. Heatmap and correlation plots were prepared by Dr. Hyobin Jeong and adapted by me.

Once the first functionality of scNOVA based on supervised cell-type classification was ready, we annotated cell types (luminal progenitor, mature luminal, or basal) to all single-cell Strand-seq libraries generated from murine mammary gland organoids. These results are further explored in Chapter 7

6.3.2 Prediction of gene expression differences between cell populations

The second functionality of scNOVA is the inference of altered gene expression based on the changes in NO at gene bodies (as NO is negatively correlated with gene expression[158]). Such functionality is particularly useful to analyze deregulated pathways in distinct subclones of one sample, or between two conditions of the same sample (for example, before or after drug treatment) regardless of their cell type of origin. scNOVA integrates deep convolutional neural network (CNN) based machine learning, and negative binomial generalized linear models[114]. The CNNs use five features (NO, NO variance, GC content, CpG content, and replication timing) on 150 genomic bins (spanning gene bodies and their ±5 kb surrounding regions) to infer the expression status of the genes (expressed/unexpressed) and then filters out genes inferred to be non-expressed. Then, for those genes predicted to be expressed, scNOVA uses negative binomial generalized linear models (from the DESeq2 package[159]) to compare NO changes at gene bodies and accordingly infer differential expression.

To create the training and test datasets needed to generate the scNOVA model for mice genome and benchmark, I performed both bulk RNA sequencing (to provide the ground truth of gene expression of a particular cell population), and Strand-seq (to generate the data on cell-type-specific NO pattern) on two different cell types: mouse embryonic stem cells (mESCs) from 129 x C57BL/6J strain, and mouse embryonic fibroblasts (MEFs) isolated from FVB/NJ strain. We obtained 62 high-quality Strand-seq libraries from MEFs, and 33 from mESCs. By analyzing pooled NO profiles from single cells of each sample, we observed, as expected, a negative correlation between NO along the gene body and gene expression level (Figure 15A), particularly strong at the transcription start site (TSS) (Figure 15B, for clarity only data from MEFs is shown in panels A and B). We also analyzed the differences in NO at gene bodies using Strand-seq data from MEFs and publicly available scMNase-seq datasets from 48 NIH3T3 single cells (highly rearranged mouse embryonic fibroblast cell line) and 278 mouse naïve CD4 T cells[160]. Based on unsupervised clustering, MEFs locate closely with NIH3T3 cells rather than T cells (Figure 15C), reassuring that Strand-seq-derived NO tracks are consistent with scMNase-seq experiments. Strand-seq libraries from MEFs were then split into two sets (each with 31 cells) and used to build the CNN model. scNOVA version for mice

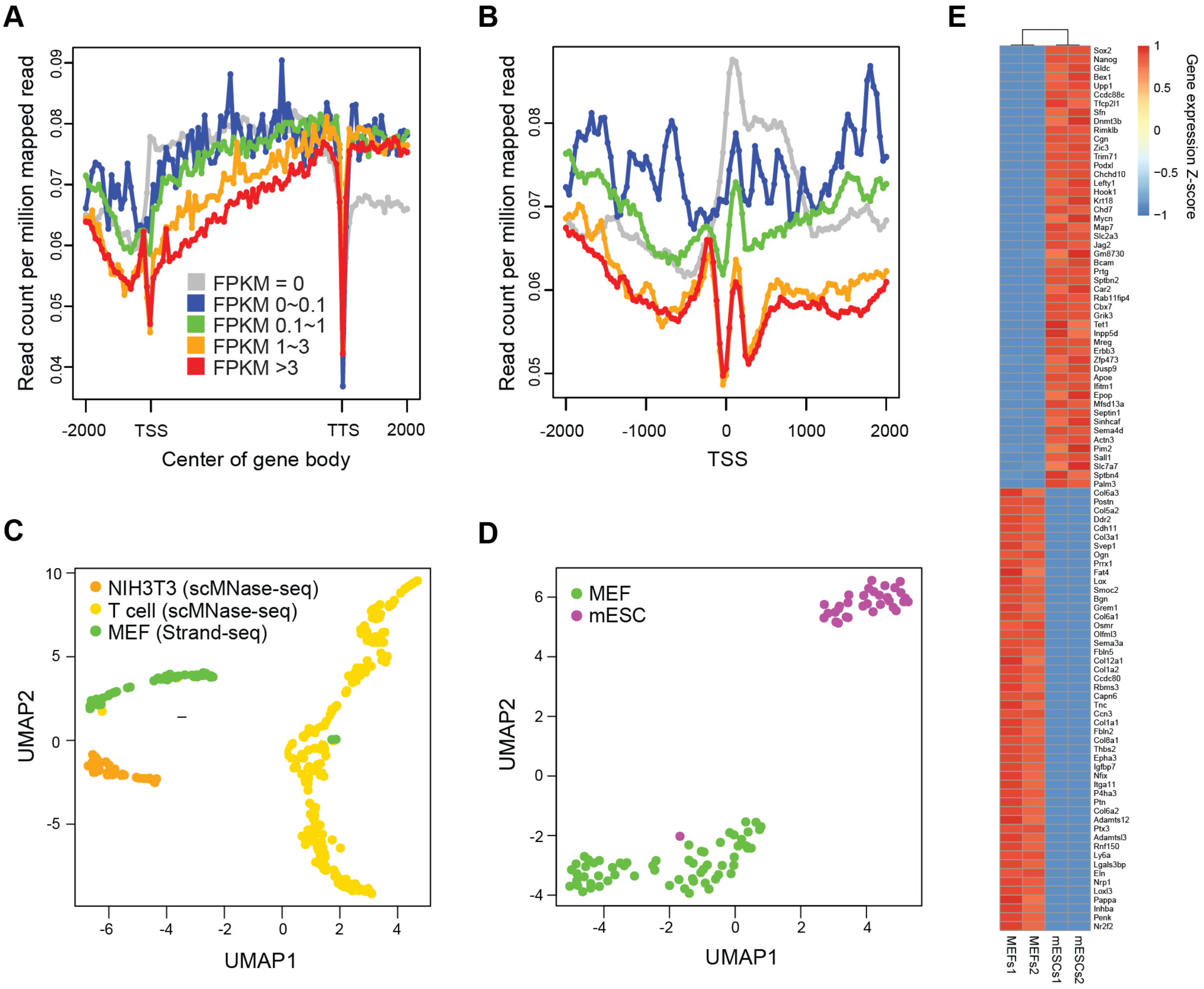


Figure 15 Extending the functionality of scNOVA to infer global patterns of gene expression in murine cells using single-cell NO profiles.

(A) NO along gene bodies ±2 kb for protein-coding genes of MEFs based on 62 pooled Strand-seq libraries (shown on the y-axis as reads per million). Genes were grouped and color-coded based on their expression level from bulk RNA-seq data (FPKM stands for fragments per kilobase of exon per million mapped fragments; from grey: unexpressed to red with FPKM>3: highly expressed). TTS: transcription termination site (B) Nucleosomes are strongly depleted on the TSS of expressed genes. (C) Based on NO at gene bodies derived from Strand-seq (for MEFs) and scMNase-seq (NIH3T3 and T cells), freshly isolated MEFs resemble more NIH3T3 cell line than T cells. (D) The NO at gene bodies obtained by scNOVA correctly separates two different cell type populations (MEFs and mESCs). (E) Top differentially expressed genes between MEFs and mESCs predicted by scNOVA are confirmed by RNA-seq data (two experimental replicates per cell type). Analysis and figures A-D by Dr. Hyobin Jeong.

genome was then benchmarked using Strand-seq-derived NO profiles from MEFs and mESCs. scNOVA correctly predicted changes in gene activity between these two different cell types by analyzing NO at gene bodies (AUC of 0.7593 for the 10 most differentially expressed genes). MEFs and mESCs create separate clusters (apart from one outlier mESC) (Figure 15D). The levels of the 50 most differentially expressed genes between MEFs and mESCs (AUC of 0.7311) predicted by scNOVA were confirmed in RNA-seq datasets (Figure 15E). These genes include well-known markers of stem cells (for example, *Sox2*, *Nanog*, *Tfcp2l1*[161], *Gldc*[162]) and fibroblasts (such as genes encoding proteins of extracellular matrix scaffold from collagen and lysyl oxidase gene families[163]) confirming that scNOVA accurately infers gene activity changes. As an additional proof of principle, we tested if scNOVA will correctly predict gene expression changes in mammary gland organoids (derived from TetO-CMYC/TetO-Neu/MMTV-rtTA mice) following doxycycline induction.

The addition of doxycycline to the medium activates two strong oncogenes (HER2 and CMYC) which induces drastic changes in cell phenotype[102]. We applied scNOVA on Strand-seq libraries generated from never-induced cells and induced for 5 days with doxycycline (described in chapter 6.1). After applying a 10% FDR cut-off, 68 genes showed a change in NO in doxycycline-induced cells compared to not-induced cells. To validate if the predicted genes with decreased or increased NO show expected changes also at the transcriptome level, I analyzed a public dataset containing bulk RNA-seq results from 16 samples of mammary gland organoids of the same inbred strain (8 never-induced, 4 tumor and 4 residual in which doxycycline was removed from the culture medium to silence oncogene overexpression)[103].

However, the RNA-seq results should be considered a proxy of the actual changes in expression level as the organoids used in these experiments have not been seeded and collected at the same time, and the organoids used in the RNA-seq experiment were induced with doxycycline for 2 more days (7 days in total) comparing to the ones used for Strand-seq. I performed differential expression analysis on bulk RNA-seq data using DESeq2[159] including the animal and condition as confounding factors in the model design. 55 out of 68 genes predicted by scNOVA were expressed in dox-induced and never-induced cells (37 genes with statistically significant diffExp), from which 19 showed expected pattern (if a gene shows decreased NO, we expect that it will be highly expressed; increase in NO, should decrease the expression of the gene, accordingly) (Figure 16). For example, among activated genes, there are known Myc-target genes (like *Dctd*, *Itpk1*)[164], and two of the downregulated genes are other members of the EGFR family (*Egfr*, *Erbb4*). Despite certain differences between scNOVA predictions and RNA-seq results, these data indicate that we can indeed infer gene expression changes from single-cell NO profiles using scNOVA.

Taken together, the results presented in this chapter demonstrate the construction of the scNOVA-based supervised cell type classifier for mammary gland organoids. We succeeded in updating scNOVA functionality to infer altered gene expression in Strand-seq data from murine cells. In Chapter 7 I show for the first time in the mammary gland field[110] how this tool, together with scTRIP, can be used to analyze SVs in different cell types or murine mammary gland organoids at a single-cell level.

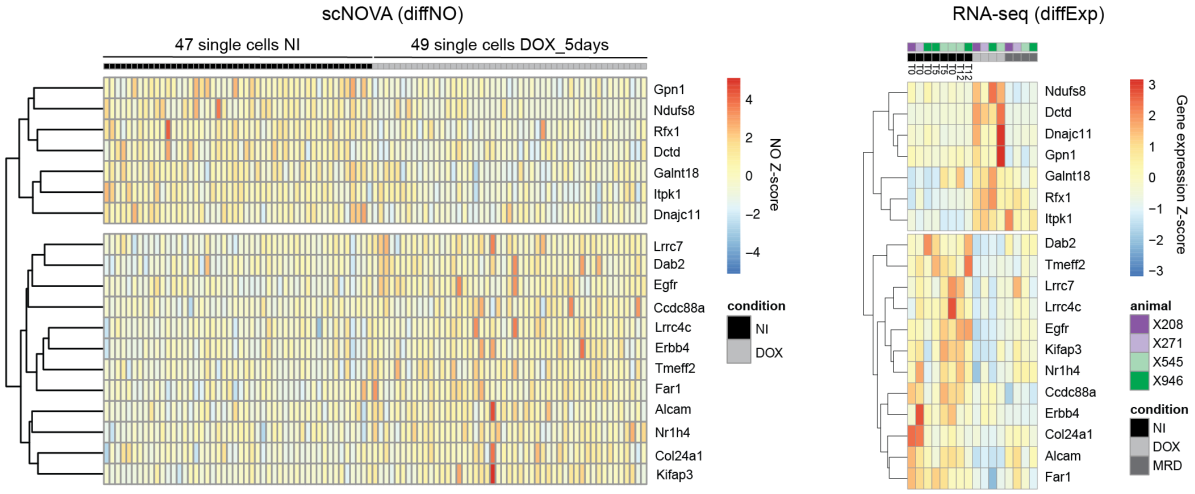


Figure 16 scNOVA confirms changes in NO on cMYC target genes and EGFR family members after doxycycline induction.

scNOVA was used to infer genes with differential NO in cells before and after induction with doxycycline. The actual changes in the expression of the predicted genes were confirmed using previously reported bulk RNA-seq of mammary gland organoids from the same transgenic model[103]. (Left) A heatmap representing NO on genes predicted by scNOVA in single cell Strand-seq libraries before (NI) and after (DOX_5 days) induction with doxycycline. NO is anticorrelated with gene expression, as shown on a heatmap summarizing RNA-seq results (right). Bulk RNA-seq was performed on murine mammary gland organoids: never induced with doxycycline (NI), induced with doxycycline (DOX), or after doxycycline removal (MRD) derived from four different mice (mice ID numbers on the plot). NI cells were collected at three different timepoints: before the organoid seeding (T0), and after 5 and 12 days of organoid culture (T5 and T12, respectively).

Chapter 7 Results: assessing SV burden in murine mammary gland organoids after doxorubicin treatment.

Having optimized the experimental workflow of Strand-seq for organoids, adapted scTRIP for murine genome, and created a scNOVA-based cell-type classifier for murine mammary gland cells, I wanted to combine all these techniques and computational workflows to analyze the SV landscape after doxorubicin treatment in a cell-type specific manner. The annotation of mutational signatures of DNA-damaging and cytotoxic agents has been concentrated primarily on base substitutions and small indels detected in whole genome sequencing data. Here, I present the results based on single-cell sequencing with a particular focus on SVs.

7.1 Overview of the samples profiled with Strand-seq

In chapter 6 I reported generating the first Strand-seq libraries from cells forming the organoids derived from the mammary glands of a murine strain TetO-CMYC/TetO-Neu/MMTV-rtTA. The organoids were either never induced with doxycycline or induced with doxycycline for 5 and 12 days representing healthy tissue or in a cancer state (these samples are labeled as NI, DOX_5days, and DOX_12days). In addition, to optimize scNOVA for mice genome, I made Strand-seq libraries from other murine cells: mouse embryonic stem cells from 129 x C57BL/6J strain and mouse embryonic fibroblasts from FVB/NJ strain. The data from these samples would allow me to detect the germline SVs characteristic for the strain used in this study, assess the baseline level of SVs present in cells of mammary gland organoids (without a drug perturbation) and compare it to other murine cell types.

In order to investigate single-cell SV patterns following exposure to doxorubicin, I performed Strand-seq experiments on doxycycline-induced organoids treated with different concentrations of doxorubicin for 72 hours and then incubated without a drug for up to 7 days (Figure 17A). Strand-seq protocol requires that cells divide exactly once after incorporating BrdU, therefore it was crucial to let the cells recover. Without the 'recovery time', the cells that managed to survive the drug treatment were unable to divide, presumably because of the cell cycle arrest induced by DNA damage. I successfully sorted nuclei after treatment with 10 nM and 100 nM doxorubicin (the following 'recovery time' was 96 h or 7 days, respectively). As an experimental control, for both conditions, I also isolated single nuclei from

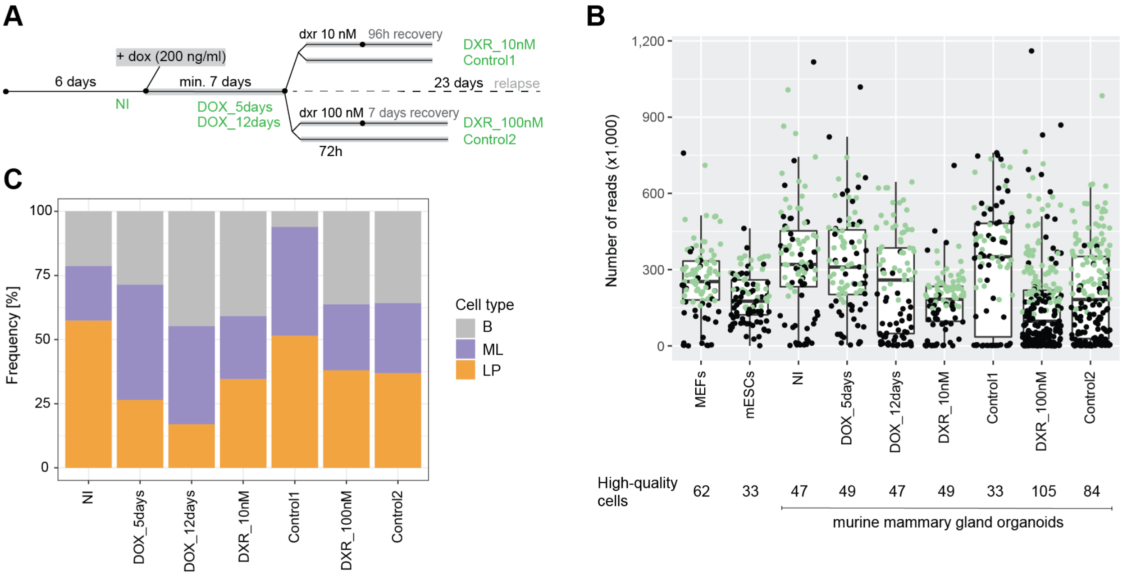


Figure 17 Successful SV annotation and cell-type prediction in more than 400 cells from mammary gland organoids.

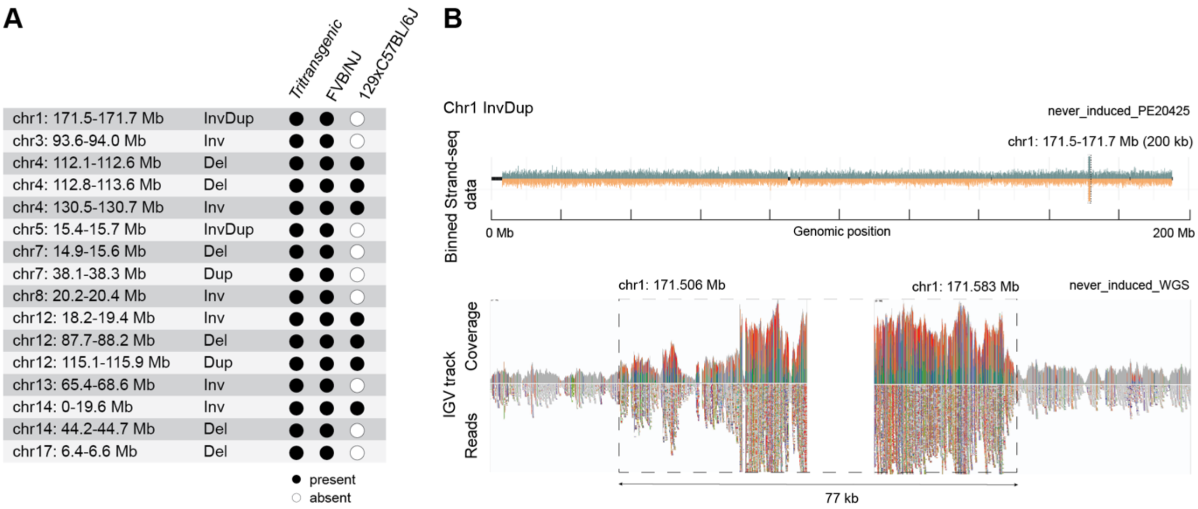
(A) Experimental workflow: mammary glands are isolated from TetO-CMYC/TetO-Neu/MMTV-rtTA and cells are used for seeding the organoids. After 6 days of unperturbed growth, the organoids are treated with doxycycline for 7 days to induce oncogene overexpression (doxycycline is afterward always added to the medium, highlighted in grey). Cells are then treated with two different concentrations of doxorubicin for 72 h (or DMSO in case of controls) and then left to recover (up to 7 days) until they start proliferating again. At the end of the experiment, Strand-seq libraries are generated. Samples collected at different time points corresponding to different stages of tumorigenesis are in green. With the transgenic model used in this study, it is also possible to mimic an idealized situation of relapse: cells are first induced with doxycycline, then doxycycline is removed from the medium and the cells remain in the MRD state until they spontaneously start dividing again and enter cancer phenotype without oncogene addiction. (B) Quality control of Strand-seq cells. Each single-cell Strand-seq library from all different samples and conditions was scored and only high-quality cells (assessed based on the factors described in chapter 6.2) were included for further analysis. Libraries from 0-cell and 100-cell controls are not included in the plot. (C) Each Strand-seq library generated from murine mammary gland organoids was annotated with the most probable cell type (B, LP, ML) based on the outcome of the scNOVA cell-type classifier.

cells (hemi-substituted with BrdU) forming doxycycline-induced organoids that were seeded at the same time and kept in culture for the same period as doxorubicin-treated but were treated with DMSO instead of the drug (I will refer to these samples later as DXR_10nM and its matching control: Control1, similarly: DXR_100nM and Control2). Strand-seq libraries were processed as described in chapter 6.1. The stringent quality control resulted in including the following number of cells for downstream analysis: 62 MEFs, 33 mESCs, and more than 400 cells from mammary gland organoids at different stages of tumorigenesis: 47 NI, 49 DOX_5days, 47 DOX_12days, 49 DXR_10nM, 33 Control1, 105 DXR_100nM, 84 Control2 (Figure 17B). In single-cell genomics each cell can be considered a separate biological replicate but to increase the number of sequenced cells I performed the experiment with a higher concentration of doxorubicin and the matching control three times (using mammary glands from three different mice for organoid seeding). The results from these three experiments are presented here together (in all these replicates I observed a similar frequency of SVs and cell-type contribution).

To analyze Strand-seq data, I used scTRIP, described in detail in chapter 6.2, and PloidyAssignR (developed by a previous student in the group, Tania Christiansen; manuscript in preparation). PloidyAssignR automatically infers the ploidy state of a cell based on the binomial distribution of strand state segregation patterns. As a result, it allows for characterizing subclonal aneuploidy events beyond the diploid genome. After manual curation of scTRIP and PloidyAssignR outputs, I confidently called SVs in single cells ranging in size (from 200 kb to whole-chromosome copy number alterations) and complexity (deletions, duplications, inversions, and complex rearrangements). I also applied the scNOVA-based cell-type classifier to the samples from murine mammary gland organoids, so that each Strand-seq library was annotated with one of the three possible cell types: luminal progenitor, mature luminal, and basal (Figure 17C). In the next subchapters, I focus on the analysis of the germline SVs, somatic SVs induced by doxorubicin treatment, and another marker of genomic instability: SCEs.

7.2 Determination of strain-specific SVs

Germline genomic rearrangements are present in all cells derived from an organism, and in the case of an inbred mouse strain, they are shared by all animals. The inbred mouse strain used in this study (TetO-CMYC/TetO-Neu/MMTV-rtTA) has an FVB/NJ background, and its germline SVs could be also detected in Strand-seq libraries derived from FVB/NJ mouse embryonic fibroblasts, but not mouse embryonic stem cells from 129 x C57BL/6J strain (Figure 18A). These observations imply that the annotated genomic rearrangements are germline SVs unique to FVB/NJ strain and not an error in the mouse reference genome which is based on C57BL/6J strain. In a previous study that aimed to characterize ‘private’ SNPs and SVs of FVB/NJ strain using short-read whole genome sequencing (WGS), the authors reported only the presence of indels (up to 50 bp) and two deletions (between 6-10 kb)[165] which are below the detection limit of Strand-seq (200 kbp). Despite the widespread usage of the FVB/NJ strain, especially for the production of transgenic animals, this strain was not included in WGS studies aiming to catalog SVs in multiple mice genomes[166–169] or in the most recent project utilizing long-read sequencing[170]. Therefore, to support the results from Strand-seq, I isolated genomic DNA from never induced mammary gland organoids of the tritransgenic strain which was then sequenced with 30x coverage using the short-read WGS method. With WGS data, I confirmed the coordinates of germline SVs called with Strand-seq (excluding inversions which are particularly difficult to detect in WGS results) (Figure 18B). Because the resolution of WGS data is much better than that of Strand-seq (1 bp vs 200 kb), there were some differences in the exact size or precise location of SVs which did not influence the overall quality of Strand-seq results (Sup. Table 1). As germline SVs are shared by all cells coming from a particular strain, they were ignored in further analysis.



		Tritransgenic	FVB/NJ	129xC57BL/6J
chr1: 171.5-171.7 Mb	InvDup	●	●	○
chr3: 93.6-94.0 Mb	Inv	●	●	○
chr4: 112.1-112.6 Mb	Del	●	●	●
chr4: 112.8-113.6 Mb	Del	●	●	●
chr4: 130.5-130.7 Mb	Inv	●	●	●
chr5: 15.4-15.7 Mb	InvDup	●	●	○
chr7: 14.9-15.6 Mb	Del	●	●	○
chr7: 38.1-38.3 Mb	Dup	●	●	○
chr8: 20.2-20.4 Mb	Inv	●	●	○
chr12: 18.2-19.4 Mb	Inv	●	●	●
chr12: 87.7-88.2 Mb	Del	●	●	●
chr12: 115.1-115.9 Mb	Dup	●	●	●
chr13: 65.4-68.6 Mb	Inv	●	●	○
chr14: 0-19.6 Mb	Inv	●	●	●
chr14: 44.2-44.7 Mb	Del	●	●	○
chr17: 6.4-6.6 Mb	Del	●	●	○

● present
○ absent

Figure 18 Germline SVs in different mice strains can be characterized using Strand-seq.
(A) Summary of SVs detected in three different strains used in this study. Tritransgenic mice (TetO-CMYC/TetO-Neu/MMTV-rtTA) were created in FVB/NJ background, therefore they share all germline SVs present in the parental line but not 129xC57BL/6J strain. InvDup- inverted duplication, Inv- inversion, Del- deletion (B) Example of an SV (InvDup) detected in Strand-seq data (top) in cells from mammary gland organoids derived from tritransgenic strain. The presence of this SV was confirmed in WGS data (bottom) and visualized in Integrative Genomics Viewer (IGV)[171].

7.3 Doxorubicin induces a wide spectrum of SVs in cells forming murine mammary gland organoids

As next, I searched for SVs present in MEFs and mESCs to get an overview of SV frequency in murine cells. I identified 6 SVs in 4 cells of MEFs (out of 62), and 13 SVs in 10 cells of mESCs (out of 33). Apart from one duplication, all SVs detected in MEFs were either chromosome losses or gains. Similarly, in mESCs majority of events were whole-chromosome copy number changes. Interestingly, 2 cells of mESCs had trisomy 8, and 1 cell had trisomy 8 and 11. Trisomy of these two chromosomes has been shown to be clonally selected in mESCs giving growth advantage (while preserving their differentiation potential)[172,173]. A higher frequency of chromosomal abnormalities detected in mESCs passaged for a longer period indicates that these cells are more chromosomally unstable compared to MEFs, freshly isolated from embryos and kept in culture for short time.

I then focused on somatic SVs present in the cells of murine mammary gland organoids before and after doxorubicin treatment. I intended to determine if exposure to doxorubicin could be associated with a specific SV pattern and whether different cell types would show different frequencies of SVs. There were no *de novo* SVs present in the NI sample which represents normal cells of the murine mammary gland (Figure 19A). Cells induced with doxycycline for 5 and 12 days had 5 and 6 SVs, respectively (present in 3 out of 49 cells, and in 4 out of 47 cells). Cells treated with doxorubicin were characterized by the highest number of SVs: in DXR_10nM 6 cells out of 49 had at least one SV (7 SVs in total) and in its matching control (Control1) only 1 cell out of 33 had an SV. In DXR_100nM 39 cells out of 105 had at least one genomic rearrangement with a total number of 80 somatic SVs, while in Control2 the frequency of SVs was much lower: 4 cells out of 84 with at least one SV (6 SVs in total). Apart from the samples DXR_10nM and Control1, chromosome gains and/or losses were present in all conditions (Figure 19B). Together with the data from MEFs and mESCs, this suggests that many SVs present in normal or cancer cells are a consequence of mitotic errors that lead to chromosome number changes. Deletions and duplications were present in all samples (excluding NI) with different frequencies (Figure 19C). Complex events occurred exclusively after doxorubicin treatment (Figure 19D), and deletions were the most common SV type induced by doxorubicin. All the events that I identified in all samples from murine mammary gland organoids were singleton SVs (detected only in a single cell). This indicates that treatment with doxorubicin induces genomic heterogeneity within the population of mammary gland cells as new potential subclones emerge.

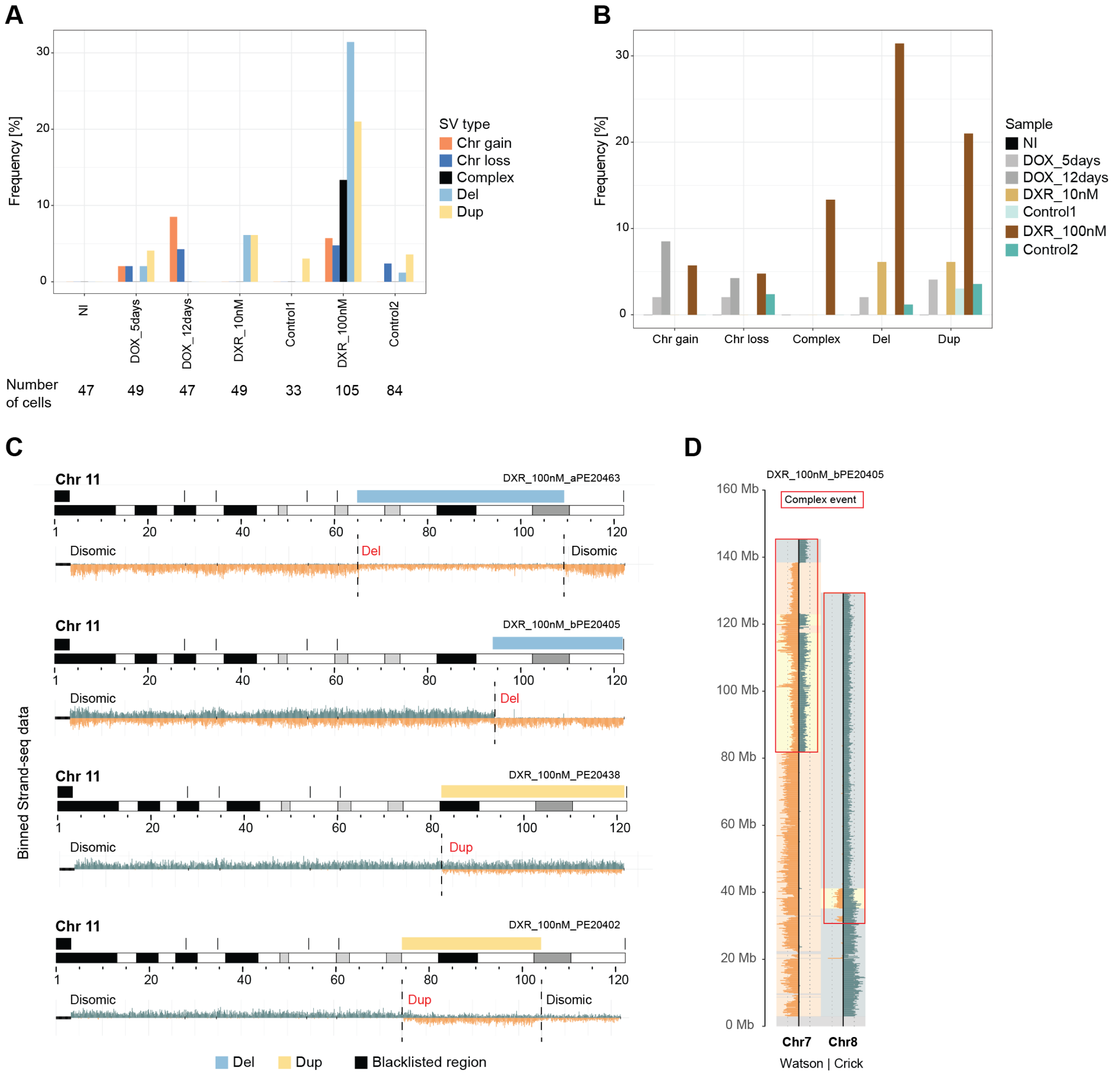


Figure 19 Doxorubicin-induced SVs vary in type and complexity.

(A) Cells treated with a higher concentration of doxorubicin (DXR_100nM) had the highest frequency of SVs among all the samples that had been sequenced for this project. As the number of high-quality libraries differed between the samples, the data was normalized per condition. Even in the cells that were induced only with doxycycline (DOX_5days, DOX_12days), or were used as controls for the drug treatment (Control1, Control2), there were already some SVs detected but at very low frequency (and as expected, the majority of them were mitotic errors that led to chromosome gains or losses).

(B) Summary of different types of SVs detected in Strand-seq libraries from murine mammary gland organoids. The data presented in this graph is the same as in (A) but such visualization allows to highlight that complex events are present only after doxorubicin treatment and that deletions are the most common SVs induced by the drug. (C) Examples of intrachromosomal and terminal deletions and duplications discovered on chromosome 11 in different cells of DXR_100nM sample. (D) Examples of complex rearrangements involving clustered deletions, amplifications and/or inversions discovered on two chromosomes in one of the cells from the DXR_100nM sample.

Deletions and duplications reported for 105 cells of DXR_100nM were present on all chromosomes and there was no bias toward a certain chromosome having a higher number of SVs than the other chromosomes (permutation test, 10,000 permutations, *p*-value 0.7469) (Figure 20A). Such an observation suggests that all chromosomes are equally vulnerable to DNA damage caused by doxorubicin. Among deletions and duplications, there were both terminal and intrachromosomal affecting one haplotype. The size of doxorubicin-induced deletions did not follow a normal distribution (Figure 20B) but it is difficult to assess whether this effect is an actual consequence of doxorubicin treatment that could be associated with its mechanism of action, or whether we are simply more biased towards the discovery of bigger SVs as they are more obvious to detect. 14 complex events detected in 10 cells were affecting the entire chromosome suggesting that chromothripsis occurred, or parts of chromosomes indicating other complex and ongoing DNA rearrangement processes.

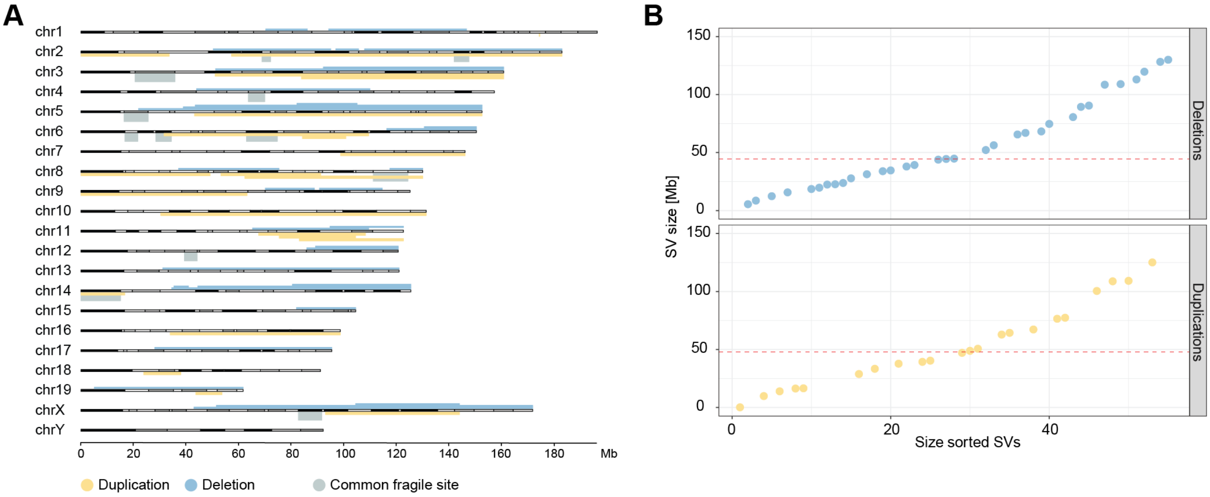
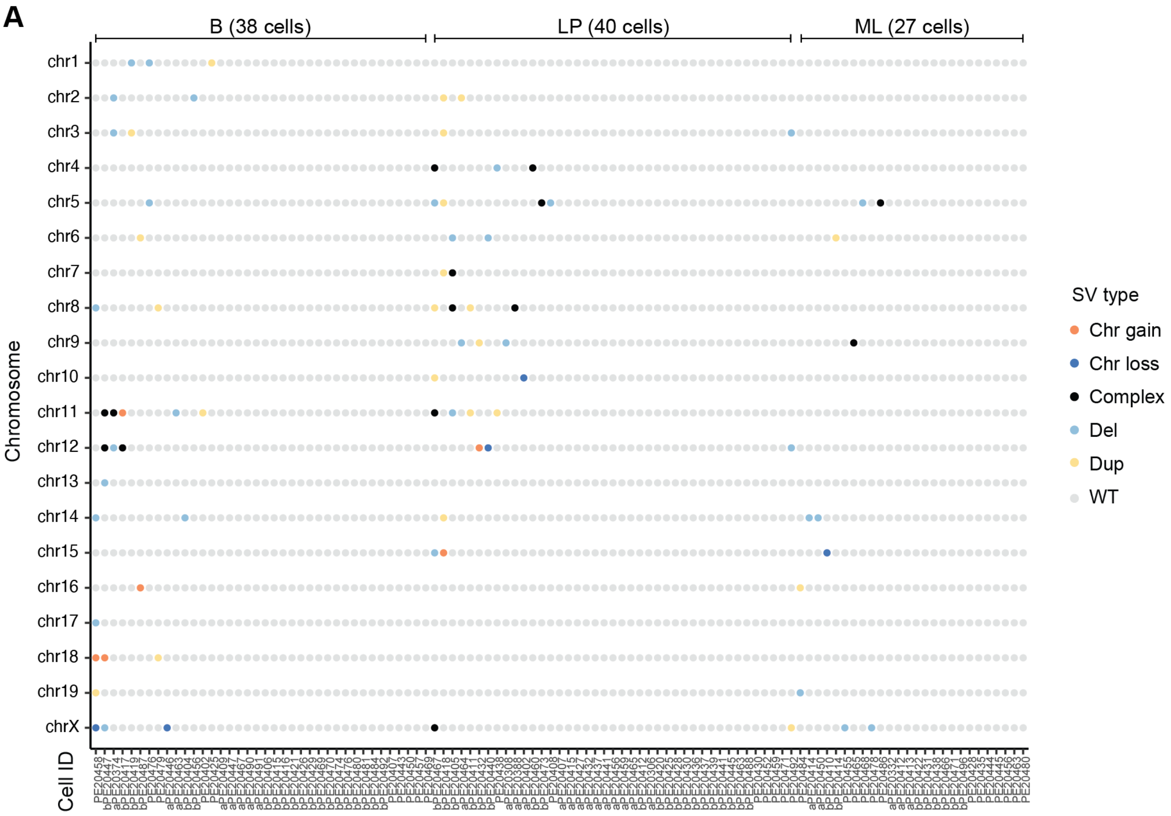


Figure 20 After doxorubicin treatment the cells are affected by deletions and duplications of different sizes.

(A) Karyogram showing the location of 33 deletions (in blue) and 22 duplications (yellow) detected in 105 cells of DXR_100nM. Common fragile sites of the murine genome are highlighted in grey. All cells were derived from female mice, therefore no SVs were called on the Y chromosome. (B) The size distribution of copy-number variants (deletions in blue, duplications in yellow) presented in panel A. Chromosome gains and losses are not included. The median size for each SV class is indicated as a red line. Individual SVs are plotted along the X-axis in order of increasing size. Doxorubicin-induced deletions do not follow a normal distribution (Shapiro-Wilk normality test: for deletions *p*-value=0.01633, for duplications *p*-value: 0.3603).

Taking advantage of the data from the scNOVA-based cell-type classifier, I was able to summarize the frequency of SVs in different cell types. In all analyzed samples libraries were coming from three different mammary epithelial cells: LP, ML, and B. All different SV types were detected in all three mammary epithelial cells of DXR_100nM (Figure 21A). In samples exposed to two different concentrations of doxorubicin and their corresponding controls, the frequency of SVs was similar in all cell types (Figure 21B). For example, after the treatment with a higher concentration of doxorubicin, between 35-40% of cells from each cell type had at least one SV. The frequency of SVs was much lower in the cells of Control2 but shared between all three cell types present in that sample. These observations imply that all three epithelial cell types of mammary glands in the cancer state are equally susceptible to DNA damage and resulting genetic alterations.

Taken together, this data shows that the increased frequency of deletions and the presence of complex events can be considered a mutational signature of doxorubicin. At the genomic level, different cell types present in the murine mammary gland share a much alike response to the drug.



B

DXR_10nM

Cell type	Total	Cells with SV	% cells with SV
LP	17	2	11.8%
ML	12	2	16.7%
B	20	2	10%

DXR_100nM

Cell type	Total	Cells with SV	% cells with SV
LP	40	14	35%
ML	27	11	40%
B	38	14	37%

Control1

Cell type	Total	Cells with SV	% cells with SV
LP	17	1	5.9%
ML	14	0	0%
B	2	0	0%

Control2

Cell type	Total	Cells with SV	% cells with SV
LP	31	1	3.2%
ML	23	2	8.7%
B	30	1	3.3%

Figure 21 Increased genomic instability after doxorubicin treatment is not cell-type-specific.

(A) Dotplot summarizing all SVs on different chromosomes detected in 105 cells of DXR_100nM. Cells are grouped based on the predicted cell type, and within each group, they are ordered according to the total number of SVs. (B) Frequency of SVs detected in different cell types of doxorubicin-treated and control samples. There is a statistically significant association between the frequency of SVs and treatment with higher (but not lower) doxorubicin concentrations compared to control (2x2 contingency tables, Fisher's exact test, *p*-value <0.001) but in all the samples there is no correlation between the frequency of SVs and cell types (2x3 contingency tables, the Freeman-Halton extension to Fisher's exact test).

7.4 Increased frequency of SCEs after doxorubicin treatment

The increased prevalence of genomic rearrangements is a clear example of genomic instability. SCEs, which occur during the S phase and mitosis when DSBs are repaired by homologous recombination pathways, are another marker of genomic stress[122]. In the next step, I wanted to check if the treatment with doxorubicin correlates with increased SCE frequency. SCEs are typically undetectable by single-cell sequencing technologies but Strand-seq was actually developed to help map such events. SCEs can be detected as points on chromosome plots where reads mapping to both Watson and Crick strands switch to reads mapping to either the Watson or the Crick strand (without affecting the average read count). Rare events of complete switches from Watson-only to Crick-only template strand reads might be an indication of misorientation in the reference genome rather than SCEs. I observed a complete switch at exactly the same location in chromosome 14 (between 19.3-19.6 Mb) in every Strand-seq library in which that region inherited both Watson or both Crick template strands (Figure 22A). As reported for the mm9 reference genome[111], the region of the template switch corresponds to an unbridged gap of unknown sequence that is difficult to map, and the relative orientation of contigs around it is not confirmed.

The observation that this event happened in multiple cells coming from different murine strains included in this study (a total of 37 libraries from MEFs, 14 from mESCs, 206 from mammary gland organoids) supports the hypothesis that the orientation of the contigs in this area might be erroneous also in the mm10 reference genome. Indeed, the orientation of the contigs on chromosome 14 (reported in the previous study to be wrong in mm9) has not been corrected in mm10.

For all the samples, including MEFs and mESCs, I annotated the number of SCEs in each single-cell Strand-seq library. Regardless of the experimental condition, SCEs were detected on all chromosomes, and their number per chromosome directly correlated with the chromosome length suggesting that all chromosomes are prone to DNA breaks repaired via SCE (Figure 22B).

Libraries from MEFs, mESCs, NI, DOX_5days, DOX_12days, Control1, and Control2 had on average four-five SCEs per cell, while treated with doxorubicin more than eight (with up to 24) (Figure 22C).

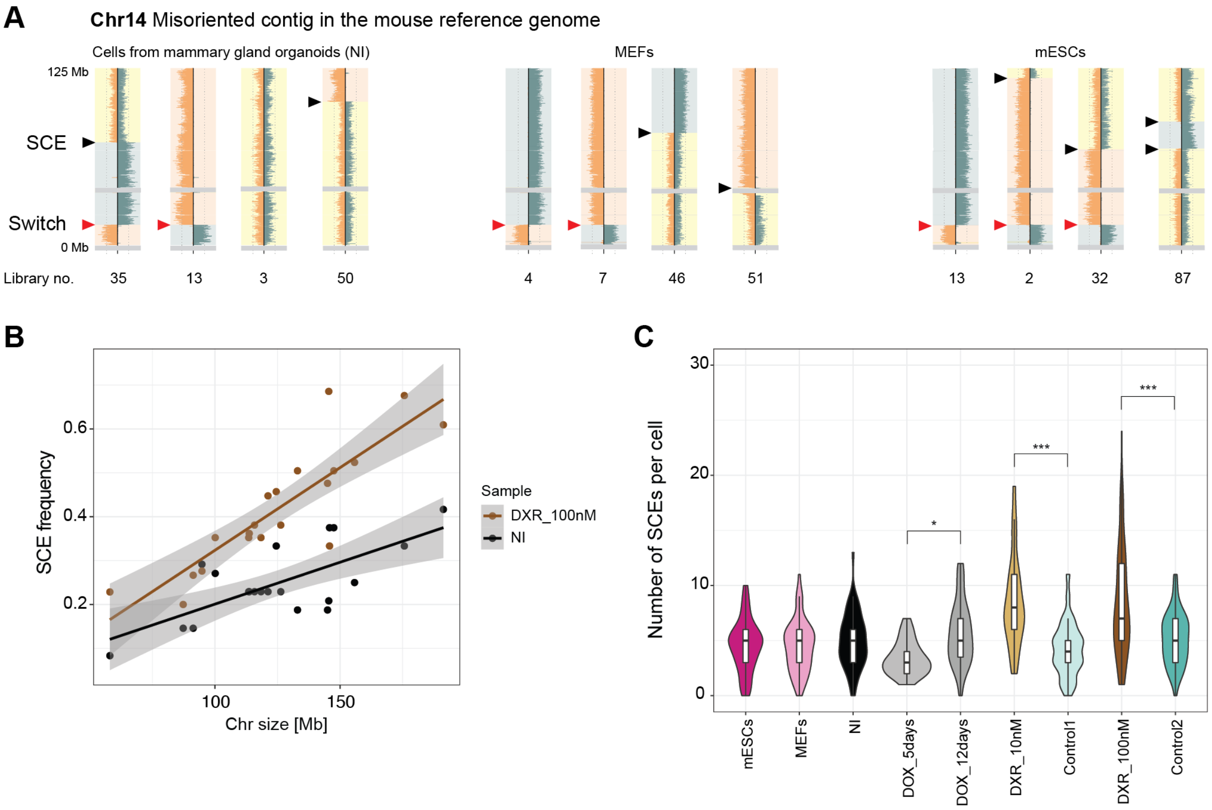


Figure 22 Identification of SCE events in Strand-seq data.

(A) Ideograms of chr14 from different exemplary single-cell libraries (indicated with library number) generated from NI, MEFs, and mESCs (all three samples are derived from different murine strains). Orange and green lines represent reads aligning to the Watson and Crick strands, respectively. Grey areas correspond to blacklisted regions. A switch from Crick/Crick (or Watson/Watson) to Watson/Crick reads can be seen in libraries 35 and 50 from NI, 46 and 51 from MEFs, and 2, 32, and 87 from mESCs indicating that SCE events have occurred (black arrow). A complete switch from Watson to Crick reads (red arrow) is present in all Strand-seq libraries from murine cells in which two Watson or Crick template strands are inherited but is not apparent if both Watson and Crick templates are inherited (libraries 3 from NI, 46 from MEFs and 87 from mESCs). (B) In all the analyzed conditions, the frequency of SCEs per chromosome increases with chromosome size. SCE frequency is equal to the total number of SCEs detected on each chromosome divided by the number of libraries in the sample. Chromosome sizes (corrected for blacklisted regions) are plotted on the X-axis. For clarity, only data from two samples is shown on the graph. (C) Total number of SCEs counted in each single Strand-seq library generated from mESCs, MEFS, and different conditions of murine mammary gland organoids. Only statistically significant differences are indicated (one-way analysis of variance (ANOVA), *p*-value=3.29E-26, with Tukey post-hoc tests). Different *p*-value thresholds are represented as follows: *p*<0.05 (*), *p*<0.001 (***).

The frequency of SCEs in MEFs, mESCs, and cells of mammary gland organoids not exposed to doxorubicin, corresponds to the previously reported values in mESCs (based on cytogenetic studies) and to the values observed in hematopoietic stem cells from the bone marrow of healthy donors (data from Strand-seq experiments, shared by Dr. Karen Grimes, unpublished).

Focusing more on the data from the DXR_100nM sample, cells with at least one SV (39 cells) had a similar frequency of SCEs to the ones in which I did not detect any genomic rearrangement (66 cells) (Figure 23A). Different cell types present in mammary gland organoids showed within each sample a similar frequency of SCEs (Figure 23B). This observation, corresponding to the one about the frequency of SVs in different cell types, supports the claim that the genomic response to doxorubicin is independent of the cell type present in the mammary gland.

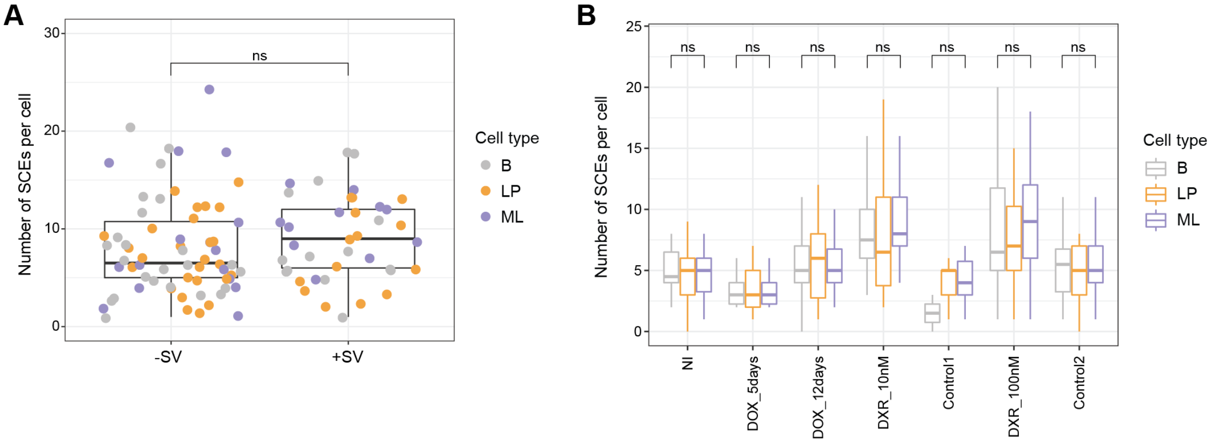


Figure 23 SCE frequency does not depend on the presence of SVs or cell type.

(A) Based on the data from the sample DXR_100nM, cells without SVs (-SV; 66 cells) and cells with at least one SV (+SV, 39 cells) have a similar frequency of SCEs per cell. Each dot represents a single library, additionally color-coded based on the annotated cell type. Statistical significance was calculated with the Mann-Whitney U test (p-value=0.14156). (B) All single-cell libraries from different samples of murine mammary gland organoids were annotated with a cell type, and then within each sample, the frequency of SCEs was compared between different cell types. In all conditions, the differences in SCE frequency between cell types were non-significant (ns) according to one-way ANOVA.

As a following step, I checked if the SV breakpoints as well as the location of SCEs correlated with common fragile sites (CFSs). CFSs are specific genomic loci associated with a higher tendency to form gaps or breakages leading to chromosomal rearrangements and copy number variation. They usually span a genomic region of between several kilobases to megabases, and many of them are located within, or in close proximity to large genes. The activity of CFSs is very often cell-type-specific and dependent on the chemical that acts as an inducer[174]. Contrary to CFSs in the human genome, CFSs in the murine genome are understudied and their exact location is not precise. Two previous studies that focused on murine CFSs used an old mouse reference genome (mm9 or older) and the authors reported the location of CFSs as entire cytogenetic bands[175,176]. This creates a problem as cytogenetic bands do not always correspond between reference genomes. Therefore, I updated the annotation of common fragile sites in mm10. Based on the literature review, I identified 12 murine CFSs, their human homologs, and if possible, large genes that are associated with these fragile sites. I then revised whether the chromosome bands containing each CFS from mm9 should be corrected in a new reference genome. As an additional control, I performed a liftover (conversion of coordinates between assemblies) of the coordinates from human CFSs[177] to mice genome and confirmed a proper annotation of murine CFSs (Sup. Table 2, Sup. Figure 7). Deletions and duplications after doxorubicin treatment are not enriched within CFSs. Based on the scTRIP output, I annotated the location (±100 kb) of all SCEs from DXR_100nM (874 SCEs in 105 cells) and Control2 samples (430 SCEs in 84 cells), and then checked if they cluster in certain CFSs. Only after doxorubicin treatment, there was a tendency of SCEs to be enriched in FraXC1 (Figure 24) (permutation test, 10,000 permutations, *p*-value 0.0848).

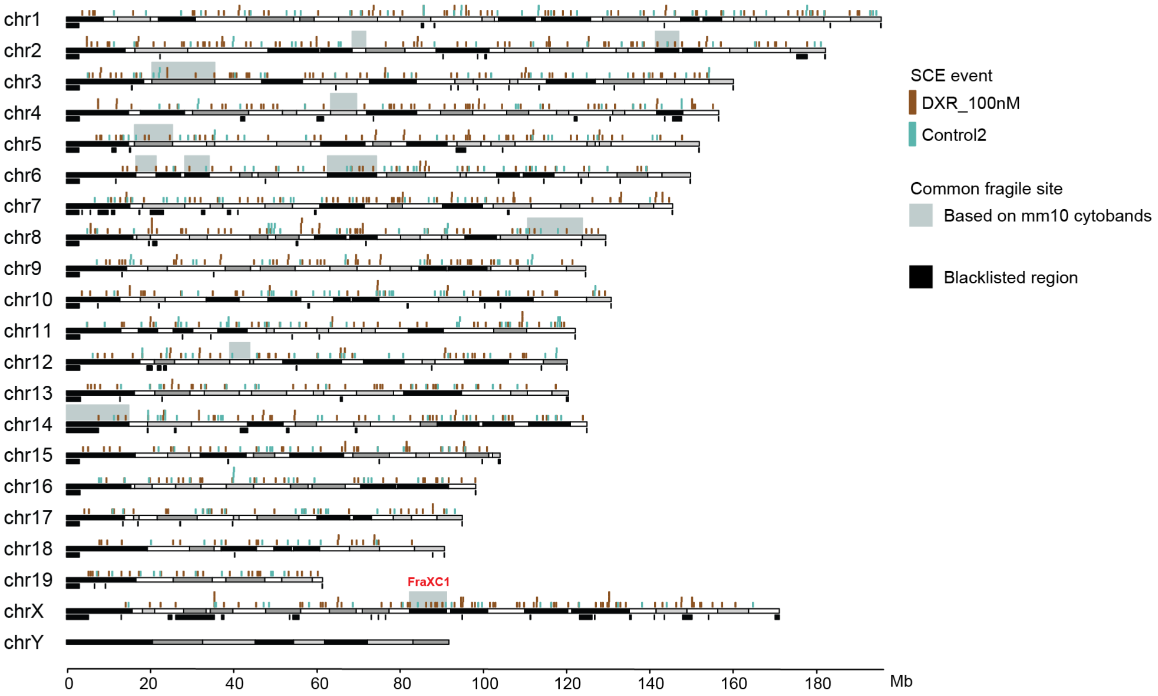


Figure 24 SCE events are not more frequent in murine CFS after doxorubicin treatment.

The location of SCE events (±100 kb) was annotated for all the libraries from DXR_100nM and Control2 samples (874 events from 105 cells of DXR_100nM and 430 events from 84 cells of Control2) and visualized on a karyogram. All cells in this study were derived from the mammary glands of female mice, therefore chromosome Y was not included in the analysis. To check if the SCEs are enriched within common fragile sites (indicated in grey), for each CFS I performed a permutation test (with 10,000 permutations) to assess if the overlap is higher than expected by chance. Only for the CFS FraXC1 (highlighted in red), there is a tendency of SCEs in DXR_100nM samples to be enriched (*p*-value=0.0848; the result of the test for Control2 and this locus was insignificant). In the permutation tests, the blacklisted regions (in black) were masked from the mm10 genome.

Together the data included in this chapter provides evidence that doxorubicin promotes the formation of SVs, including complex rearrangements, and its effect is similar in three main cell types present in mammary gland organoids. The increased genomic instability, indicated also by a higher frequency of SCE events, results in karyotypic diversity and may be one of the factors accelerating tumor evolution under cancer treatment.

Chapter 8 Discussion

Genomic instability, defined as an increased rate of acquisitions of chromosomal alterations and mutations, is a phenomenon observed in most human cancers. High-level of genomic instability affects tumor evolution and correlates negatively with patient prognosis in numerous tumor types[178]. Chemotherapy, a standard-of-care against cancers, targets cells of high genome instability inducing additional DNA damage and putting overwhelming pressure on the DNA repair pathways, eventually leading to cell death[179]. DNA-damaging cytotoxic agents can significantly change the mutational profile of cells that survive the treatment, leading to further genomic instability and promoting heterogeneity. In this sense, chemotherapy can be a double-edged sword, which on one side destroys cancer cells, but on the other sets the stage for drug resistance and a future relapse. Indeed, the relapsing clones carry mutational signatures associated with previous chemotherapeutic exposure. However, the role played by such mutations is still underexplored and this is especially true for SVs, as most studies focused on mutational signatures have been so far restricted to base substitutions and indels. In addition to the DNA damage itself, the cell-type effect and transcriptional states can contribute to resistance, survival, and the eventual establishment of a minimal residual clone.

In this thesis, I explored the consequences of doxorubicin treatment on the genome and transcriptome of cancer cells, with a particular interest in cell-type-specific consequences, using an organoid model of HER2-positive breast cancer and single-cell technologies.

8.1 Organoid technology in disease modelling and drug testing

Breast cancer is highly diverse and in fact, each of its subtypes can be considered a separate disease. One of them, characterized by overexpression of two strong oncogenes: HER2 and CMYC, has an aggressive phenotype and particularly poor prognosis. After the initial treatment, many patients will experience relapse, and such cases are difficult to treat as the cancer cells have been already exposed to very potent therapy regimens that promoted the pro-survival response despite the toxic effect of the drugs through genomic, epigenomic, transcriptomic and/or metabolic changes. In this project, I aimed to identify, on a genomic and transcriptomic level, how a chemotherapeutic drug influences an adaptation potentially leading to therapy failure. To do so, I used an organoid model of HER2-positive breast cancer with CMYC overexpression[102,103]. In this system, primary epithelial cells are harvested from mammary glands of TetO-cMYC/TetO-Neu/MMTV-rtTA mice (adult, nulliparous females) and cultured in 3D, with Matrigel mimicking the extracellular matrix. The organoid cultures have been shown to better represent the physiological conditions and cell-to-cell interactions than standard 2D cell lines. With the system described in this thesis, it is possible to recapitulate different stages of tumorigenesis depending on whether doxycycline is added to the medium, from normal tissue to tumor initiation and progression (after induction of HER2 and CMYC overexpression). Importantly, the transgenic murine strain used in this study comes from an inbred line providing a homogenous genetic background and a reproducible, controlled model.

Apart from disease mimicking, the organoids are suitable for drug toxicity testing. I performed a small-scale drug screen on murine mammary gland organoids to identify a drug for follow-up experiments and the concentration range in which it would be active. In the screen, I included three drugs that are particularly relevant for breast cancer patients: two chemotherapeutics: doxorubicin and paclitaxel, and an example of targeted therapy: lapatinib. Among the three tested drugs, only doxorubicin showed a clear dose-dependent cytotoxic effect. Interestingly, the viability of cells forming murine mammary gland organoids, both not-induced and induced with doxycycline, was not strongly affected even with very high concentrations of paclitaxel. This indicates that the cells may have intrinsic resistance to this particular drug and such mechanisms of resistance do not protect the cells from the toxic effect of doxorubicin. The other possibility is that doxorubicin and paclitaxel have different efficiency of diffusion within Matrigel. It has been reported that doxorubicin diffuses well within 3D spheroids derived from human breast cancer cell lines growing in Matrigel[180] but it may not be the case for paclitaxel. Such observation additionally strengthens the point of using organoids rather than the conventional 2D cell cultures for drug discovery, repurposing, and testing as 3D cultures with the extracellular matrix of various stiffness and concentration gradients can better resemble the complex environment in which tumor cells grow[181].

The other surprising result was connected with the treatment with lapatinib. Lapatinib, a small molecule tyrosine kinase inhibitor of both HER2 and EGFR, approved for advanced HER2-amplified breast cancer[131], was similarly active against cells with and without HER2 and CMYC overexpression. The results of additional RT-qPCR experiments presented in chapter 4.3 did not explain why this drug had a cytotoxic effect on cells with a basal level of HER2 expression. Considering the high variability across experiments with paclitaxel and lapatinib, I decided to focus on doxorubicin, which showed a more consistent cellular response and higher experimental reproducibility.

8.2 Single-cell transcriptome profiling after doxorubicin treatment reveals shared responses of distinct cell types to the drug

As a first step, I wanted to assess the transcriptomic heterogeneity of the population shortly after doxorubicin treatment. Characterizing the phenotypic diversity provides additional insight into the potential genetic changes that can be associated with the drug as the transcriptional response has a more immediate effect on cellular phenotypes. I performed scRNA-seq on murine mammary gland organoids in a cancer state (induced with doxycycline) treated with 100 nM doxorubicin for 72 hours and then let to recover for the next 72 hours in a drug-free medium. Such an approach, although does not fully represent the situation experienced by the patients in the clinics, allows to analyze the initial response of cells to doxorubicin after a single exposure to the drug. Another strategy applied often in the studies focusing on resistance, including doxorubicin resistance[79,182], is to perform a continuous treatment with increasing concentrations of a drug, with or without recovery phases until the cells are adapted to a maximum tolerated dose. Such a method is also not perfect as most chemotherapeutic regimens include repeated dosing at regular intervals over months, optimized for a particular patient, their body weight, and their condition. Transferring a clinical scheme for *in vitro* experiments is challenging, if not impossible[183], therefore reductionist models are needed.

With the results of the scRNA-seq experiment, I wanted to annotate, for the first time, which cell types are present in the murine mammary gland organoids, which genes and pathways are deregulated in doxorubicin-treated cells compared to DMSO-treated control, and whether there are differences between cell types how they react to the drug.

Based on the expression of previously reported cell-type specific markers[136–139], I identified 3 main clusters of epithelial cells in murine mammary gland organoids: luminal progenitors, mature luminal, and basal. Within a basal compartment, I detected cells that resemble a more differentiated or specialized myoepithelial state. In addition, the dataset contained a very low number of fibroblasts. These cells were probably transferred from the stroma during the isolation of mammary glands and survived the culture in Matrigel. As shown also by immunostaining, both luminal and basal cells are present in a single organoid and the luminal cells are more common than the basal. These results confirm that mammary gland organoids represent well the cell-type composition of the tissue from which they are derived.

Treatment with doxorubicin had a profound effect on the transcriptome of the cancer cells that survived. Products of downregulated genes were involved in fundamental cellular processes like respiration, energy production, translation, or peptide synthesis. Cells did not progress in the cell cycle but upregulated their stress response. They did not show the resistance or senescence-associated phenotype presumably because a single drug treatment is not strong enough to promote such a reaction. Interestingly, among upregulated processes in doxorubicin-treated cells were the ones connected with cell motility and migration. Similar observations were recently reported by different groups and the authors demonstrated using human breast cancer cell lines that sublethal doses of doxorubicin activate pro-invasive programs[184–186].

The population of basal cells was particularly decreased after the drug treatment which may indicate that these cells are more sensitive to the drug than the luminal lineage. However, the cell-type-specific analysis did not provide a clear explanation of why luminal cells would be better adapted to the toxic effect of doxorubicin. After the drug treatment, all cell types downregulated the expression of *Top2a*, which product is a direct target of doxorubicin. There are indications that basal and luminal cells differed in the expression of the enzyme involved in the detoxification of doxorubicin (*Cbr3*) but additional experiments would be needed to validate this hypothesis. The data from the scRNA-seq experiment showed that after a single doxorubicin treatment, cells drastically altered their transcriptome in a largely similar way within their population.

8.3 Progress and challenges of multi-omics profiling of heterogeneous cell populations

Considering that the main mechanisms of action of doxorubicin involves DNA damage, I intended to assess the impact of the drug on the genomes of cancer cells by measuring the frequency of *de novo* SVs. It is now possible to do that with the updated versions of computational frameworks, scTRIP[113] and scNOVA[114], made to facilitate and advance the analysis of Strand-seq data.

In Chapter 6 I showed that Strand-seq libraries can be generated from cells forming organoids. This is the first application of Strand-seq in organoids, and it opens up the potential use of the method in solid tumors. Also in Chapter 6 I presented a joint effort to adapt scTRIP workflow for the detection of SVs in the mouse genome. We established a scNOVA-based cell-type classifier using publicly available scATAC-seq data from murine mammary cells[140] as a training set. As a result, we are now able to annotate one of the cell types (LP, ML, basal) to Strand-seq libraries prepared from cells of murine mammary gland organoids. Such a multi-omic approach allows the discovery of SVs in a cell-type aware manner. This is significant progress for the mammary gland field as so far, the single-cell DNA-sequencing methods alone were limited in providing the information about a cell state and there is no need for an additional indexing or selection step with the scNOVA-based classifier.

However, there are some limitations to how Strand-seq, scTRIP, and scNOVA can be used. As the resolution of Strand-seq is currently set to 200 kb, rearrangements smaller than that are not detected with this method. Strand-seq protocol requires that the cells are BrdU-labelled, therefore the libraries can be made only from mitotically-active cells and not fixed samples. Because the isolation of single nuclei relies on FACS, there is also a minimum number of cells needed as input. The scNOVA-based cell-type classifiers have been generated only for human bone marrow and umbilical cord blood hematopoietic stem and progenitor cells, and murine mammary glands. Expanding this functionality of scNOVA to different cell types would require additional reference training datasets. In this project, we applied the classifier to the Strand-seq libraries from cells of murine mammary-gland organoids before and after doxycycline induction, so in normal and cancer state. We additionally validated that the accessibility of the motifs used to create the classifier is shared between corresponding cell types from the reference and scATAC-seq data from not-induced murine mammary gland organoids. However, we cannot exclude that the longer the culture period in presence of doxycycline or a drug, the motif accessibility patterns present in normal cell types would change over time. Taken together, despite certain constraints, this method offers a unique means to study SVs in single cells while preserving the information about their cell type.

8.4 Identification of germline and somatic SVs in mouse genomes

Throughout this project, I analyzed Strand-seq data coming from different murine samples: MEFs, mESCs, and mammary gland organoids. Strand-seq libraries from mESCs and MEFs were generated to create and test one of the scNOVA functionalities, while with Strand-seq experiments on mammary gland organoids, I wanted to assess the mutagenic effect of doxorubicin. MEFs, mESCs, and mammary glands were derived from three different murine strains, and for each of them, I summarized germline SVs specific to a particular line. I excluded these germline SVs in the further analysis as they are shared by all the cells coming from a specific mouse strain. It was previously shown that the mm9 mouse reference genome assembly contains some misoriented contigs that span nearly 1% of the genome[111]. I was able to confirm that one of the reported contigs on chromosome 14 is still misoriented in the mouse reference genome mm10.

After the successful annotation of germline SVs, I focused on the discovery of somatic SVs in the samples. MEFs were freshly isolated from normal embryos, while mESCs had been in culture for some time before. The differences in genomic instability of these two cell types were represented by the frequency of SVs. As expected, MEFs had a very low SV burden, while mESCs were characterized by a higher number of chromosomal abnormalities, including previously reported trisomies 8 and 11 that give a growth advantage[172,173]. The majority of the detected SVs were aneuploidies which are likely a consequence of mitotic errors[187]. Even though changes in the chromosome copy number may be detrimental to cell fitness[188], it has been shown that normal cells in human bodies also experience aneuploidies and the fluctuations in ploidy are common in cells growing in culture[187].

8.5 Increased karyotypic heterogeneity induced by doxorubicin

Strand-seq libraries from mammary gland organoids were prepared at different time points: in the normal state before the induction with doxycycline, after the doxycycline-induced overexpression of HER2 and CMYC, including after the treatment with two different concentrations of doxorubicin (10 nM and 100 nM). As shown with the results of scRNA-seq, doxorubicin induces cell-cycle arrest, and a recovery period in a drug-free medium was needed after the exposure to the drug until the cells started proliferating again (a prerequisite for Strand-seq). Overall, I analyzed more than 400 single cells from mammary gland organoids at different stages of tumorigenesis. For every single library, I listed SVs (if present) and assigned the most probable cell type based on the predictions of scNOVA-based classifier.

In all the samples, apart from the never-induced condition, cells had at least one SV but the frequency differed, with doxorubicin-treated cells showing the highest frequency (with 37% of cells with at least one SV in their genome). Deletions of various sizes were the most frequent type of SVs detected after doxorubicin treatment and complex rearrangements occurred exclusively in cells exposed to the drug. In the cells treated with the higher concentration of the drug, the frequency of SVs was higher compared to the treatment with 10 nM doxorubicin. All the events that I detected were singletons, present only in one cell indicating that the genetic heterogeneity within the population increased. This suggests that new clones, potentially with growth advantage, may have a chance to expand. At the same time, extremely high levels of genomic instability, either intrinsic or drug-induced, may have an adverse effect on the fitness of cancer cells and prevent them from further replication possibly leading to dormancy. The results of scRNA-seq experiment showed that after a single drug treatment, the cells did not have the phenotype of senescence or dormancy. It can be speculated that multiple rounds of therapy or longer exposure to the drug would be necessary to induce such extreme reactions.

For cell-type-specific SV discovery, we applied the scNOVA-based classifier to all Strand-seq libraries from murine mammary gland organoids and annotated the most likely cell type (LP, ML, basal) to each cell. Interestingly, after doxorubicin treatment, all three different cell types were similarly affected by SVs and up to 40% of cells from each cell type had at least one SV. Such an observation was surprising, as the results of both scRNA-seq and Strand-seq indicate that both genomic and transcriptomic responses were largely similar across cell types, and yet basal cells were more sensitive to the cytotoxic effect of doxorubicin. Potential explanations are discussed in section 8.6.

With Strand-seq data I was also able to correlate doxorubicin treatment with a higher frequency of SCE events in all three cell types, an additional indicator of genomic stress and instability. For the purpose of this project, I updated the coordinates of common fragile sites for the mm10 reference genome as they have been annotated only in the previous reference assemblies. Neither SVs nor SCEs induced by doxorubicin were enriched within CFS. As the expression of CFSs is cell-type-dependent[174], it might be that the loci previously reported as murine CFSs are not breakage-prone in mammary cells. It has to be noted that BrdU is among a few chemical agents that are known to be capable of inducing CFS breakages[189]. However, BrdU is associated only with rare fragile sites in the human genome[177], and the impact of BrdU on the integrity of the genome was thoroughly tested when the Strand-seq protocol was being established. It was also demonstrated that variable BrdU concentrations in the cell culture medium (10-200 μM) have no effect on the frequency of SCEs in normal human cells (fibroblasts and lymphoblasts) and in cells from patients with Bloom syndrome, which is characterized by high levels of genomic instability and SCEs[190].

8.6 Lineage-specific differential sensitivity to doxorubicin

As summarized in the previous subchapter, all three different cell types present in the murine mammary gland organoids experienced similar levels of genomic damage (measured by the frequency of SVs and SCEs). Such observation would indicate that distinct cell types present in mammary glands have a comparable capacity for DNA repair. Different organs or tissues are exposed to different mutagens leading to various types of DNA damage. Therefore, depending on the type of the DNA lesion, particular DNA repair pathways might be involved to fix the damage[191]. However, cell-type-specific differences in DNA repair strategies within an organ are understudied. In a recent report, the authors showed that mammary epithelial lineage influences the choice of DNA repair pathway after DNA DSBs induced by irradiation[192]. According to their data, all mammary epithelial cells are capable of non-homologous end joining but homologous recombination is predominant only in luminal cells. Such claims are not fully consistent with the results presented in this thesis. Certain deficiencies in DNA repair of basal cells which would lead to cell death if the DSBs overload is too high may explain why basal cells are more sensitive to doxorubicin (as seen in the scRNA-seq experiment) but on the contrary, the basal cells show the same frequency of SCEs, resolved by HR, as luminal cells (according to Strand-seq data). Both irradiation and doxorubicin induce DSBs in DNA and are associated with the generation of ROS (which may lead to DNA base damage), but irradiation causes also single-strand breaks[193] that have not been reported for doxorubicin. More detailed research would be needed to determine if the choice of DNA repair pathway is more influenced by the type of damage rather than dictated by the cell lineage in the mammary gland. I hypothesize that the toxic effect of doxorubicin on basal cells could be connected with the faulty metabolism of the drug in this particular cell type. Cbr3, an enzyme that catalyzes the conversion of doxorubicin to toxic metabolites, is downregulated upon doxorubicin treatment in luminal cells but not basal. Unsuccessful detoxification of a drug and accumulation of toxic metabolites leading to cell death may explain the lineage-specific sensitivity to doxorubicin.

8.7 Outlook

Taken together, in this thesis I showed that treatment with doxorubicin induces SVs and promotes genomic heterogeneity and that the overall genomic and transcriptomic response to the drug is shared by the three major cell types of mammary lineage. To my knowledge, this is the first study in which chemotherapy-induced mutational signatures were systematically analyzed beyond base substitutions and indels. The results indicate that deletions and complex genomic rearrangements emerge as potential mutational signatures of doxorubicin. To strengthen this claim, SV patterns should be ideally validated with the statistical analysis of cancer genomes in doxorubicin-treated patients. Such analysis may have a true clinical relevance as the concentration of doxorubicin used in this study falls in the range of values detected in the blood of breast cancer patients undergoing chemotherapy [194,195].

Beyond biological significance, one of the key outcomes of this study is expanding the utility of Strand-seq and scNOVA. For the first time in the mammary gland field, I was able to couple the detection of structural variants in single cells while preserving the information about their cell type. There is still more to explore with scNOVA as one of its functionalities allows, based on Strand-seq data, to predict gene expression differences between specified cell populations. This module of scNOVA was previously used to identify and characterize the functional consequences of dysregulated genes in subclones bearing different SVs present in lymphoblastoid cell lines and patient-derived leukemia samples[114]. In this project, we created this functionality for mice genome and used it to infer genes with differential NO in cells before and after induction with doxycycline. In the future, scNOVA could be applied to look for NO changes between cells treated with doxorubicin (also considering the annotated cell type) and their corresponding controls, and the obtained results could be integrated with the already created scRNA-seq dataset. However, in the context of this project, it would not be possible to dissect the functional effect of doxorubicin-induced SVs purely using Strand-seq data as all the events were singletons and to apply scNOVA one would need at least two single cells that share a common SV.

The results included in this thesis highlight that doxorubicin, a drug widely used in the clinic for many years, may promote the formation of new mutations in the tumor. For many cancer patients, the benefits of receiving the drug will outweigh the side effects and potential mutagenic risks. However, as cancer patients live longer thanks to therapeutic advancements, it will be necessary in the future to focus on the effect of chemotherapeutics on healthy cells and the long-term consequences of chemotherapy. By using Strand-seq, a method that allows the detection of a wide variety of DNA rearrangement processes, in organoids, it will now be possible to study such potentially harmful consequences of cancer therapy.

Chapter 9 Materials and methods

If not indicated otherwise, all buffers were prepared by Media Kitchen at EMBL Heidelberg.

9.1 Animals and mouse cell culture

9.1.1 Animals

Mouse colonies used in this study (strain TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB background, and FVB/NJ strain[102]) were bred and maintained in LAR (Laboratory Animal Resources) facility at EMBL Heidelberg, under veterinarian supervision and in accordance to the guidelines of the European Commission, revised Directive 2010/63/EU and AVMA Guidelines 2007. Animals were kept on a 12-hour light/12-hour dark cycle, with constant ambient temperature (23±1°C) and humidity (60±8%), and pellet food and water were provided *ad libitum*.

9.1.2 Genotyping

The correct genotype of tri-transgenic mice was determined by PCR on genomic DNA from the tail tissue. Genomic DNA was extracted by digestion of the tail in 75 µl of digestion buffer (NaOH 25 mM + EDTA 0.2 mM) at 98°C followed by neutralization with 75 µl Tris-HCl (40 mM, pH 5.5) and centrifugation at 4000 rpm for 3 min. 2 µl of the supernatant was used in the PCR reaction summarized in the Table 1. Primer sequences and PCR programs are included in Table 2 and Table 3, respectively. Agarose gel electrophoresis (1.5% agarose gel, 100 V, 60 minutes using 1 kb or 100 bp DNA ladders (Thermo Fisher Scientific) as size markers) was used to detect PCR products of expected sizes (TetO-Myc: 630 bp, TetO-Neu: 386 bp, MMTV-rtTA: 380 bp).

Table 1 Reaction mix for genotyping.

Component	**Stock**	**Final concentration**
Forward primer	10 µM	0.25 µM
Reverse primer	10 µM	0.25 µM
dNTP mix	10 mM each	200 µM each
Taq polymerase	x	1 µl per 50 µl reaction
DreamTaq™ buffer (Thermo Fisher Scientific)	10x	1x

Taq polymerase was produced by the Protein Expression Facility at EMBL Heidelberg.

Table 2 Primer sequences for genotyping.

Transgene	Forward primer (5'-3')	Reverse primer (5'-3')
TetO-CMYC	TAGTGAACCGTCAGATCGCCTG	TTTGATGAAGGTCTCGTCGTCC
TetO-Neu	GACTCTCTCTCCTGCGAAGAATGG	CCTCACATTGCCAAAAGACGG
MMTV-rtTA	GTGAAGTGGGTCCGCGTACAG	GTACTCGTCAATTCCAAGGGCATCG

Table 3 PCR programs for genotyping.

Transgene	*TetO-CMYC*	*TetO-Neu*	*MMTV-rtTA*
PCR program	94ºC- 3 min 10x: 94ºC- 10s 62ºC- 30s 68ºC- 1:30 min 28x: 94ºC- 10s 60ºC- 30s 68ºC- 2 min 72ºC- 10 min 10ºC- hold	95ºC- 1 min 2x: 95ºC- 15 s 64ºC- 15 s 72ºC- 1:30 min 2x: 95ºC- 15 s 61ºC- 15 s 72ºC- 1:30 min 23x: 95ºC- 15 s 58ºC- 15 s 72ºC- 1:30 min 18x: 95ºC- 15 s 55ºC- 15 s 72ºC- 1:30 min 72ºC- 10 min 10ºC- hold	94ºC- 5 min 39x: 94ºC- 30 s 57ºC- 30 s 72ºC- 30 s 72ºC- 5 min 10ºC- hold

9.1.3 3D culture of mouse mammary gland organoids

6- to 10-week-old female virgin mice were euthanized by CO_2 overdose or cervical dislocation, and all mammary glands were harvested. The tissue was digested overnight (max. 16 hours) in 5 ml of digestion medium (DMEM/F12 (Lonza) with 25 mM HEPES, supplemented with 1% penicillin/streptomycin (Thermo Fisher Scientific), 150 U Collagenase type 3 (Worthington Biochemical Corporation) and 20 µg Liberase (Roche)) at 37°C and 5% CO_2, in a loosely capped 50 mL polypropylene conical tube. The glands were then mechanically disrupted by pipetting with a 5 ml pipette, washed with phosphate-buffered saline (PBS), and centrifuged at 300xg for 5 minutes. The layer of fat and medium with PBS was removed, and 5 ml of 0.25% Trypsin-EDTA (Thermo Fisher Scientific) was added to the cell pellet. After incubation for 45 minutes at 37°C and 5% CO_2 the enzymatic reaction was stopped by the addition of 40 ml of serum-supplemented media (DMEM/F12 (Thermo Fisher Scientific) with 25 mM HEPES, 1% penicillin/streptomycin (Thermo Fisher Scientific), and 10% fetal bovine serum (FBS) (Thermo Fisher Scientific)). The cells were centrifuged again, the pellet was resuspended in mammary epithelial cell basal medium (MEBM) (PromoCell) enriched with MEpiCGS (ScienCell Research Laboratories), and the cell suspension was transferred to collagen-coated plates (Corning). During overnight incubation at 37ºC and 5% CO_2, epithelial cells adhere to the surface of the plate while cells of other cell types and dead cells float in the medium. On the following day, the medium with non-epithelium cells was removed, and after washing in PBS and trypsinization (incubation with 0.25% Trypsin-EDTA for 5-7 minutes at 37ºC and 5% CO_2, followed by inactivation with serum-supplemented media), the epithelial cells were detached from collagen-coated plates. The cells were centrifuged, resuspended in MEBM with MEpiCGS, and counted. If not indicated otherwise, 10,000 cells were seeded per 1 well of a 12-well plate in 90 µl of ice-cold solution 4:1 of Matrigel (Corning): PBS. Matrigel-PBS-cell solution droplets were dispensed into the bottom of wells and incubated for 30-45 minutes at 37ºC and 5% CO_2 until the mixture solidified. Then 1.5 ml of MEBM with MEpiCGS was added to each well, and the organoids were let to grow at 37ºC and 5% CO_2. For drug screen, BrdU, and DMSO titration experiments, primary murine mammary gland epithelial cells were isolated as described above but they were seeded on 96-well plates (Falcon® 96-well Black/Clear Flat Bottom TC-treated Imaging Microplate with Lid, Corning) (600 cells in 10 µl gels) and incubated with 200 µl MEBM with MEpiCGS. The medium was changed one day after seeding, and then every second day (every 3 days in case the organoids were growing in a 96-well plate). If required by experiment conditions, the organoids were induced with 200 ng/ml of doxycycline (doxycycline hyclate, Sigma) 7 days after seeding. 200 ng/ml of doxycycline was then always added to the medium.

9.1.4 Dissociation of 3D structures

Gels containing organoids (growing in single wells of 12-well plates) were incubated for 2 hours at 37°C with 75 U Collagenase type III (Worthington Biochemical Corporation) and 10 µg Liberase (Roche), and then disintegrated completely by mechanical disruption with pipetting up and down with a 1000 µl pipette. The suspension from each well was transferred to its 15 ml falcon and centrifuged at 300xg for 5 minutes. After washing once in pre-warmed PBS and removing any leftovers of Matrigel that remained above the cell pellets, the pellets were resuspended in 200 µl of 0.25% Trypsin-EDTA (Thermo Fisher Scientific) and incubated for 5 minutes in 37ºC water bath. Trypsin was then deactivated with 5 ml medium containing

DMEM/F12 (Thermo Fisher Scientific) with 25 mM HEPES, 1% penicillin/streptomycin (Thermo Fisher Scientific), and 10% FBS (Thermo Fisher Scientific). Suspensions of single cells were centrifuged at 300xg for 5 minutes, and the pellets were washed once in PBS. For both scRNA-seq and scATAC-seq all gels seeded in a 12-well plate were collected and pooled, for Strand-seq two gels from a 12-well plate containing organoids pulsed with BrdU at the same time were collected and pooled.

9.1.5 Isolation and culture of MEFs

Fibroblasts were isolated from 13.5-day FVB/NJ mouse embryos. A pregnant female was euthanized by cervical dislocation and the uterus was dissected. The yolk sac was opened and individual fetuses were exposed. Head, liver, and heart were removed from each embryo. The remainings were then cut into fine pieces, and all material was transferred into 15 ml falcon tubes. The tissue from up to 5 embryos coming from one mother was pooled into one falcon, and incubated overnight at 4ºC in 15 ml of ice-cold 0.25% trypsin-EDTA (Thermo Fisher Scientific). On the next day, most of the trypsin solution was aspirated (leaving an amount equivalent to approximately two volumes of the tissue) and the tube was incubated for 30 minutes in a 37ºC water bath. 25 ml of MEF culture medium (high-glucose DMEM (Thermo Fisher Scientific), with 10% FBS (Thermo Fisher Scientific) 1% penicillin/streptomycin (Thermo Fisher Scientific)) was added to the tube, and the digested tissue was broken up into a cell suspension by vigorously pipetting the solution up and down. More MEF culture medium was added, and cells were plated in T75 tissue culture flasks. Cells were maintained at 37ºC with 5% CO_2. MEFs were passaged every second day and cells from up to the first 5 passages were used for experiments. For RNA-seq and Strand-seq, cells were seeded at a density of $1x10^5$ cells/well in a 6-well plate, and for both experiments they were collected on the same day, 3 days after seeding. For RNA-seq cells from two wells were pooled as one technical replicate and two technical replicates were submitted for sequencing. For Strand-seq MEFs were incubated with 40 µM BrdU (Sigma, B5002) for 18h before nuclei isolation and single-nuclei sorting.

9.1.6 Culture of mESCs

A plate of mESCs (129 x C57BL/6J) was a gift from Noh lab, EMBL Heidelberg. To reduce the risk of contamination with feeder cells, mESCs were transferred into feeder-free gelatin-coated plates (0.1% sterile gelatin solution) one passage before seeding for experiments. Gelatin coating was performed on all plates used for experiments with mESCs. The medium, containing KnockOut DMEM (Thermo Fisher Scientific) with 15% EmbryoMax FBS (Merck Millipore) and 20 ng/ml leukemia inhibitory factor (produced by the Protein Expression Facility at EMBL Heidelberg), 1% nonessential amino acids (Thermo Fisher Scientific), 1% GlutaMAX (Thermo Fisher Scientific), 1% penicillin/streptomycin (Thermo Fisher Scientific) and 1% of 55 mM β-mercaptoethanol solution (Sigma), was changed every day. Cells were maintained at 37ºC with 5% CO_2. For RNA-seq and Strand-seq, cells were seeded at a density of $1.6x10^5$ cells/well in a 6-well plate, and for both experiments they were collected on the same day, 2 days after seeding. For RNA-seq cells from two wells were pooled as one technical replicate and two technical replicates were submitted for sequencing. For Strand-seq mESCs were incubated with 20 µM BrdU (Sigma, B5002) for 13h before nuclei isolation and single-nuclei sorting.

9.2 Human cell culture

Human hTERT-immortalized retinal pigment epithelial cells RPE-1 were purchased from ATCC, and cultured in DMEM/F12 medium (Thermo Fisher Scientific) supplemented with 10% FBS (Thermo Fisher Scientific).

Human mammary gland cell line MCF10a (purchased from ATCC) were cultured in DMEM/F12 medium without HEPES and phenol red (Thermo Fisher Scientific) supplemented with 5% horse serum (Thermo Fisher Scientific), 20 ng/ml EGF (Biotrend), 0.5 µg/ml hydrocortisone (Sigma), 100 ng/ml cholera toxin (Sigma) and 0.01 mg/ml human insulin (Sigma).

Human invasive ductal carcinoma cell line BT-474 was a gift from Dr. Matt Boucher, Jechlinger group, EMBL Heidelberg. The cells were cultured in high glucose DMEM (Thermo Fisher Scientific) supplemented with 10% FCS (Thermo Fisher Scientific), 1% non-essential amino acids (Thermo Fisher Scientific), 1% sodium puryvate (Thermo Fisher Scientific), 1% L-glutamine (Thermo Fisher Scientific), 10 mM HEPES.

All cell lines were cultured in standard conditions (37ºC, 5% CO_2).

9.3 Cytotoxicity analysis

9.3.1 Drug screen on murine mammary gland organoids

To test the effect of drugs on murine mammary gland organoids, different chemotherapeutics were added to the media of both not-induced organoids and structures induced with doxycycline for 7 days (after 7 days of normal growth). Each drug was tested at five different concentrations with four to five technical replicates for 72 hours. Depending which solvent was used to dissolve the drugs, cells were treated with either DMSO (Sigma) or water as a control. Following the incubation with the drugs, both cytotoxicity and cell viability assays were performed, and then an IC50 value for each of the drugs was calculated.

The following drugs were included in the screen: lapatinib (GW-572016) ditosylate (Selleckchem), paclitaxel (NSC 1259733, Selleckchem), InSolution Paclitaxel (Sigma), doxorubicin-hydrochlorid (D1515, Sigma).

9.3.1.1 Bright-field imaging

To visually assess the growth of organoids and the impact of drugs on their morphology, bright-field imaging was performed at different time points (on day 1 after seeding, on day 8 before doxycycline induction, on day 15 before drug treatment and on day 18 (72 hours after drug treatment) at the end of the experiment). All 96-well plates used in the drug screen were imaged on the ScanR (High Content Screening Station, Olympus) with 4x objective, in 4 quadrant fields of view per well. To acquire the data from the entire gel drop with organoids, 21 images were recorded at z-steps intervals of 100 μm in each four fields of view. Using custom-made macros and scripts (provided by Dr. Sylwia Gawrzak and Dr. Matt Boucher, Jechlinger group, EMBL Heidelberg) with modifications, the z-stacks were processed to show a maximum projection image of each field of view, and then the four projections were joined together to represent each well.

9.3.1.2 Cell cytotoxicity and cell viability assays

The cytotoxic effect of drugs was quantified using a fluorescence-based commercially available kit CellTox^{TM} Green Cytotoxicity Assay (Promega) that measures changes in membrane integrity following cell death. After 72 hours of drug treatment, 100 μl of media was removed and 20 μl of the working solution was added to each well. The plate was then gently shaken at room temperature for 1 hour, and after the incubation, the fluorescent signal (at an excitation wavelength of 485-500 nm and an emission of 520-530 nm) proportional to cell death in each well was measured at the Chemical Core Facility at EMBL Heidelberg, using an EnVisionTM Multilabel Plate Reader (PerkinElmer).

Cell cytotoxicity measurement was multiplexed with a luminescence-based cell viability assay (CellTiter-Glo® 3D Cell Viability Assay, Promega) that quantifies the ATP present, an indicator of metabolically active cells. Once the data from cell cytotoxicity was recorded, 75 μl of CellTiter-Glo® 3D Cell Viability Assay was added to the media per well and the whole plate was again gently shaken at room temperature for 1 hour. The luminescence signal was recorded using the Infinite® M1000 microplate reader (Tecan). The raw data from the plate reader was analyzed using GraphPad6 (Prism) to generate drug-response curves and calculate IC50 values for each drug and assay.

9.3.2 BrdU and DMSO titration experiments

Similarly to the drug screen, mammary gland organoids were seeded on 96-well plates with 600 cells in 10 µl in each well. The organoids were grown for 72 hours unperturbed (apart from a change of medium one day after seeding), and then they were treated with increasing concentrations of BrdU (Sigma, B5002) or DMSO (Sigma) for the next 72 hours. Five different concentrations of each substance were tested (with three technical replicates per condition). Organoids not exposed to BrdU or DMSO were used as a control. Cell viability assay alone (without cell cytotoxicity assay) was performed as the endpoint of the BrdU and DMSO titration experiment.

9.3.3 Drug screens on human cell lines

MCF10a and BT-474 cells were seeded on 96-well plates (Falcon® 96-well Black/Clear Flat Bottom TC-treated Imaging Microplate with Lid, Corning), 1×10^4 cells per well. One day later they were treated with 4 different concentrations of either doxorubicin, lapatinib, or paclitaxel with 3 technical replicates for each tested drug and dilution. Cells treated with DMSO (solvent for all the drugs) were used as a control. Cell viability assay (CellTiter-Glo® Luminescent Cell Viability Assay, Promega) was performed after 72 hours of drug exposure, and additionally after 96 hours and 120 hours of paclitaxel treatment. Similarly to the assay used for the readout in the drug screens on organoids, this kit quantifies the ATP present in the cells, proportional to the number of cells alive. 75 µl of the reagent was added to the media in each well and the whole plate was gently shaken at room temperature for 20 minutes. The luminescence signal was recorded using the Infinite® M1000 microplate reader (Tecan) and the raw data were analyzed using Excel to quantify the proportion of alive cells following drug treatment in comparison to DMSO control.

9.4 Molecular biology methods

9.4.1 RT-qPCR (murine mammary gland organoids)

Total RNA was extracted from mammary gland organoids derived from tri-transgenic mice (TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB/NJ background) at three different time points: 6 days after seeding, after 7 days of doxycycline induction and after 7 days after the deinduction (on the day of the deinduction the gels were washed for 10 minutes in PBS, once for 10 minutes in STOP media (DMEM/F12 (Thermo Fisher Scientific) with 25mM HEPES, 1% penicillin/streptomycin (Thermo Fisher Scientific)), and again for 10 minutes in PBS; after that, they were cultured in the MEBM with MEpiCGS but without doxycycline). Organoids from two wells were pooled for each condition for each biological replicate. After the dissociation of 3D structures (as described above), the pellets containing single cells were used as a starting material for the RNA isolation performed with the kit innuPREP DNA/RNA Mini Kit (Analytik Jena) according to the manufacturer's protocol. RNA concentration was measured using the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific). Complementary DNA (cDNA) was prepared from 50 ng of RNA using SuperScript™ III First-Strand-Synthesis SuperMix for qRT-PCR (Invitrogen) in 20 µl. The synthesized cDNA was diluted 4x, and 2 µl were used for qPCR reaction (with 10 µM primers) on a StepOne™ Real-Time PCR System (Thermo Fisher Scientific). Primer sequences are shown in Table 4. Three technical replicates were included per each analyzed gene, condition, and biological replicate. Standard 'no template' and 'no reverse transcriptase' controls were always performed. Agarose gel electrophoresis (2% agarose gel, 100 V, 30-45 minutes using 1 kb or 100 bp DNA ladders (Thermo Fisher Scientific) as size markers) was performed to confirm the presence of single amplicons of the correct size. Beta-actin (*Actb*) was used as a reference. The fold changes in gene expression were calculated using the $\Delta\Delta C_t$ method, and the results are presented as fold change of values obtained for never-induced organoids.

Table 4 Primer sequences for RT-qPCR.

Target	**Forward primer (5'-3')**	**Reverse primer (5'-3')**
mouse *Erbb2*	GAGACAGAGCTAAGGAAGCTGA	ACGGGGATTTTCACGTTCTCC
rat *TetO-Neu*	GAATCCCTGCTGGGGCACC	CAGTGCCTGGGGTAGGGTCC
mouse *Actb*	AGAGCTACGAGCTGCCTGAC	AGCACTGTGTTGGCGTACAG

9.4.2 RNA-seq (MEFs and mESCs for scNOVA)

Before RNA isolation, MEFs and mESCs growing in 6-well plates were washed once in PBS and trypsinized for 5 minutes (incubation with 0.25% Trypsin-EDTA at 37ºC and 5% CO_2). The cells were then resuspended in their respective medium and centrifuged (5 minutes, 1200 rpm). The cell pellets were washed once in PBS and the solutions were centrifuged again. The cell pellets were used as the starting material for RNA extraction using RNeasy Mini Kit (QIAGEN). Once the pellets were resuspended in the lysis buffer, the cells were additionally disrupted mechanically by passing the solutions through needles with syringes (20G) (each sample 10 times). After elution, RNA concentrations were measured with the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific) and high-quality RNA was submitted to Genomics Core Facility at EMBL Heidelberg for library preparation. All 4 libraries (2 technical replicates for MEFs and 2 for mESCs) were multiplexed and sequenced together on a NextSeq 500 High sequencer (75 single-ends).

9.4.3 RNA-seq (murine mammary gland organoids for validation of scNOVA)

A table containing raw gene counts from RNA-seq experiments performed on murine mammary gland organoids was downloaded from Array Express under accession number E-MTAB-8834. The differential expression analysis was performed with DESeq2[159] following the authors' manual. Genes with fewer than 10 counts across all samples were filtered out. The animal and condition were used for multifactor design. Genes with adjusted *p*-values (after a Bonferroni correction for multiple testing) smaller than 0.1 were considered significantly differentially expressed.

9.4.4 Isolation of genomic DNA and WGS

Mammary gland organoids (tissue derived from TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB/NJ background mice) were cultured for 8 days and then dissociated as described before. Genomic DNA from all organoids growing in one plate was isolated using QIAAmp DNA Mini and Blood Mini kit (QIAGEN) according to the manufacturer's protocol (including a recommended additional step of RNase and protease treatment). DNA concentration was measured using the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific) and with a Qubit™ 3.0 Fluorometer following the staining with Qubit™ dsDNA HS-Assay Kit (Thermo Fisher Scientific). DNA was submitted to Genomics Core Facility at EMBL Heidelberg for library preparation. The completed library was sequenced on a HiSeq4000 sequencer (150 paired ends).

9.4.5 Analysis of bulk WGS data

Raw whole genome sequencing data (in fastq format) was aligned to the mice reference genome mm10, sorted, marked for duplicates, and indexed. The high quality of data was confirmed with Alfred, a command-line application that computes quality control metrics (for example, GC bias, base composition, insert size, and sequencing coverage distributions). The presence of each germline SV called by Strand-seq was manually verified in WGS data by checking the differences in the read count around the predicted SV breakpoints in the Integrative Genomics Viewer (IGV) (version 2.13.2)[171].

9.5 Immunofluorescence

9.5.1 Anti-BrdU staining on RPE-1 cells

Before cell seeding, cover glasses were immersed in 70% EtOH and incubated for 1 hour under UV light in a laminar flow hood. Each sterile cover glass was then placed at the bottom of a well of a 6-well plate immediately after plating the cells. The RPE-1 cells were seeded at the density of 1.5x10^5 cells/well. One day later, the cells were treated with 40 µM BrdU (Sigma, B5002) (excluding one well) and incubated for 24 h. After that time, the medium was removed from both BrdU-containing and BrdU-negative wells, and the cover glasses were washed twice with PBS. The cells were fixed with 4% paraformaldehyde (PFA) (Electron Microscopy Sciences) for 10 minutes and washed twice with PBS. To permeabilize the cell membranes, the cells were treated with 0.2% Triton X-100 in PBS for 15 minutes at room temperature and washed again twice with PBS. Depending on the experimental conditions, the cells were treated with 2 M HCl for 5, 15, or 30 minutes, and then incubated for 30 minutes in a neutralization buffer (boric acid/potassium chloride/sodium hydroxide, pH 9.0) (Merck). Cells treated with PBS (without treatment with HCl and the neutralization buffer) were used as a control. All cover glasses were transferred to a humidified chamber for downstream steps. Blocking was performed in 10% goat serum (Merck) in 0.2% Triton X-100 in PBS for 30 minutes at room temperature. After washing with PBS, the cover glasses were incubated with 1:250 dilution of anti-BrdU antibody (ab6326) (Abcam) for 1 hour at room temperature. The cells were washed three times with PBS, and incubated with 1:1000 dilution of anti-rat IgG Alexa Fluor 488-conjugated antibody in 10% goat serum in 0.2% Triton X-100 in PBS, for 1 hour at room temperature. The cells were washed three times with PBS before DNA staining with Hoechst 33258 (1:10000) in 10% goat serum in 0.2% Triton X-100 in PBS, for 20 minutes at room temperature. The cells were washed once in PBS and once in the ultrapure water (Milli-Q® EQ 7000 quality) before fixing with Vectashield® Antifade Mounting Medium (Vector Laboratories) on microscopy slides. The samples were imaged at 6 random locations per tested condition, using the same settings, with Zeiss Cell Observer HS (Zeiss) fluorescence microscope. Fiji[196] was used for image processing.

9.5.2 Cryosections

3D culture gels for Click-iT EdU imagining or immunofluorescence staining were collected from the wells of the plate and transferred to Tissue-Tek® Cryomold® moulds (Sakura) filled with Tissue-Tek® O.C.T.™ Compound (Sakura). The moulds were stored on dry ice for 10 minutes until the matrix solidified and then transferred to -80ºC for long-term storage. The samples were then cut on a cryostat to obtain 8 µm sections (2-3 sections per slide), and the slides with sections were processed within a week (stored at -80ºC in the meantime).

9.5.3 Click-iT EdU staining on cryosections (murine mammary gland organoids)

Mammary gland organoids derived from tri-transgenic mice (TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB/NJ background) were cultured for 7 days and then induced with doxycycline. After 3 or 4 days they were treated with 20 µM EdU for 48 h or 24 h or left untreated before collecting and processing for cryosections as described above. After thawing, the slides with sections were washed once in PBS, and then the cells were fixed with 4% PFA (Electron Microscopy Sciences) for 10 minutes and washed once with PBS. The cell membranes were permeabilized with 0.2% Triton X-100 in PBS for 10 minutes at room temperature. The slides were washed three times with PBS. A label mix containing PBS, 2 mM $CuSO_4 \cdot 5H_2O$ (Sigma), 20 mg/ml ascorbic acid (Sigma), and 8 µM sulfo-Cy3-azide (Lumiprobe) was prepared freshly and added to the cells. After 30 minutes of incubation, the cells were washed three times with PBS and then stained with 1:1000 DAPI (Thermo Fischer Scientific) for 20 minutes. After one wash in PBS and then water, cryosections were mounted with Vectashield® Antifade Mounting Medium (Vector Laboratories) and protected with a cover slip. The images were collected with a confocal microscope Zeiss LSM 900 (Zeiss). All image processing was performed with Fiji[196].

9.5.4 Anti-keratin staining on cryosections (murine mammary gland organoids)

Mammary gland organoids derived from tri-transgenic mice (TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB/NJ background) were cultured for 7 days and then induced with doxycycline for 5 days. The gels containing 3D structures were collected and processed for cryosections as described above. After thawing, the slides with sections were transferred to a humidified chamber and were washed once in PBS. Then the cells were fixed with 4% PFA (Electron Microscopy Sciences) for 10 minutes and washed twice with PBS. The cell membranes were permeabilized with 0.2% Triton X-100 in PBS for 15 minutes at room temperature. The slides were washed twice with PBS and incubated with a blocking solution (10% goat serum (Merck) in 0.2% Triton X-100) at room temperature. After 30 minutes, the blocking agent was removed and a 1:200 dilution of primary antibodies (anti-keratin 14 (ab7800) and anti-keratin 19 (ab52625), both from Abcam) in the blocking solution was added to the sections. One slide was incubated without the primary antibodies as a negative control. After 1 hour, the slides were washed three times with PBS and all of them were stained with 1:1000 dilution of secondary antibodies (Donkey anti-Mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 555, Donkey anti-Rabbit IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 647, both from Thermo Fisher Scientific) in the blocking solution for 1 hour. Then the slides were washed three times with PBS and stained with a 1:5000 dilution of Hoechst 33258 dye. Finally, the slides were washed once in PBS, and once in the water. The cryosections were mounted with Vectashield® Antifade Mounting Medium (Vector Laboratories) and protected with a cover slip. The images were collected with a confocal microscope Zeiss LSM 900 (Zeiss). All image processing was performed with Fiji[196].

9.6 Flow cytometry

9.6.1 Click-iT EdU flow cytometry labeling

Mammary gland organoids derived from tri-transgenic mice (TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB/NJ background) were cultured for 7 days and then induced with doxycycline. After 3 or 4 days they were treated with 20 µM EdU for 48 h or 24 h or left untreated before dissociating into single cells as described above. Organoids used for Click-iT microscopy staining and flow cytometry labeling were seeded and collected on the same day. The suspension of single cells isolated from the mammary gland organoids was stained using Click-iT® EdU Flow Cytometry Assay Kit (with Alexa Fluor® 488 azide; Thermo Fisher Scientific) and following manufacturer's protocol. The frequency of EdU-positive cells was recorded using flow cytometer LSR Fortessa Analyser (BD Biosciences) in the EMBL Flow Cytometry Core Facility, and the raw data was analyzed using FlowJo™ v10 software (BD Biosciences).

9.6.2 Detection of BrdU-positive nuclei with flow cytometry

To isolate nuclei, the pellets containing PBS-washed single cells (from murine mammary gland organoids, MEFs, or mESCs), both BrdU-pulsed and BrdU-negative, were resuspended in 300 µl nuclei isolation buffer each (the composition summarized in Table 5) and incubated on ice for 20 minutes in dark. Nuclei were then analyzed using BD FACS Melody™ Cell Sorter (BD Biosciences) or BD LSRFortessa™ Cell Analyzer (BD Biosciences). After removing debris (based on forward and side scatter) and doublets (based on size), the nuclei from the negative control were used to establish the Hoechst profile and identify the G1 peak. The sorting gate for BrdU-positive cells was set on the peak showing half Hoechst fluorescence compared to no-BrdU control.

Table 5 Composition of the nuclei staining buffer (NSB)

Stock solution	Final concentration	For 10 ml of NSB
1 M Tris-HCl pH 7.4	100 mM	1 ml
5 M NaCl	154 mM	308 µl
1 M $MgCl_2$	0.5 mM	5 µl
1 M $CaCl_2$	1 mM	10 µl
7.5% BSA solution	0.20%	267 µl
10% NP-40 substitute	0.1%	100 µl
dH_2O	-	8.3 ml
10 mg/ml Hoechst 33258	10 µg/ml	10 µl

NP-40 subsitute was purchased from MillporeSigma.

9.7 Single-cell sequencing technologies

9.7.1 scRNA-seq

9.7.1.1 scRNA-seq library preparation

Mammary gland organoids derived from tri-transgenic mice (TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB/NJ background) were cultured for 7 days and then induced with doxycycline. After 6 days they were treated for 72 hours with 100 nM doxorubicin, 0.5% DMSO or left untreated, and then the medium was changed and the organoids were left to recover for the next 72 hours. For each sample, the organoids were extracted from Matrigel and dissociated as described before. The pellet of single cells was resuspended in 400 µl of filtered solution of 0.4% BSA in PBS. The cells were stained with DRAQ7™ dye (Thermo Fisher Scientific) and IncuCyte® Caspase-3/7 Green Apoptosis Assay Reagent (IncuCyte) to detect dead and early-apoptotic cells, and the double-negative population was immediately sorted to 15 ml falcons containing 1 ml of 0.4% BSA in PBS at the bottom. Depending on the sample, up to 1x10^6 cells were collected. After sorting the cells were immediately counted and then resuspended in app. 50 µl of 0.4% BSA in PBS. The cells were counted again and then a 100 µl solution of cells was prepared with a desired concentration of 1000 cells/µl. All counting steps were performed using a Countess II FL Automated Cell Counter (Thermo Fisher Scientific). scRNA-seq libraries for each sample were then prepared following the user guide of 10x Genomics Chromium Next GEM Single Cell 3' Reagent Kit v3.1 (dual index) with the expected sequencing depth of 20,000 read pairs per cell and 8,000-10,000 cells per sample. 1:10 dilution of each constructed library was run on the Agilent Bioanalyzer High Sensitivity DNA chip (Agilent) to confirm their high-quality and assess the concentration. The concentration of libraries was additionally measured with a Qubit™ 3.0 Fluorometer following the staining with Qubit™ dsDNA HS-Assay Kit (Thermo Fisher Scientific). Completed libraries were pooled and sequenced on NextSeq 2000 sequencer (using P3 reagent).

9.7.1.2 scRNA-seq data processing

The Cell Ranger Single Cell Software (10x Genomics), version 6.1.2, was used for sample demultiplexing and conversion of raw base call files into FASTQ files generated by Illumina sequencer (function: cellranger mkfastq), alignment to the mm10 mouse reference transcriptome (mm10-2020-A), filtering, barcode and unique molecular identifier (UMI) counting (function: cellranger count). Samples were then aggregated to normalize the runs to the same sequencing depth and then the feature-barcode matrices were recomputed (function: cellranger aggr) and after importing to R (version 4.2.0) they were used for further analysis. R packages Seurat[134] (version 4.1.1) and sctransform[135] (version 0.3.3) were used, following the recommended settings, for filtering of high-quality cells, normalization, principal component analysis, variable genes finding, clustering analysis, and UMAP dimensional reduction. Cells were grouped into 5 cell types based on the expression of signature genes identified in the previous studies[136,138,139] (on data following sctransform normalization). Differential gene expression analysis, GO and GSEA were performed on the log-normalized data using available functions from Seurat and Cluster profiler 4.0[142,143]. Differential gene expression analysis with cell type as a confounding factor was performed with EdgeR[148–150].

9.7.2 Strand-seq and SV discovery

9.7.2.1 Strand-seq library preparation

Strand-seq libraries were prepared according to a previously published protocol[112]. Briefly, cells (mammary gland organoids, MEFs, or mESCs) were incubated first with BrdU for exactly one cell division, then the single nuclei were isolated and analyzed by flow cytometry using nuclei from a BrdU-negative control to set up gates. Then single BrdU-containing nuclei were sorted into individual wells of a 96-well plate (100 μm nozzle, BD LSRFortessa™ Cell Analyzer (BD Biosciences)) containing 5 μl of freeze buffer (the composition summarized in Table 6), centrifuged for 5 minutes at 4ºC at full speed, and immediately frozen at -80ºC. Final libraries were prepared using a Beckman Coulter Biomek FxP liquid-handling robotic system (Beckman coulter) and then sequenced on a NextSeq Mid sequencer (75 paired ends).

Table 6 Composition of the freezing buffer.

Stock solution	For 5 ml freezing buffer
ProFreeze™-CDM (Lonza)	2.125 ml
DMSO	375 μl
PBS	2.5 ml

9.7.2.2 Bioinformatic analysis (pre-processing)

Raw sequencing data from Strand-seq libraries were demultiplexed and aligned to the mouse reference genome mm10 using BWA (0.7.15). PCR duplicates were marked before sorting and generating indexed bam files. Reads were then allocated to 200 kb bins across the genome (excluding blacklisted regions) and plotted as Strand-seq ideograms that show directional read distribution across all chromosomes in a cell. Low-quality libraries were removed based on manual curation of the ideograms (excluded libraries include the ones with fewer than 150,000 uniquely mapped reads, showing low or excess BrdU incorporation), and only high-quality libraries were then analyzed with scTRIP.

9.7.2.3 GC correction

In certain cases, if Strand-seq libraries were particularly affected by uneven read distribution along the chromosome (due to sequence bias of MNase) (visible as 'waviness' on Strand-seq ideograms), the data was GC corrected using a custom-made script by Dr. Marco Cosenza from Korbel lab. GC-corrected data was only used for a clearer presentation of Strand-seq results on ideograms.

9.7.2.4 Blacklisting

A blacklist for mm10 was created following the approach presented in the original scTRIP publication[113]. The list of excluded genomic regions was generated based on the data from 5 different Strand-seq experiments performed on murine cells (Strand-seq libraries derived from samples described in this thesis: murine mammary gland organoids never induced with doxycycline and induced with doxycycline for 5 days, and samples previously generated in the lab: B6xMEFxP5x02, B6CAST1F1, CAST1D1). The genome was divided into 100 kb windows, and bins with consistently distorted mean coverage across sequenced cells (<50% or >200% of the mean coverage in all bins) were blacklisted.

9.7.2.5 scTRIP

SV discovery (of duplications, deletions, inversions, and inverted duplications, as well as chromosome gains and losses) and annotation in high-quality Strand-seq libraries was performed with scTRIP that integrates template strand, read depth, and haplotype phase[113]. The pipeline relies on binned read counting, normalization of coverage, segmentation, strand state, SCE detection, and haplotype-aware SV classification. For this project, the original version of scTRIP for the human genome was updated for mm10. The 'strict' SV caller, optimized for the detection of SVs with clonal frequency ≥ 5%, was used to annotate germline SVs present in murine mammary gland organoids, while the 'lenient' caller was applied to detect SVs present in a single cell only.

9.7.2.6 Ploidy AssignR

Aneuploidies detected in Strand-seq libraries by scTRIP or through manual investigation were additionally confirmed using Ploidy AssignR, a newly developed tool from Korbel lab that accurately identifies aneuploidies in single cells independent of ploidy reference based on Strand-seq strand inheritance patterns.

9.7.2.7 SCE detection in individual cells

SCEs were identified in single cells as points on chromosome plots where reads mapping to both Watson and Crick strands switch to reads mapping to either the Watson or the Crick strand (without affecting the average read count). Coordinates of SCEs were extracted from the output of scTRIP, and to correct for Strand-seq resolution (200 kb), each SCE breakpoint was converted into a region ±100kb.

9.7.3 scNOVA

Generation of scNOVA, a computational framework using Strand-seq data for functional characterization of somatic SVs and prediction of gene activity changes based on differences in NO, is covered in detail in the original publication[114] (including definition of NO and CNN model). Here, I briefly describe how scNOVA was adapted for this project.

9.7.3.1 Cell-type classification

Cell-type classifier was built and trained on the scATAC-seq reference data[140]. The count matrix (peak by cells) was converted into motif accessibility matrix (motifs by cells) using chromVAR package[156]. The motif accessibility was then used as a feature to build a classifier based on PLS-DA. For feature selection, VIP values measuring discriminant power for each motif were calculated, and motifs with significant VIP values compared to null distribution (FDR 10%) were used to finalize the model and assess the performance with leave-one-out cross-validation. To apply the classifier on the Strand-seq data from murine mammary gland organoids, the NO count matrix (peak by cells) was converted into motif NO matrix (motifs by cells) using chromVAR. This matrix was then converted to motif accessibility matrix (motifs by cells): motif accessibility=(1)*motif NO Z-score. Based on the provided motif accessibility matrix, the classifier outputs the most likely cell-type of each Strand-seq library.

9.7.3.2 Inference of genome-wide changes in gene activity

To infer gene dysregulation, scNOVA follows two steps: in the first step, genes unlikely to be expressed are filtered out. This is performed based on the analysis of both NO and gene-context-specific features using CNNs. In the second step, the dysregulated (differentially expressed) genes between subclones are inferred using a generalized liner model.

The CNN model was trained (including leave-one-chromosome-out cross validation) and parametrized on the NO computed from Strand-seq data of MEFs and further validated using data from MEFs and mESCs. Ground-truth labels of not-expressed genes (NEs) and expressed genes (EGs) were defined based on bulk RNA-seq data from these two murine cell types (sample preparation described in 9.4.2). Reads were aligned to mm10 with STAR aligner (v2.6.0)[197] using gene annotations from ENSEMBL GTF. The relative gene expression was assessed using the FPKM values, with genes with FPKM>1 considered as EGs, and all the others as NEs.

9.7.4 scATAC-seq (for validation of scNOVA-based cell-type classifier)

9.7.4.1 scATAC-seq library preparation

Cells forming 1-week-old organoids derived from TetO-CMYC/TetO-Neu mice were harvested as described before. The solution of single cells was resuspended in 1 ml of 0.04% BSA in PBS and passed through a 30 μm MACS SmartStrainer (Miltenyi Biotec) into a 15-ml conical tube. 1 ml more of 0.04% BSA (Thermo Fisher Scientific) in PBS was passed through the strainer and the flowthrough was collected in the same conical tube. Cells were counted using a Countess II FL Automated Cell Counter (Thermo Fisher Scientific) and, depending on the sample, 200,000-300,000 cells were used for nuclei isolation according to the demonstrated protocol from 10x Genomics ('Nuclei Isolation from Mouse Brain Tissue for Single Cell ATAC Sequencing, Rev B, CG000212) with modifications (nuclei isolation with 1x Lysis Buffer instead of 0.1x, and for 9 minutes instead of 5 minutes). scATAC-seq libraries for each sample were prepared as per the standard 10x Genomics Chromium Next GEM Single Cell ATAC (v1.1) protocol (Rev F, CG000209). Targeted nuclei recovery was set for 7,000 per sample, and 13 cycles were included in sample index PCR. 1 μl of each constructed library was run on the Agilent Bioanalyzer High Sensitivity DNA chip (Agilent) to determine fragment size and confirm the presence of peaks indicative of the periodicity of the chromatin structure (nucleosome-free, mononucleosome, dinucleosome, and multinucleated fragments). The concentration of libraries was additionally measured with a Qubit™ 3.0 Fluorometer following the staining with Qubit™ dsDNA HS-Assay Kit (Thermo Fisher Scientific). Completed libraries were sequenced on a NextSeq2000 platform (50 bp paired-ends).

9.7.4.2 scATAC-seq data processing and cluster analysis

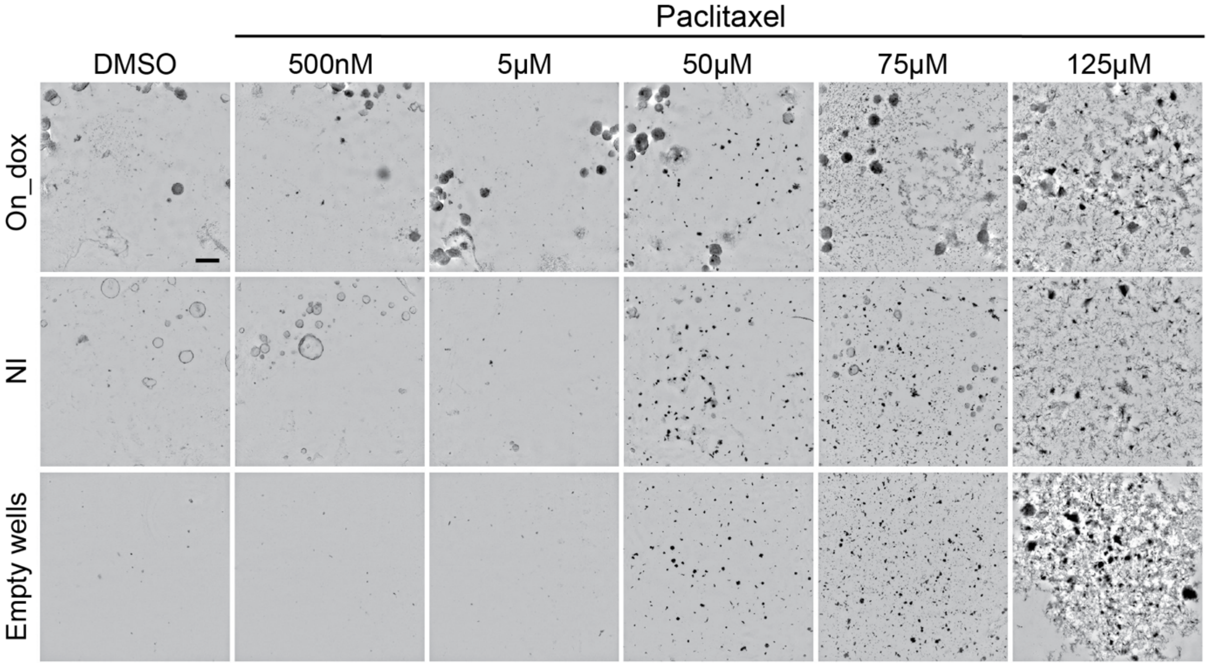
Raw base call files generated during sequencing were demultiplexed into FASTQ files using cellranger-atac (10x Genomics, version 2.1.0) mkfastq pipeline, and the downstream steps such as read filtering, alignment to the mouse reference genome (mm10) and peak calling were performed with the cellranger-atac count. Further data processing was performed using Signac[157], an R-based package for the analysis of scATAC-seq. The data from two samples were merged after creating a unified set of peaks containing promoter-proximal (-1000 bp, +100 bp of any TSS), promoter-distal (within 200 kb of the closest TSS or overlapping a gene body) or intergenic regions (not mapped to any gene). Cells with fewer than 3000 reads, cells with reads in peak ratio <15%, cells with nucleosomal signal >4, and cells with the TSS enrichment score <2 were considered of low quality and removed from the dataset.

Normalization, feature selection, and dimensionality reduction were carried out using the recommended settings. The correlation between each latent semantic indexing (LSI) component and sequencing depth was assessed, and the first LSI component was removed from the downstream analysis as it captured sequencing depth (technical variation) rather than biological variation. A low-dimensional visualization of the DNA accessibility was constructed using UMAP. Cell type identification was performed by cell type label transfer from the annotated scRNA-seq assay (of mice mammary gland organoids) to a new gene activity assay derived from the scATAC-seq data. This strategy relies on summing the fragments that span the body and promoter region of each gene and using the resulting value as a proxy of gene expression (assuming a general correspondence between the chromatin accessibility of a specific gene and its expression). To validate that the cell-type-specific motif accessibility between cell types was shared between the reference and scATAC-seq dataset described here, the scATAC-seq peaks were annotated depending on the presence of TF motifs using chromVAR[156]. The correlation matrix was restricted to 23 motifs used in the cell-type classifier.

9.8 Statistical analysis

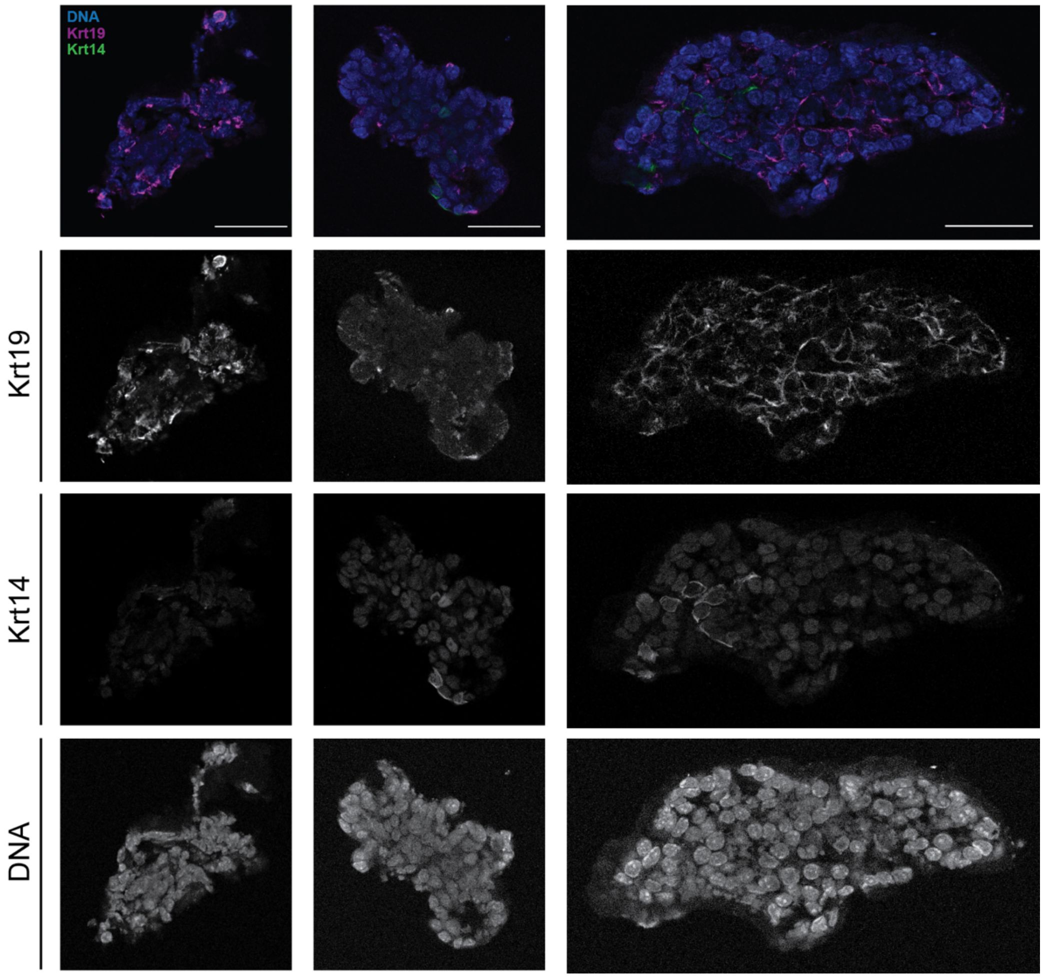
Statistical analysis was performed using R (versions 4.0.0, 4.0.5, 4.1.1, 4.1.3) or the free online version of GraphPad Software (Dotmatics). Permutation tests were performed with an R package regioneR[198]. The information about a number of experimental/biological and technical replicates, as well as the statistical test used, is provided for each experiment in the Methods section and/or on corresponding figures.

Chapter 10 Supplementary data



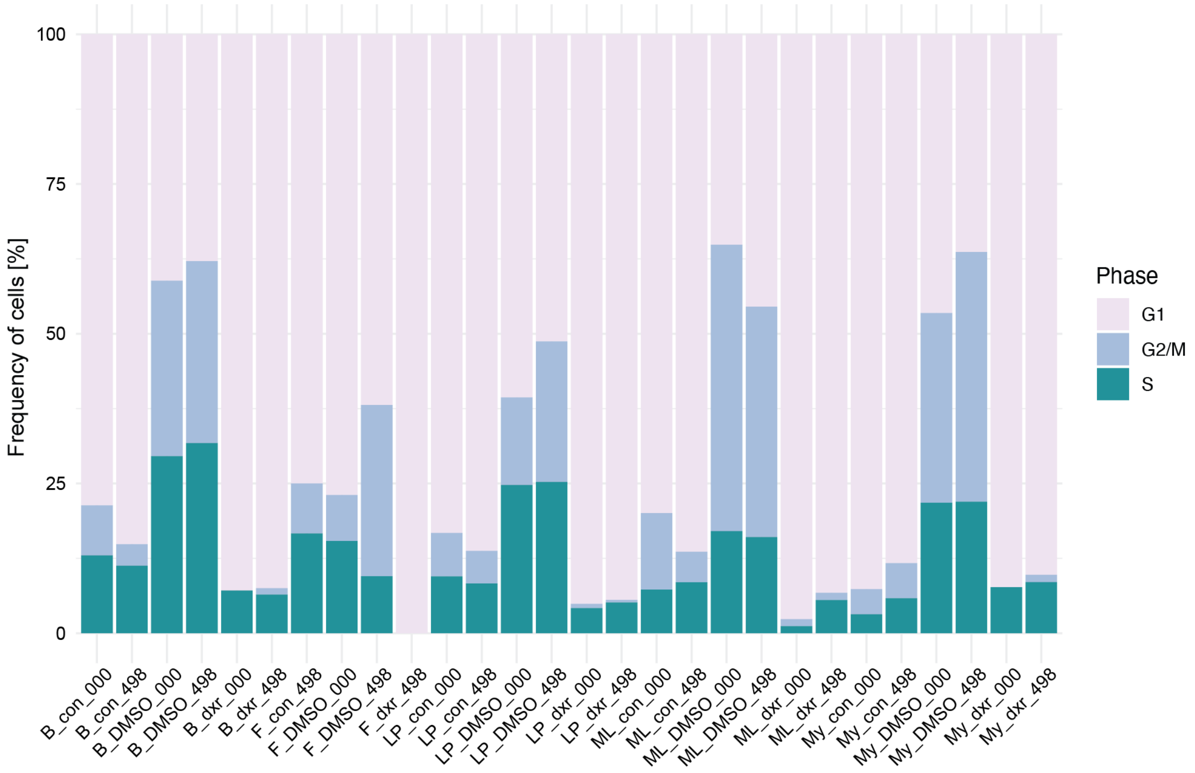
Sup. Figure 1 Paclitaxel precipitates in the medium when added in high concentrations.

Summary panel from one of the biological replicates of drug screen performed on murine mammary gland organoids (both never induced (NI) or induced with doxycycline (On_dox)). Paclitaxel crystals or precipitates are visible to the naked eye with concentrations above 50 µM. Each representative image included in the panel is a projection of 21 images taken per well (one well corresponds to one technical replicate).



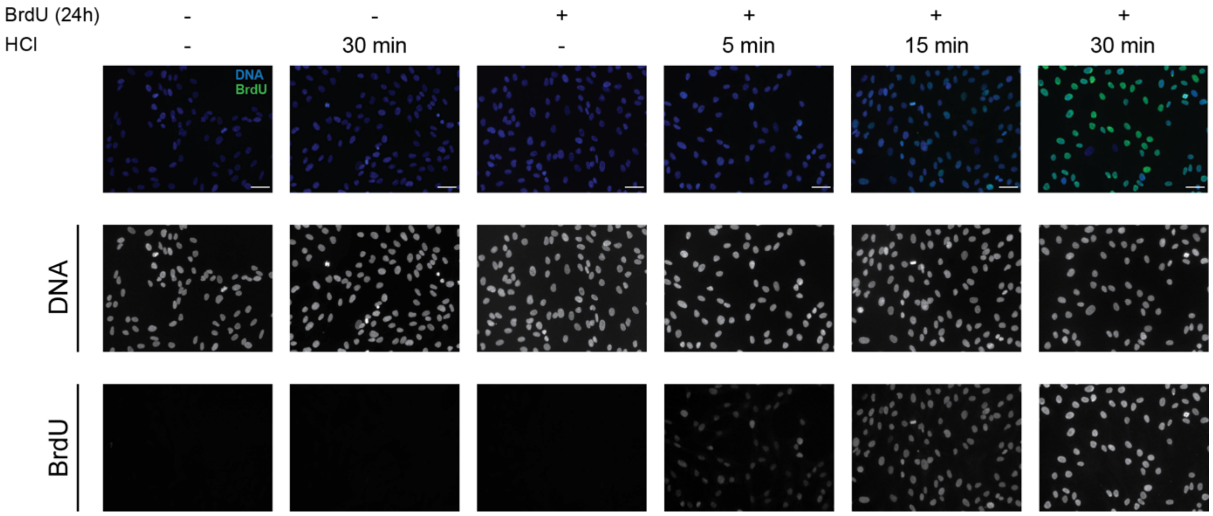
Sup. Figure 2 Keratin staining patterns of murine mammary gland organoids.

Mammary cells can be classified based on the cell surface markers and expression patterns of keratins as visualized on microscopic images. The gels containing the doxycycline-induced organoids (corresponding to cancer state) were cut into 8 µm cryosections. The cryosections were then fixed and stained for Krt19 (magenta) and Krt14 (green), and DNA was counterstained with (DAPI) (blue). Example images of single organoids show that luminal cells characterized by expression of Krt19 are more frequent in the organoids than basal cells expressing Krt14. Scale bar 50 µm.

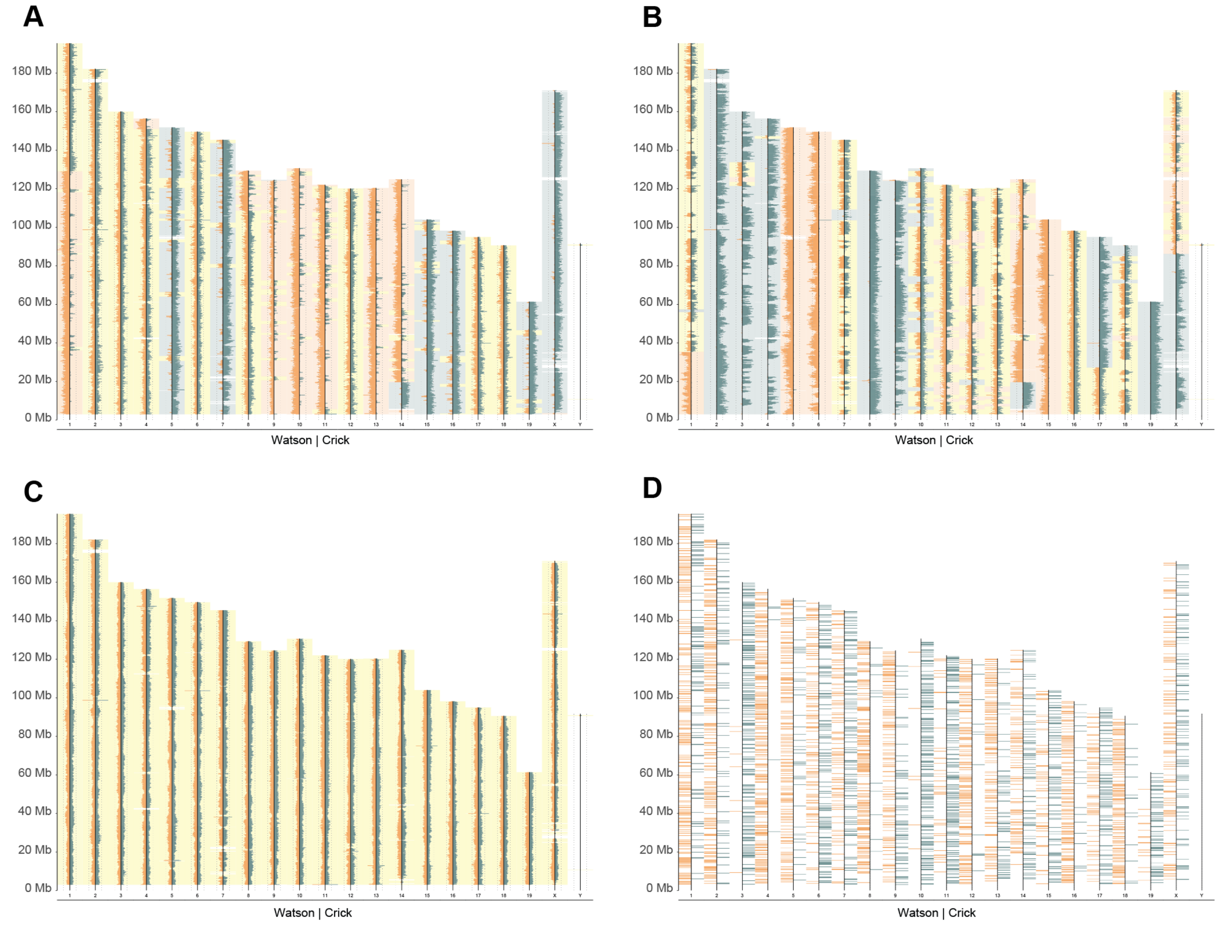


Sup. Figure 3 Differences in cell cycle progression after doxorubicin treatment are shared by cell types present in murine mammary gland organoids.

Relative frequency bar chart of cell types (B, F, LP, ML, My) present in murine mammary gland organoids treated with doxorubicin (dxr), DMSO (DMSO), or left untreated (con). Data from two biological replicates (000 and 498 are ID numbers of mice from which mammary glands were extracted). In both biological replicates, the population of fibroblasts was marginal (below 1% of all analyzed cells in the sample) or absent (as was the case for con_000 and dxr_000).

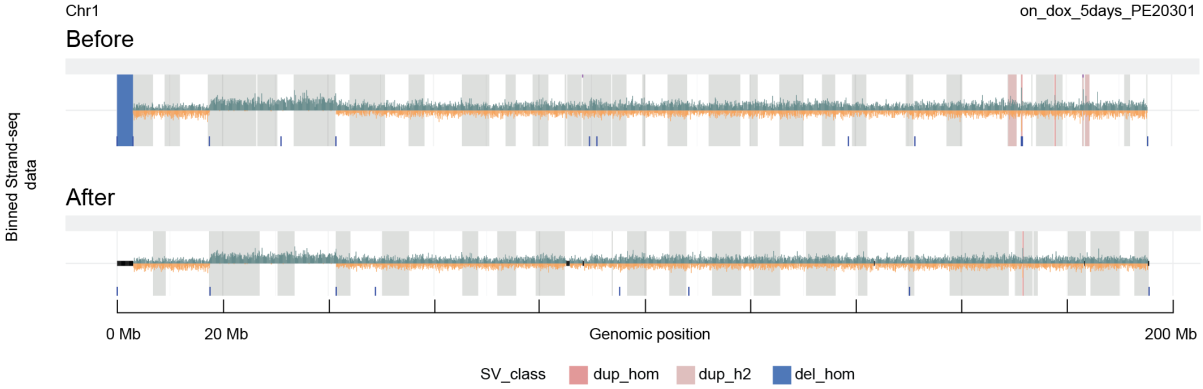


Sup. Figure 4 Successful anti-BrdU staining requires acid treatment during sample preparation. RPE-1 cells were plated on coverslips and exposed for 24 hours to BrdU. Coverslips were then fixed and processed for immunofluorescence. The BrdU incorporated into DNA was detected only if the samples were pretreated with HCl before the incubation with the primary antibody (green). Already 5 minutes of incubation with HCl allows detecting BrdU, with longer exposure (within the tested range of 5 to 30 minutes) giving a better signal. Nuclei were counterstained with Hoechst 33258 (blue). Scale bar: 50 µm.



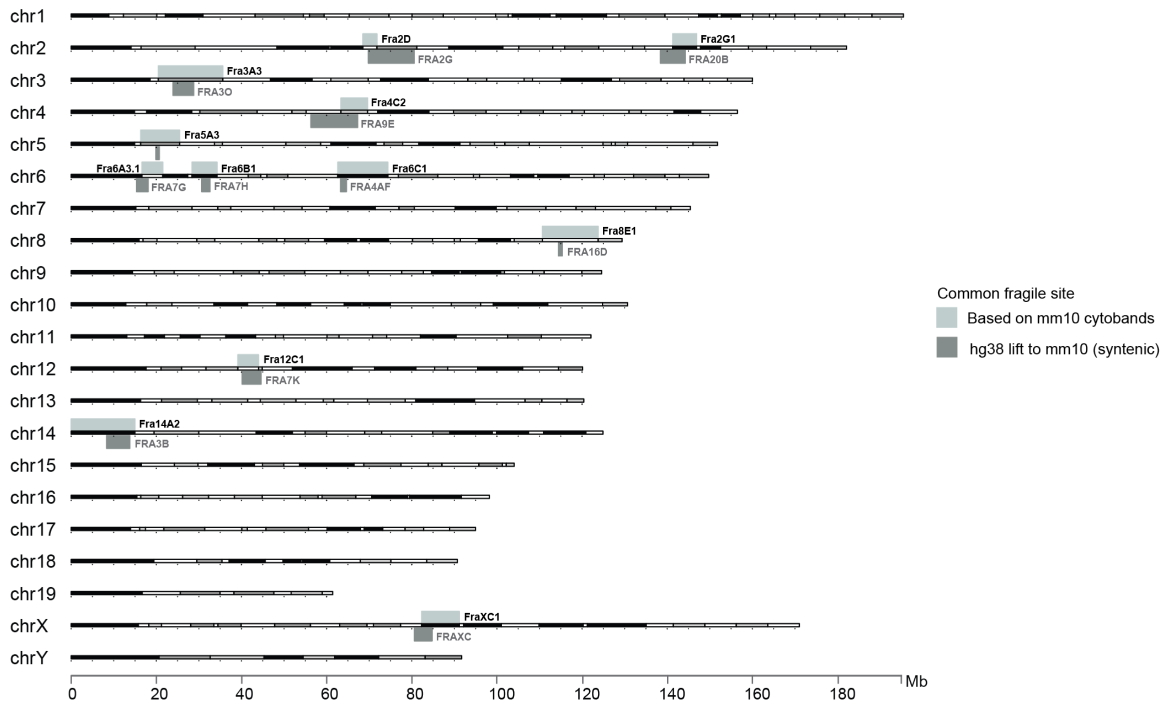
Sup. Figure 5 Examples of Strand-seq libraries of low quality and library preparation controls.

Depending on the length of BrdU exposure and gating conditions during the FACS, some libraries may exhibit signs of incomplete BrdU incorporation indicated by regions of increased reads of the opposite strand (A), or they have gone through more than a single round of BrdU incorporation which results in missing reads for one or both templates across all chromosomes (B). During the sort, two controls are introduced in each 96-well plate: 100-cell control (positive control) and no cells (negative control). (C) The 100-cell control shows much higher coverage than any single-cell library, and it resembles a whole-genome sequence pattern with all chromosomes represented by Watson and Crick reads. (D) The 0-cell control (as well as libraries lost during preparation on the robot) show no or extremely low number of sequencing reads. All ideograms were generated using the Strand-seq plotting pipeline. Examples come from different single-cell libraries from never-induced murine mammary gland organoids.



Sup. Figure 6 Blacklisting and normalization are necessary for proper SV calling with scTRIP.

Example of chr1 from a Strand-seq library of one cell (ID: PE20301) derived from murine mammary gland organoids induced with doxycycline for 5 days. Before the corrections, the scTRIP pipeline detected several SVs in that chromosome (indicated as colored boxes, color-coding according to the SV class) which were technical artifacts. Only one inverted duplication (between 171.5 and 171.7 Mb) was a true call. For the correct SV calling, strand-specific read counts need to be normalized and the regions of low mappability removed (indicated with a black thick line).



Sup. Figure 7 A karyogram summarizing the location of murine common fragile sites in mm10. Human CFS and their murine homologs were identified based on a literature review. The cytobands containing murine CFSs were reported only for the reference genome mm9 and needed to be updated for mm10 (shown in light grey, names of CFSs in black). To check that the updated annotation is correct, I converted the coordinates of human homologous CFSs (coordinates extracted from a database[177]) from hg38 to mm10 based on synteny (physical co-localization of genetic loci on the same chromosome between species). The lifted CFS in the murine genome are shown in dark grey, and the names of CFS in grey and in capital letters correspond to the human CFS homolog.

Sup. Table 1 Germline SVs present in murine strain TetO-CMYC/TetO-Neu/MMTV-rtTA in FVB/NJ background, detected with Strand-seq and validated with whole genome sequencing.

Strand-seq			**WGS**		
SV type	Coordinates (mm10)	Size	SV type	Coordinates (mm10)	Size
InvDup	chr1: 171.5-171.7 Mb	200 kb	Dup	chr1: 171,505-171,583 kb	78 kb
Inv	chr3: 93.6-94.0 Mb	400 kb	-	-	-
Del	chr4: 112.1-112.6 Mb	500 kb	Del	chr4: 112,057-112,600 kb	543 kb
Del	chr4: 112.8-113.6 Mb	800 kb	Del	not detected	-
Inv	chr4: 130.5-130.7 Mb	200 kb	-	-	-
InvDup	chr5: 15.4-15.7 Mb	300 kb	Dup	chr5: 15,457-15,716 kb	259 kb
Del	chr7: 14.9-15.6 Mb	700 kb	Del	chr7: 14,990-15,115 kb	125 kb
Dup	chr7: 38.1-38.3 Mb	200 kb	Dup	chr7: 38,175-38,201 kb	26 kb
Inv	chr8: 20.2-20.4 Mb	200 kb	-	-	-
Inv	chr12: 18.2-19.4 Mb	1.2 Mb	-	-	-
Del	chr12: 87.7-88.2 Mb	500 kb	Del	chr12: 87,655-87,756 kb	101 kb
Dup	chr12: 115.1-115.9 Mb	800 kb	Dup	chr12: 115,100-116,000 kb	900 kb
Inv	chr13: 65.4-68.6 Mb	400 kb	-	-	-
Inv	chr14: 0-19.6 Mb	19.6 Mb	-	-	-
Del	chr14: 44.2-44.7 Mb	500 kb	Del	chr14: 44,200-44,700 kb	500 kb
Del	chr17: 6.4-6.6 Mb	200 kb	Del	chr17: 6,420-6,590 kb	170 kb

Del: deletion, Dup: duplication, InvDup: inverted duplication, Inv: inversion

Sup. Table 2 List of murine common fragile sites and their coordinates in the reference genome mm10.

Large genes bigger than 0.5 Mb located within CFS are indicated with bold font. In red cytogenetic bands containing murine CFS that are different than the originally reported cytobands with CFS; for example, murine CFS Fra12C1 was reported to be located in the cytogenetic band 12qC1 but in mm10, this region corresponds to the band 12qB1. Coordinates of human CFS were extracted from a database[177].

LITERATURE				Cytogenetic band mm10			Coordinates of human homolog (hg38)			Liftover to mice genome hg38-mm10		
Murine CFS	Human CFS	Large gene **(>0.5 Mb)**	Cyt. band mm10	Chr	Start	End	Chr	Start	End	Chr	Start	End
Fra6C1	FRA4F	**Grid2**	6qC1	chr6	62634895	74378937	chr4	92303966	93810456	chr6	63256027	64668285
Fra8E1	FRA16D	**Wwox**	8qE1	chr8	110647414	123775073	chr16	24200000	28100000	chr8	114439655	115352708
Fra12C1	FRA7K	**Immp2l**	12qB1	chr12	39162941	44003304	chr7	107400000	114400000	chr12	40169806	44612010
Fra14A2	FRA3B	**Fhit** **Ptprg**	14qA2	chr14	1	14988269	chr3	58600000	63700000	chr14	8327179	13773275
Fra2D	FRA2G	Igrp, **Lrp**	2q2C	chr2	68527140	71812688	chr2	169700000	183000000	chr2	69795384	80550637
Fra6A3.1	FRA7G	Cav1, Cav2, Lpa	6qA2	chr6	16637394	21530745	chr7	114600000	117400000	chr6	15356232	18087378
Fra6B1	FRA7H	Sec8 (Exoc4)	6qA3.3	chr6	28381436	34253458	chr7	130400000	132600000	chr6	30656437	32616490
Fra4C2	FRA9E	**Astn2**	4qC1	chr4	63371384	69612505	chr9	108383899	118069329	chr4	56311755	67303093
FraXC1	FRAXC	**Dmd, Il1rapl1**	XqC1	chrX	82311893	91183833	chrX	31500000	37800000	chrX	80616184	84779479
Fra5A3	-	**Magi2**	5qA3	chr5	16336643	25465943	-	-	-	chr5	19907745	20702126
Fra3A3	FRA3O	**Nlgn1, Naaladl2**	3qA3	chr3	20492885	35618587	chr3	170900000	175700000	chr3	23910456	28787171
Fra2G1	FRA20B	**Macrod2**	2qG1	chr2	141278557	146910925	chr20	12000000	17900000	chr2	138366610	144218117

References

1. Szabo GK., Vandenberg LN. The male mammary gland: a novel target of endocrine-disrupting chemicals. *Reproduction* **162,** F79–F89 (2021).

2. Cristea S., Polyak K. Dissecting the mammary gland one cell at a time. *Nature Communications* **9,** 1–3 (2018).

3. Sung H., Ferlay J., Siegel RL., Laversanne M., Soerjomataram I., Jemal A., Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71,** 209–249 (2021).

4. Giordano SH. Breast Cancer in Men. *New England Journal of Medicine* **378,** 2311–2320 (2018).

5. Sauer S., Reed DR., Ihnat M., Hurst RE., Warshawsky D., Barkan D. Innovative Approaches in the Battle Against Cancer Recurrence: Novel Strategies to Combat Dormant Disseminated Tumor Cells. *Frontiers in Oncology* **11,** 1–19 (2021).

6. Zhang M., Lee A v., Rosen JM. The cellular origin and evolution of breast cancer. *Cold Spring Harbor Perspectives in Medicine* **7,** 1–15 (2017).

7. la Rosa S., Rubbia-Brandt L., Scoazec JY., Weber A. Editorial: Tumor Heterogeneity. *Frontiers in Medicine* **6,** 1–2 (2019).

8. Garattini S., Fuso Nerini I., D'Incalci M. Not only tumor but also therapy heterogeneity. *Annals of Oncology* **29,** 13–18 (2018).

9. Marusyk A., Janiszewska M., Polyak K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell* **37,** 471–484 (2020).

10. Banin Hirata BK., Oda JMM., Losi Guembarovski R., Ariza CB., Oliveira CEC de., Watanabe MAE. Molecular markers for breast cancer: Prediction on tumor behavior. *Disease Markers* **2014,** 1–13 (2014).

11. Prat A., Perou CM. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology* **5,** 5–23 (2011).

12. O'Brien KM., Cole SR., Tse CK., Perou CM., Carey LA., Foulkes WD., Dressler LG., Geradts J., Millikan RC. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clinical Cancer Research* **16,** 6100–6110 (2010).

13. Koboldt DC., Fulton RS., McLellan MD., Schmidt H., Kalicki-Veizer J., McMichael JF., Fulton LL., Dooling DJ., Ding L., Mardis ER., *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

14. Swain SM., Shastry M., Hamilton E. Targeting HER2-positive breast cancer: advances and future directions. *Nature Reviews Drug Discovery* **22,** 101–126 (2022).

15. Marchiò C., Annaratone L., Marques A., Casorzo L., Berrino E., Sapino A. Evolving concepts in HER2 evaluation in breast cancer: Heterogeneity, HER2-low carcinomas and beyond. *Seminars in Cancer Biology* **72,** 123–135 (2021).

16. Loibl S., Gianni L. HER2-positive breast cancer. *The Lancet* **389,** 2415–2429 (2017).

17. Milioli HH., Tishchenko I., Riveros C., Berretta R., Moscato P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Medical Genomics* **10,** 1–17 (2017).

18. Lüönd F., Tiede S., Christofori G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British Journal of Cancer* **125,** 164–175 (2021).

19. Milczarek M. The premature senescence in breast cancer treatment strategy. *Cancers* **12,** 1–22 (2020).

20. Dieci MV., Miglietta F., Guarneri V. Immune infiltrates in breast cancer: Recent updates and clinical implications. *Cells* **10,** 1–27 (2021).

21. McCleskey BC., Penedo TL., Zhang K., Hameed O., Siegal GP., Wei S. GATA3 expression in advanced breast cancer: Prognostic value and organ-specific relapse. *American Journal of Clinical Pathology* **144,** 756–763 (2015).

22. Ciriello G., Sinha R., Hoadley KA., Jacobsen AS., Reva B., Perou CM., Sander C., Schultz N. The molecular diversity of Luminal A breast tumors. *Breast Cancer Research and Treatment* **141,** 409–420 (2013).

23. Kwei KA., Kung Y., Salari K., Holcomb IN., Pollack JR. Genomic instability in breast cancer: Pathogenesis and clinical implications. *Molecular Oncology* **4,** 255–266 (2010).

24. Dorling L., Carvalho S., Allen J., González-Neira A., Luccarini C., Wahlström C., Pooley KA., Parsons MT., Fortuno C., Wang Q., *et al.* Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women. *New England Journal of Medicine* **384,** 428–439 (2021).

25. Kandoth C., McLellan MD., Vandin F., Ye K., Niu B., Lu C., Xie M., Zhang Q., McMichael JF., Wyczalkowski MA., *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).

26. Carvalho CMB., Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics* **17,** 224–238 (2016).

27. Raffaele Cosenza M., Rodriguez-Martin B., Korbel JO. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. *Annual Review of Genomics and Human Genetics* **23,** 123–152 (2022).

28. Dahiya R., Hu Q., Ly P. Mechanistic origins of diverse genome rearrangements in cancer. *Seminars in Cell and Developmental Biology* **123,** 100–109 (2022).

29. Marotta M., Chen X., Inoshita A., Stephens R., Thomas Budd G., Crowe JP., Lyons J., Kondratova A., Tubbs R., Tanaka H. A common copy-number breakpoint of ERBB2 amplification in breast cancer colocalizes with a complex block of segmental duplications. *Breast Cancer Research* **14,** 1–19 (2012).

30. Nishizaki T., DeVries S., Chew K., Goodson WH., Ljung BM., Thor A., Waldman FM. Genetic Alterations in Primary Breast Cancers and Their Metastases: Direct Comparison Using Modified Comparative Genomic Hybridization. *Genes Chromosomes and Cancer* **19,** 267–272 (1997).

31. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578,** 82–93 (2020).

32. Li Y., Roberts ND., Wala JA., Shapira O., Schumacher SE., Kumar K., Khurana E., Waszak S., Korbel JO., Haber JE., *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578,** 112–121 (2020).

33. Black JRM., McGranahan N. Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer* **21,** 379–392 (2021).

34. Hajdu SI. A note from history: Landmarks in history of cancer, part 1. *Cancer* **117,** 1097–1102 (2011).

35. Kerr AJ., Dodwell D., McGale P., Holt F., Duane F., Mannu G., Darby SC., Taylor CW. Adjuvant and neoadjuvant breast cancer treatments: A systematic review of their effects on mortality. *Cancer Treatment Reviews* **105,** 1–14 (2022).

36. Moo TA., Sanford R., Dang C., Morrow M. Overview of Breast Cancer Therapy. *PET Clinics* **13,** 339–354 (2018).

37. Migliaccio I., Malorni L., Hart CD., Guarducci C., di Leo A. Endocrine therapy considerations in postmenopausal patients with hormone receptor positive, human epidermal growth factor receptor type 2 negative advanced breast cancers. *BMC Medicine* **13,** 1–6 (2015).

38. Tremont A., Lu J., Cole JT. Endocrine Therapy for Early Breast Cancer: Updated Review. *Ochsner Journal* **17,** 405–417 (2017).

39. Yu S., Liu Q., Han X., Qin S., Zhao W., Li A., Wu K. Development and clinical application of anti-HER2 monoclonal and bispecific antibodies for cancer treatment. *Experimental Hematology and Oncology* **6,** 1–15 (2017).

40. Jerez Y., Márquez-Rodas I., Aparicio I., Alva M., Martín M., López-Tarruella S. Poly (ADP-ribose) Polymerase Inhibition in Patients with Breast Cancer and BRCA 1 and 2 Mutations. *Drugs* **80,** 131–146 (2020).

41. Fu X., Tan W., Song Q., Pei H., Li J. BRCA1 and Breast Cancer: Molecular Mechanisms and Therapeutic Strategies. *Frontiers in Cell and Developmental Biology* **10,** 1–11 (2022).

42. Anand U., Dey A., Chandel AKS., Sanyal R., Mishra A., Pandey DK., de Falco V., Upadhyay A., Kandimalla R., Chaudhary A., *et al.* Cancer chemotherapy and beyond: Current status, drug candidates, associated risks and progress in targeted therapeutics. *Genes and Diseases* (2022). doi:10.1016/j.gendis.2022.02.007

43. van den Boogaard WMC., Komninos DSJ., Vermeij WP. Chemotherapy Side-Effects: Not All DNA Damage Is Equal. *Cancers* **14,** 1–27 (2022).

44. Weiss F., Lauffenburger D., Friedl P. Towards targeting of shared mechanisms of cancer metastasis and therapy resistance. *Nature Reviews Cancer* **22,** 157–173 (2022).

45. Hausser J., Alon U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nature Reviews Cancer* **20,** 247–257 (2020).

46. Alexandrov LB., Kim J., Haradhvala NJ., Huang MN., Tian Ng AW., Wu Y., Boot A., Covington KR., Gordenin DA., Bergstrom EN., *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578,** 94–101 (2020).

47. Alexandrov LB., Nik-Zainal S., Wedge DC., Campbell PJ., Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3,** 246–259 (2013).

48. Liu D., Abbosh P., Keliher D., Reardon B., Miao D., Mouw K., Weiner-Taylor A., Wankowicz S., Han G., Teo MY., *et al.* Mutational patterns in chemotherapy resistant muscle-invasive bladder cancer. *Nature Communications* **8,** 1–11 (2017).

49. Boot A., Huang MN., Ng AWT., Ho SC., Lim JQ., Kawakami Y., Chayama K., Teh BT., Nakagawa H., Rozen SG. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Research* **28,** 654–665 (2018).

50. Kucab JE., Zou X., Morganella S., Joel M., Nanda AS., Nagy E., Gomez C., Degasperi A., Harris R., Jackson SP., *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177,** 821-836.e16 (2019).

51. Pich O., Muiños F., Lolkema MP., Steeghs N., Gonzalez-Perez A., Lopez-Bigas N. The mutational footprints of cancer therapies. *Nature Genetics* **51,** 1732–1740 (2019).

52. Christensen S., van der Roest B., Besselink N., Janssen R., Boymans S., Martens JWM., Yaspo ML., Priestley P., Kuijk E., Cuppen E., *et al.* 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nature Communications* **10,** 1–11 (2019).

53. Wang H., Xiao X., Xiao Q., Lu Y., Wu Y. The efficacy and safety of daunorubicin versus idarubicin combined with cytarabine for induction therapy in acute myeloid leukemia: A meta-analysis of randomized clinical trials. *Medicine* **99,** e20094 (2020).

54. Khasraw M., Bell R., Dang C. Epirubicin: Is it like doxorubicin in breast cancer? A clinical review. *Breast* **21,** 142–149 (2012).

55. Meredith AM., Dass CR. Increasing role of the cancer chemotherapeutic doxorubicin in cellular metabolism. *Journal of Pharmacy and Pharmacology* **68,** 729–741 (2016).

56. Yang F., Teves SS., Kemp CJ., Henikoff S. Doxorubicin, DNA torsion, and chromatin dynamics. *Biochimica et Biophysica Acta - Reviews on Cancer* **1845,** 84–89 (2014).

57. Uusküla-Reimand L., Wilson MD. Untangling the roles of TOP2A and TOP2B in transcription and cancer. *Science Advances* **8,** 1–16 (2022).

58. Baron B. Doxorubicin: An Overview of the Anti-Cancer and Chemoresistance Mechanisms. *Ann Clin Toxicol* **3,** 1–12 (2020).

59. Pang B., Qiao X., Janssen L., Velds A., Groothuis T., Kerkhoven R., Nieuwland M., Ovaa H., Rottenberg S., van Tellingen O., *et al.* Drug-induced histone eviction from open chromatin contributes to the chemotherapeutic effects of doxorubicin. *Nature Communications* **4,** 1–13 (2013).

60. Yang F., Kemp CJ., Henikoff S. Doxorubicin enhances nucleosome turnover around promoters. *Current Biology* **23,** 782–787 (2013).

61. Qiao X., van der Zanden SY., Wander DPA., Borràs DM., Song J-Y., Li X., van Duikeren S., van Gils N., Rutten A., van Herwaarden T., *et al.* Uncoupling DNA damage from chromatin damage to detoxify doxorubicin. *Proceedings of the National Academy of Sciences* **117,** 15182–15192 (2020).

62. Seoane JA., Kirkland JG., Caswell-Jin JL., Crabtree GR., Curtis C. Chromatin regulators mediate anthracycline sensitivity in breast cancer. *Nature Medicine* **25,** 1721–1727 (2019).

63. Szikriszt B., PótiÁdám., Pipek O., Krzystanek M., Kanu N., Molnár J., Ribli D., Szeltner Z., Tusnády GE., Csabai I., *et al.* A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biology* **17,** 1–16 (2016).

64. Pleasance E., Titmuss E., Williamson L., Kwan H., Culibrk L., Zhao EY., Dixon K., Fan K., Bowlby R., Jones MR., *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nature cancer* **1,** 452–468 (2020).

65. Mardin BR., Drainas AP., Waszak SM., Weischenfeldt J., Isokane M., Stütz AM., Raeder B., Efthymiopoulos T., Buccitelli C., Segura-Wang M., *et al.* A cell-based model system links chromothripsis with hyperploidy. *Molecular Systems Biology* **11,** 1–13 (2015).

66. Cappetta D., de Angelis A., Sapio L., Prezioso L., Illiano M., Quaini F., Rossi F., Berrino L., Naviglio S., Urbanek K. Oxidative stress and cellular response to doxorubicin: A common factor in the complex milieu of anthracycline cardiotoxicity. *Oxidative Medicine and Cellular Longevity* **2017,** 1–13 (2017).

67. Bhatia S. Genetics of Anthracycline Cardiomyopathy in Cancer Survivors. *JACC: CardioOncology* **2,** 539–552 (2020).

68. Tacar O., Sriamornsak P., Dass CR. Doxorubicin: An update on anticancer molecular action, toxicity and novel drug delivery systems. *Journal of Pharmacy and Pharmacology* **65,** 157–170 (2013).

69. Okabe M., Unno M., Harigae H., Kaku M., Okitsu Y., Sasaki T., Mizoi T., Shiiba K., Takanaga H., Terasaki T., *et al.* Characterization of the organic cation transporter SLC22A16: A doxorubicin importer. *Biochemical and Biophysical Research Communications* **333,** 754–762 (2005).

70. Calcagno AM., Ambudkar S v. Molecular mechanisms of drug resistance in single-step and multi-step drug-selected cancer cells. *Methods in molecular biology* **596,** 77–93 (2010).

71. Singhal S., Singhal J., Nair M., Lacko A., Awasthi Y., Awasthi S. Doxorubicin transport by RALBP1 and ABCG2 in lung and breast cancer. *International Journal of Oncology* **30,** 717–725 (2007).

72. Siebel C., Lanvers-Kaminsky C., Würthwein G., Hempel G., Boos J. Bioanalysis of doxorubicin aglycone metabolites in human plasma samples–implications for doxorubicin drug monitoring. *Scientific Reports* **10,** 1–7 (2020).

73. Bains OS., Karkling MJ., Lubieniecka JM., Grigliatti TA., Reid RE., Riggs KW. Naturally occurring variants of human CBR3 alter anthracycline in vitro metabolism. *Journal of Pharmacology and Experimental Therapeutics* **332,** 755–763 (2010).

74. Kassner N., Huse K., Martin HJ., Gödtel-Armbrust U., Metzger A., Meineke I., Brockmöller J., Klein K., Zanger UM., Maser E., *et al.* Carbonyl reductase 1 is a predominant doxorubicin reductase in the human liver. *Drug Metabolism and Disposition* **36,** 2113–2120 (2008).

75. Fujii J., Homma T., Miyata S., Takahashi M. Pleiotropic actions of aldehyde reductase (Akr1a). *Metabolites* **11,** 1–22 (2021).

76. Marine JC., Dawson SJ., Dawson MA. Non-genetic mechanisms of therapeutic resistance in cancer. *Nature Reviews Cancer* **20,** 743–756 (2020).

77. Hanahan D. Hallmarks of Cancer: New Dimensions. *Cancer Discovery* **12,** 31–46 (2022).

78. Ji X., Lu Y., Tian H., Meng X., Wei M., Cho WC. Chemoresistance mechanisms of breast cancer and their countermeasures. *Biomedicine and Pharmacotherapy* **114,** 1–9 (2019).

79. McGuirk S., Audet-Delage Y., Annis MG., Xue Y., Vernier M., Zhao K., St-Louis C., Minarrieta L., Patten DA., Morin G., *et al.* Resistance to different anthracycline chemotherapeutics elicits distinct and actionable primary metabolic dependencies in breast cancer. *eLife* **10,** 1–29 (2021).

80. Braga S. Resistance to Targeted Therapies in Breast Cancer. *Methods in Molecular Biology* **1395,** 105–136 (2016).

81. Phan TG., Croucher PI. The dormant cancer cell life cycle. *Nature Reviews Cancer* **20,** 398–411 (2020).

82. Wang L., Lankhorst L., Bernards R. Exploiting senescence for the treatment of cancer. *Nature Reviews Cancer* **22,** 340–355 (2022).

83. Kwon SM., Hong SM., Lee YK., Min S., Yoon G. Metabolic features and regulation in cell senescence. *BMB Reports* **52,** 5–12 (2019).

84. Gomis RR., Gawrzak S. Tumor cell dormancy. *Molecular Oncology* **11,** 62–78 (2017).

85. Triana-Martínez F., Loza MI., Domínguez E. Beyond Tumor Suppression: Senescence in Cancer Stemness and Tumor Dormancy. *Cells* **9,** 1–28 (2020).

86. Luskin MR., Murakami MA., Manalis SR., Weinstock DM. Targeting minimal residual disease: A path to cure? *Nature Reviews Cancer* **18,** 255–263 (2018).

87. Tachtsidis A., McInnes LM., Jacobsen N., Thompson EW., Saunders CM. Minimal residual disease in breast cancer: an overview of circulating and disseminated tumour cells. *Clinical and Experimental Metastasis* **33,** 521–550 (2016).

88. Sumbal J., Budkova Z., Gunnhildur &., Traustadóttir Á., Koledova Z. Mammary Organoids and 3D Cell Cultures: Old Dogs with New Tricks. *Journal of Mammary Gland Biology and Neoplasia* **25,** 273–288 (2020).

89. Fu NY., Nolan E., Lindeman GJ., Visvader JE. Stem cells and the differentiation hierarchy in mammary gland development. *Physiological Reviews* **100,** 489–523 (2020).

90. Gieniec KA., Davis FM. Mammary basal cells: Stars of the show. *Biochimica et Biophysica Acta - Molecular Cell Research* **1869,** 1–6 (2022).

91. Dontu G., Ince TA. Of Mice and Women: A Comparative Tissue Biology Perspective of Breast Stem Cells and Differentiation. *Journal of Mammary Gland Biology and Neoplasia* **20,** 51–62 (2015).

92. Twigger AJ., Khaled WT. Mammary gland development from a single cell 'omics view. *Seminars in Cell and Developmental Biology* **114,** 171–185 (2021).

93. Ren L., Li J., Wang C., Lou Z., Gao S., Zhao L., Wang S., Chaulagain A., Zhang M., Li X., *et al.* Single cell RNA sequencing for breast cancer: present and future. *Cell Death Discovery* **7,** 1–11 (2021).

94. Li M., Izpisua Belmonte JC. Organoids — Preclinical Models of Human Disease. *New England Journal of Medicine* **380,** 569–579 (2019).

95. Zhou J., Li C., Liu X., Chiu MC., Zhao X., Wang D., Wei Y., Lee A., Zhang AJ., Chu H., *et al.* Infection of bat and human intestinal organoids by SARS-CoV-2. *Nature Medicine* **26,** 1077–1083 (2020).

96. Post Y., Puschhof J., Beumer J., Kerkkamp HM., de Bakker MAG., Slagboom J., de Barbanson B., Wevers NR., Spijkers XM., Olivier T., *et al.* Snake Venom Gland Organoids. *Cell* **180,** 233-247.e21 (2020).

97. Drost J., Clevers H. Organoids in cancer research. *Nature Reviews Cancer* **18,** 407–418 (2018).

98. Rosenbluth JM., Schackmann RCJ., Gray GK., Selfors LM., Li CMC., Boedicker M., Kuiken HJ., Richardson A., Brock J., Garber J., *et al.* Organoid cultures from normal and cancer-prone human breast tissues preserve complex epithelial lineages. *Nature Communications* **11,** 1–14 (2020).

99. Mohan SC., Lee TY., Giuliano AE., Cui X. Current Status of Breast Organoid Models. *Frontiers in Bioengineering and Biotechnology* **9,** 1–7 (2021).

100. Sachs N., de Ligt J., Kopper O., Gogola E., Bounova G., Weeber F., Balgobind AV., Wind K., Gracanin A., Begthel H., *et al.* A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. *Cell* **172,** 373-386.e10 (2018).

101. Bhatia S., Kramer M., Russo S., Naik P., Arun G., Brophy K., Andrews P., Fan C., Perou CM., Preall J., *et al.* Patient-Derived Triple-Negative Breast Cancer Organoids Provide Robust Model Systems That Recapitulate Tumor Intrinsic Characteristics. *Cancer Research* **82,** 1174–1192 (2022).

102. Havas KM., Milchevskaya V., Radic K., Alladin A., Kafkia E., Garcia M., Stolte J., Klaus B., Rotmensz N., Gibson TJ., *et al.* Metabolic shifts in residual breast cancer drive tumor recurrence. *Journal of Clinical Investigation* **127,** 2091–2105 (2017).

103. Radic Shechter K., Kafkia E., Zirngibl K., Gawrzak S., Alladin A., Machado D., Lüchtenborg C., Sévin DC., Brügger B., Patil KR., *et al.* Metabolic memory underlying minimal residual disease in breast cancer. *Molecular Systems Biology* **17,** 1–21 (2021).

104. Nair R., Roden DL., Teo WS., McFarland A., Junankar S., Ye S., Nguyen A., Yang J., Nikolic I., Hui M., *et al.* C-Myc and Her2 cooperate to drive a stem-like phenotype with poor prognosis in breast cancer. *Oncogene* **33,** 3992–4002 (2014).

105. Risom T., Wang X., Liang J., Zhang X., Pelz C., Campbell LG., Eng J., Chin K., Farrington C., Narla G., *et al.* Deregulating MYC in a model of HER2+ breast cancer mimics human intertumoral heterogeneity. *Journal of Clinical Investigation* **130,** 231–246 (2020).

106. Das AT., Tenenbaum L., Berkhout B. Tet-On Systems For Doxycycline-inducible Gene Expression. *Current Gene Therapy* **16,** 156–167 (2016).

107. Dhanasekaran R., Deutzmann A., Mahauad-Fernandez WD., Hansen AS., Gouw AM., Felsher DW. The MYC oncogene — the grand orchestrator of cancer growth and immune evasion. *Nature Reviews Clinical Oncology* **19,** 23–36 (2022).

108. Nam AS., Chaligne R., Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nature Reviews Genetics* **22,** 3–18 (2021).

109. Kharchenko P v. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods* **18,** 723–732 (2021).

110. Twigger AJ., Khaled WT. Mammary gland development from a single cell 'omics view. *Seminars in Cell and Developmental Biology* **114,** 171–185 (2021).

111. Falconer E., Hills M., Naumann U., Poon SSS., Chavez EA., Sanders AD., Zhao Y., Hirst M., Lansdorp PM. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature Methods* **9,** 1107–1112 (2012).

112. Sanders AD., Falconer E., Hills M., Spierings DCJ., Lansdorp PM. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nature Protocols* **12,** 1151–1176 (2017).

113. Sanders AD., Meiers S., Ghareghani M., Porubsky D., Jeong H., van Vliet MACC., Rausch T., Richter-Pechańska P., Kunz JB., Jenni S., *et al.* Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nature Biotechnology* **38,** 343–354 (2020).

114. Jeong H., Grimes K., Rauwolf KK., Bruch P-M., Rausch T., Hasenfeld P., Benito E., Roider T., Sabarinathan R., Porubsky D., *et al.* Functional analysis of structural variants in single cells using Strand-seq. *Nature Biotechnology* (2022). doi:10.1038/s41587-022-01551-4

115. Mahmoud M., Gobet N., Cruz-Dávalos DI., Mounier N., Dessimoz C., Sedlazeck FJ. Structural variant calling: The long and the short of it. *Genome Biology* **20,** 1–14 (2019).

116. Yang L. A Practical Guide for Structural Variation Detection in the Human Genome. *Current Protocols in Human Genetics* **107,** 1–17 (2020).

117. Caswell-Jin JL., Lorenz C., Curtis C. Molecular Heterogeneity and Evolution in Breast Cancer. *Annual Review of Cancer Biology* **5,** 79–94 (2021).

118. Macaulay IC., Voet T. Single Cell Genomics: Advances and Future Perspectives. *PLoS Genetics* **10,** 1–9 (2014).

119. Falconer E., Lansdorp PM. Strand-seq: A unifying tool for studies of chromosome segregation. *Seminars in Cell and Developmental Biology* **24,** 643–652 (2013).

120. Claussin C., Porubsky D., Spierings DC., Halsema N., Rentas S., Guryev V., Lansdorp PM., Chang M. Genome-wide mapping of sister chromatid exchange events in single yeast cells using Strand-seq. *eLife* **6,** 1–17 (2017).

121. van Wietmarschen N., Lansdorp PM. Bromodeoxyuridine does not contribute to sister chromatid exchange events in normal or Bloom syndrome cells. *Nucleic Acids Research* **44,** 6787–6793 (2016).

122. Heijink AM., Stok C., Porubsky D., Manolika EM., de Kanter JK., Kok YP., Everts M., de Boer HR., Audrey A., Bakker FJ., *et al.* Sister chromatid exchanges induced by perturbed replication can form independently of BRCA1, BRCA2 and RAD51. *Nature Communications* **13,** 1–16 (2022).

123. Ebert P., Audano PA., Zhu Q., Rodriguez-Martin B., Porubsky D., Bonder MJ., Sulovari A., Ebler J., Zhou W., Mari RS., *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372,** 1–13 (2021).

124. Porubský D., Sanders AD., van Wietmarschen N., Falconer E., Hills M., Spierings DCJ., Bevova MR., Guryev V., Lansdorp PM. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Research* **26,** 1565–1574 (2016).

125. Porubsky D., Ebert P., Audano PA., Vollger MR., Harvey WT., Marijon P., Ebler J., Munson KM., Sorensen M., Sulovari A., *et al.* Fully phased human genome assembly without parental

data using single-cell strand sequencing and long reads. *Nature Biotechnology* (2020). doi:10.1038/s41587-020-0719-5

126. Porubsky D., Garg S., Sanders AD., Korbel JO., Guryev V., Lansdorp PM., Marschall T. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature Communications* **8,** 1–10 (2017).

127. Porubsky D., Sanders AD., Höps W., Hsieh PH., Sulovari A., Li R., Mercuri L., Sorensen M., Murali SC., Gordon D., *et al.* Recurrent inversion toggling and great ape genome evolution. *Nature Genetics* **52,** 849–858 (2020).

128. Porubsky D., Höps W., Ashraf H., Hsieh PH., Rodriguez-Martin B., Yilmaz F., Ebler J., Hallast P., Maria Maggiolini FA., Harvey WT., *et al.* Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185,** 1986-2005.e26 (2022).

129. Driehuis E., Kretzschmar K., Clevers H. Establishment of patient-derived cancer organoids for drug-screening applications. *Nature Protocols* **15,** 3380–3409 (2020).

130. Weaver BA. How Taxol/paclitaxel kills cancer cells. *Molecular Biology of the Cell* **25,** 2677–2681 (2014).

131. Tsang RY., Sadeghi S., Finn RS. Lapatinib, a Dual-targeted small molecule inhibitor of EGFR and HER2, in HER2-Amplified breast cancer: From bench to bedside. *Clinical Medicine Insights: Therapeutics* **3,** 1–13 (2011).

132. Lewis Phillips G., Guo J., Kiefer JR., Proctor W., Bumbaca Yadav D., Dybdal N., Shen BQ. Trastuzumab does not bind rat or mouse ErbB2/neu: implications for selection of non-clinical safety models for trastuzumab-based therapeutics. *Breast Cancer Research and Treatment* **191,** 303–317 (2022).

133. Wan Mohamad Zain WNI., Joanne B., Bateman E., Keefe D. Cytotoxic Effects of the Dual ErbB Tyrosine Kinase Inhibitor, Lapatinib, on Walker 256 Rat Breast Tumour and IEC-6 Rat Normal Small Intestinal Cell Lines. *Biomedicines* **8,** 1–14 (2020).

134. Stuart T., Butler A., Hoffman P., Hafemeister C., Papalexi E., Mauck WM., Hao Y., Stoeckius M., Smibert P., Satija R. Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888-1902.e21 (2019).

135. Hafemeister C., Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20,** 1–15 (2019).

136. Bach K., Pensa S., Grzelak M., Hadfield J., Adams DJ., Marioni JC., Khaled WT. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications* **8,** 1–11 (2017).

137. Li CMC., Shapiro H., Tsiobikas C., Selfors LM., Chen H., Rosenbluth J., Moore K., Gupta KP., Gray GK., Oren Y., *et al.* Aging-Associated Alterations in Mammary Epithelia and Stroma Revealed by Single-Cell RNA Sequencing. *Cell Reports* **33,** 1–23 (2020).

138. Pal B., Chen Y., Milevskiy MJG., Vaillant F., Prokopuk L., Dawson CA., Capaldo BD., Song X., Jackling F., Timpson P., *et al.* Single cell transcriptome atlas of mouse mammary epithelial cells across development. *Breast Cancer Research* **23,** 1–19 (2021).

139. Saeki K., Chang G., Kanaya N., Wu X., Wang J., Bernal L., Ha D., Neuhausen SL., Chen S. Mammary cell gene expression atlas links epithelial cell remodeling events to breast carcinogenesis. *Communications Biology* **4,** 1–16 (2021).

140. Chung CY., Ma Z., Dravis C., Preissl S., Poirion O., Luna G., Hou X., Giraddi RR., Ren B., Wahl GM. Single-Cell Chromatin Analysis of Mammary Gland Development Reveals Cell-State Transcriptional Regulators and Lineage Relationships. *Cell Reports* **29,** 495-510.e6 (2019).

141. Luecken MD., Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15,** 1–23 (2019).

142. Yu G., Wang LG., Han Y., He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* **16,** 284–287 (2012).

143. Wu T., Hu E., Xu S., Chen M., Guo P., Dai Z., Feng T., Zhou L., Tang W., Zhan L., *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2,** 1–10 (2021).

144. Bagge J., Oestergaard VH., Lisby M. Functions of TopBP1 in preserving genome integrity during mitosis. *Seminars in Cell and Developmental Biology* **113,** 57–64 (2021).

145. Cox J., Weinman S. Mechanisms of doxorubicin resistance in hepatocellular carcinoma. *Hepatic Oncology* **3,** 57–59 (2016).

146. Kumar U., Castellanos-Uribe M., May ST., Yagüe E. Adaptive resistance is not responsible for long-term drug resistance in a cellular model of triple negative breast cancer. *Gene* **850,** 1–9 (2023).

147. Fridman AL., Tainsky MA. Critical pathways in cellular senescence and immortalization revealed by gene expression profiling. *Oncogene* **27,** 5975–5987 (2008).

148. Robinson MD., McCarthy DJ., Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).

149. McCarthy DJ., Chen Y., Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40,** 4288–4297 (2012).

150. Chen Y., Lun ATL., Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* **5,** 1–51 (2016).

151. Law CW., Chen Y., Shi W., Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15,** 1–17 (2014).

152. Ritchie ME., Phipson B., Wu D., Hu Y., Law CW., Shi W., Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** 1–13 (2015).

153. Maria Maggiolini FA., Sanders AD., Shew CJ., Sulovari A., Mao Y., Puig M., Catacchio CR., Dellino M., Palmisano D., Mercuri L., *et al.* Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Research* **30,** 1680–1693 (2020).

154. Tolstorukov MY., Kharchenko P v., Park PJ. Analysis of the primary structure of chromatin with next-generation sequencing. *Epigenomics* **2,** 187–197 (2010).

155. Yan F., Powell DR., Curtis DJ., Wong NC. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21,** 1–16 (2020).

156. Schep AN., Wu B., Buenrostro JD., Greenleaf WJ. ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods* **14,** 975–978 (2017).

157. Stuart T., Srivastava A., Madad S., Lareau CA., Satija R. Single-cell chromatin state analysis with Signac. *Nature Methods* **18,** 1333–1341 (2021).

158. Struhl K., Segal E. Determinants of nucleosome positioning. *Nature Structural and Molecular Biology* **20,** 267–273 (2013).

159. Love MI., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 1–21 (2014).

160. Lai B., Gao W., Cui K., Xie W., Tang Q., Jin W., Hu G., Ni B., Zhao K. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **562,** 281–285 (2018).

161. Ghimire S., van der Jeught M., Neupane J., Roost MS., Anckaert J., Popovic M., van Nieuwerburgh F., Mestdagh P., Vandesompele J., Deforce D., *et al.* Comparative analysis of naive, primed and ground state pluripotency in mouse embryonic stem cells originating from the same genetic background. *Scientific Reports* **8,** 1–11 (2018).

162. Tian S., Feng J., Cao Y., Shen S., Cai Y., Yang D., Yan R., Wang L., Zhang H., Zhong X., *et al.* Glycine cleavage system determines the fate of pluripotent stem cells via the regulation of senescence and epigenetic modifications. *Life Science Alliance* **2,** 1–18 (2019).

163. Muhl L., Genové G., Leptidis S., Liu J., He L., Mocci G., Sun Y., Gustafsson S., Buyandelger B., Chivukula I v., *et al.* Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nature Communications* **11,** 1–18 (2020).

164. Lachmann A., Xu H., Krishnan J., Berger SI., Mazloom AR., Ma'ayan A. ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26,** 2438–2444 (2010).

165. Wong K., Bumpstead S., van der Weyden L., Reinholdt LG., Wilming LG., Adams DJ., Keane TM. Sequencing and characterization of the FVB/NJ mouse genome. *Genome Biology* **13,** 1–12 (2012).

166. Yalcin B., Wong K., Bhomra A., Goodson M., Keane TM., Adams DJ., Flint J. The fine-scale architecture of structural variants in 17 mouse genomes. *Genome Biology* **13,** 1–12 (2012).

167. Doran AG., Wong K., Flint J., Adams DJ., Hunter KW., Keane TM. Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biology* **17,** 1–16 (2016).

168. Srivastava A., Morgan AP., Najarian ML., Sarsani VK., Sigmon JS., Shorter JR., Kashfeen A., McMullan RC., Williams LH., Giusti-Rodríguez P., *et al.* Genomes of the mouse collaborative cross. *Genetics* **206,** 537–556 (2017).

169. Lilue J., Doran AG., Fiddes IT., Abrudan M., Armstrong J., Bennett R., Chow W., Collins J., Collins S., Czechanski A., *et al.* Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature Genetics* **50,** 1574–1583 (2018).

170. Ferraj A., Audano PA., Balachandran P., Czechanski A., Flores JI., Radecki AA., Mosur V., Gordon DS., Walawalkar IA., Eichler EE., *et al.* Resolution of structural variation in diverse

mouse genomes reveals chromatin remodeling due to transposable elements. *bioRxiv* 1–25 (2022). doi:10.1101/2022.09.26.509577

171. Robinson JT., Thorvaldsdóttir H., Winckler W., Guttman M., Lander ES., Getz G., Mesirov JP. Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26 (2011).

172. Kim YM., Lee J-Y., Xia L., Mulvihill JJ., Li S. Trisomy 8: a common finding in mouse embryonic stem (ES) cell lines. *Molecular Cytogenetics* **6,** 1–5 (2013).

173. Gaztelumendi N., Nogués C. Chromosome Instability in mouse Embryonic Stem Cells. *Scientific Reports* **4,** 1–8 (2014).

174. Ji F., Zhu X., Liao H., Ouyang L., Huang Y., Syeda MZ., Ying S. New Era of Mapping and Understanding Common Fragile Sites: An Updated Review on Origin of Chromosome Fragility. *Frontiers in Genetics* **13,** 1–13 (2022).

175. Helmrich A., Stout-Weider K., Matthaei A., Hermann K., Heiden T., Schrock E. Identification of the human/mouse syntenic common fragile site FRA7K/Fra12C1 - Relation of FRA7K and other human common fragile sites on chromosome 7 to evolutionary breakpoints. *International Journal of Cancer* **120,** 48–54 (2007).

176. LeTallec B., Millot GA., Blin ME., Brison O., Dutrillaux B., Debatisse M. Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Reports* **4,** 420–428 (2013).

177. Kumar R., Nagpal G., Kumar V., Usmani SS., Agrawal P., Raghava GPS. HumCFS: A database of fragile sites in human chromosomes. *BMC Genomics* **19,** 1–8 (2019).

178. Andor N., Maley CC., Ji HP. Genomic instability in cancer: Teetering on the limit of tolerance. *Cancer Research* **77,** 2179–2185 (2017).

179. Kiwerska K., Szyfter K. DNA repair in cancer initiation, progression, and therapy—a double-edged sword. *Journal of Applied Genetics* **60,** 329–334 (2019).

180. Lovitt CJ., Shelper TB., Avery VM. Doxorubicin resistance in breast cancer cells is mediated by extracellular matrix proteins. *BMC Cancer* **18,** 1–11 (2018).

181. Langhans SA. Three-dimensional in vitro cell culture models in drug discovery and drug repositioning. *Frontiers in Pharmacology* **9,** 1–14 (2018).

182. Wang X., Yan J., Shen B., Wei G. Integrated Chromatin Accessibility and Transcriptome Landscapes of Doxorubicin-Resistant Breast Cancer Cells. *Frontiers in Cell and Developmental Biology* **9,** 1–18 (2021).

183. Howard GR., Jost TA., Yankeelov TE., Brock A. Quantification of long-term doxorubicin response dynamics in breast cancer cell lines to direct treatment schedules. *PLoS Computational Biology* **18,** 1–26 (2022).

184. Liu CL., Chen MJ., Lin JC., Lin CH., Huang WC., Cheng SP., Chen SN., Chang YC. Doxorubicin promotes migration and invasion of breast cancer cells through the upregulation of the RHOA/MLC pathway. *Journal of Breast Cancer* **22,** 185–195 (2019).

185. Mohammed S., Shamseddine AA., Newcomb B., Chavez RS., Panzner TD., Lee AH., Canals D., Okeoma CM., Clarke CJ., Hannun YA. Sublethal doxorubicin promotes migration and invasion of breast cancer cells: role of Src Family non-receptor tyrosine kinases. *Breast Cancer Research* **23,** 1–20 (2021).

186. Canals D., Salamone S., Santacreu BJ., Aguilar D., Hernandez-Corbacho MJ., Ostermeyer-Fay AG., Greene M., Nemeth E., Haley JD., Obeid LM., *et al.* The doxorubicin-induced cell motility network is under the control of the ceramide-activated protein phosphatase 1 alpha. *FASEB Journal* **35,** 1–17 (2021).

187. Klaasen SJ., Kops GJPL. Chromosome Inequality: Causes and Consequences of Non-Random Segregation Errors in Mitosis and Meiosis. *Cells* **11,** 1–14 (2022).

188. Ben-David U., Amon A. Context is everything: aneuploidy in cancer. *Nature Reviews Genetics* **21,** 44–62 (2020).

189. Kuwano A., Sugio Y., Murano I., Kajii T. Common fragile sites induced by folate deprivation, BrdU and aphidicolin: Their frequency and distribution in Japanese individuals. *Journal of Human Genetics* **33,** 355–364 (1988).

190. van Wietmarschen N., Lansdorp PM. Bromodeoxyuridine does not contribute to sister chromatid exchange events in normal or Bloom syndrome cells. *Nucleic Acids Research* **44,** 6787–6793 (2016).

191. Sun S., Osterman MD., Li M. Tissue specificity of DNA damage response and tumorigenesis. *Cancer Biology and Medicine* **16,** 396–414 (2019).

192. Kim H., Casey AE., Palomero L., Aliar K., Parsons M., Narala S., Mateo F., Hofer S., Kislinger T., Pujana MA., *et al.* Mammary lineage dictates homologous recombination repair and PARP inhibitor vulnerability. *bioRxiv* 1–77 (2021). doi:10.1101/2021.05.14.444217

193. Biau J., Chautard E., Verrelle P., Dutreix M. Altering DNA repair to improve radiation therapy: Specific and multiple pathway targeting. *Frontiers in Oncology* **9,** 1–10 (2019).

194. Wihlm J., Limacher JM., Levêque D., Duclos B., Dufour P., Bergerat JP., Methlin G. [Pharmacokinetic profile of high-dose doxorubicin administered during a 6 h intravenous infusion in breast cancer patients]. *Bulletin du cancer* **84,** 603–8 (1997).

195. Harahap Y., Ardiningsih P., Winarti AC., Purwanto DJ. Analysis of the doxorubicin and doxorubicinol in the plasma of breast cancer patients for monitoring the toxicity of doxorubicin. *Drug Design, Development and Therapy* **14,** 3469–3475 (2020).

196. Schindelin J., Arganda-Carreras I., Frise E., Kaynig V., Longair M., Pietzsch T., Preibisch S., Rueden C., Saalfeld S., Schmid B., *et al.* Fiji: An open-source platform for biological-image analysis. *Nature Methods* **9,** 676–682 (2012).

197. Dobin A., Davis CA., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras TR. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

198. Gel B., Díez-Villanueva A., Serra E., Buschbeck M., Peinado MA., Malinverni R. RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32,** 289–291 (2016).