The development of COIN-seq: A method to explore molecular synergy through combinatorial interventions in breast cancer

Eftyhios Kirbizakis

A Thesis
in
The Department
of Biology

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Science (Biology) at
Concordia University
Montreal, Quebec, Canada

November 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Eftyhios Kirbizakis

Entitled: The development of COIN-seq: a method to explore molecular synergy through combinatorial perturbation in breast cancer

and submitted in partial fulfilment of the requirements for the degree of

Master of Science (Biology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair

Dr. Selvadurai Dayanandan

_____ Examiner

Dr. Sylvie Mader

_____ Examiner

Dr. Aashiq Kachroo

_____Supervisor

Dr. Michael Hallett

Approved by _____

Dr. Robert Weladji, Graduate Program director

_____2022 _____

Dr. Pascale Sicotte, Dean of Faculty

Abstract

The development of COIN-seq: A method to explore molecular synergy through combinatorial perturbation in breast cancer

Eftyhios Kirbizakis

Breast cancer is a heterogeneous disease that has been extensively profiled by high throughput technologies and can be classified into different subtypes. However, the underlying causes of subtype differentiation remain unclear. Traditional experimental approaches to understand causality in breast cancer subtype differentiation are limited. This is due to the correlative nature of the data produced from profiling biological systems without perturbations, the use of low throughput single gene or gene product perturbation systems, and the time and resource bottleneck associated with traditional experimental approaches. To address these problems, we designed and implemented COIN-seq, a pooled COmbinatorial INtervention sequencing (COIN-seq) screen. COIN-seq combines single cell RNA sequencing and clustered regularly interspaced short palindromic repeats (CRISPR) based genetic interventions to perform massively parallel multi-locus gene perturbations. I contribute to the development of COIN-seq by aiding in the design and implementation of the different components involved with the system. In this thesis, we begin by describing the background information on breast cancer and the biological targets. We then explore the different technologies COIN-seq is predicated upon as well as the overall design and implementation of the system. Lastly, we discuss the successes and failures involved in this development process. Although a full implementation of this method was not achieved here, the successes and failures reported in this thesis can nevertheless serve as a guide for future development.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Michael Hallett, for giving me the opportunity of being both a student and member of his lab. Under his guidance and support, I have learned more than I ever thought possible. I am forever grateful. Thank you, Mike. I would also like to thank my committee members Dr. Sylvie Mader and Dr. Aashiq Kachroo for their never-ending support and contribution throughout this project. Furthermore, I thank Dr. Vanessa Dumeaux, for her invaluable help throughout the entirety of the project.

I would also like to thank the members of the Hallett Lab for their constant support. Without them, none of this would be possible. I am proud not only to have you all on my team, but to call you my friends. Thank you for your help Sanny Khurdia, Abdelrahman Ahmed, Van Bettauer, Shawn Simpson, and Samira Massahi. I look forward to all the great things you will all accomplish.

Similarly, I would like to extend my thanks to the members of the Mader lab, who took me under their wing when I needed it the most. Thank you, Leanna Canapi, Madline Sauvage, Salwa Haidar, Elham Dianati Ajibisheh, and Lucas Porras for all of your help and for supporting me through my never-ending stream of questions.

Finally, I would like to thank those closest to me. My parents and siblings for their unwavering presence and support from the beginning, my partner, Nathalie Jreidini, who has been my rock throughout this whole experience, and my friends, who do not have the background to understand what I do but endured my ceaseless jabbering out of love. I dedicate this work to all of you.

Contribution of Authors

The table below describes the contributions of each member of this project. Mammalian cell culture training was provided by L Canapi, M Sauvage, and S Haidar (UdeM - IRIC). During this time, they assisted with the generation of cell lines used and preparation of samples. S Simpson (Concordia) assisted with the construction of the Poseidon single cell sequencing device. S Khurdia (Concordia) implemented and optimized the microparticle modification and Drop-seq based single cell sequencing protocols used throughout the project. A Ahmed (Concordia) aided in constructing plasmid vectors used, sample preparation, and single cell sequencing. Samples were sequenced by the IRIC Genomics Platform. V Dumeaux (Concordia) assisted with our analytic pipelines. M Hallett contributed to the analyses and manuscript preparation. M Hallett and S Mader (UdeM - IRIC) designed, supervised, and funded the project.

Name	Cell line prep.	Poseidon construction	Sample prep.	Microparticle prep.	Single cell sequencing	Analysis	Manuscript prep.	Funding
Kirbizakis								
Khurdia								
Ahmed								
Simpson								
Canapi								
Sauvage								
Haidar								
Dumeaux								
Mader								
Hallett								

Table of Contents

Lists of Figures

List of Tables

List of Abbreviations

AP2γ	Activating enhancer-binding protein 2γ
BC	Breast cancer
BM	Basement membrane
bZIP	Basic region leucine zipper
Cas	CRISPR associated
CBC	Cell barcode
COIN-seq	COmbinatorial INtervention sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference
crRNA	CRISPR-RNA
DBD	DNA binding domain
dCas9	Dead Cas9
DHT	Dihydrotestosterone
dRNA	Dead RNA
E2	17β-estradiol
ECM	Extracellular matrix
ER	Estrogen receptor
ERα	estrogen receptor alpha
ERE	Estrogen response element
ERS	Endoplasmic reticulum stress
FHD	Forkhead DNA-binding domain
FOX	Forkhead box
FOXA1	Forkhead box A1
GATA3	GATA binding protein 3
GBC	Guide barcode

GH	Growth hormone
HER2	Human epidermal growth factor 2
KO	Knockout
LBD	Ligand binding domain
MaSc	Mammary stem cell
MHC	Major histocompatibility complex
MOI	Multiplicity of infection
MPH	MS2-p65-HSF1 fusion protein
NAC	NIpple areola complex
NHEJ	Nonhomologous end-joining
NGS	Next-generation sequencing
NLS	Nuclear localization signal
NTD	Amino-terminal domain
OE	Overexpression
PAM	Protospacer adjacent motif
PAM50	Prediction Analysis of Microarray 50
PDMS	Polydimethylsiloxane
PR	Progesterone receptor
SAM	Synergistic Activation Mediator
sgRNA	Single guide RNA
STAMPs	Single-cell transcriptomes attached to microparticles
TAD	Transactivation domain
TEB	Terminal end buds
TracrRNA	Transactivating crRNA
TF	Transcription factor
UMI	Unique molecular identifier
UPR	Unfolded protein response

Chapter 1. Introduction

Breast cancer (BC) is the most commonly diagnosed cancer in women, contributing to 13% of all cancer deaths. The prevalence and death rate of BC carry a significant socioeconomic burden, as BC creates an enormous pressure on individuals and the global health care system as a consequence of expensive targeted therapies, long-term chemotherapies, and lifelong follow-ups[1]. The Canadian Cancer Society estimates that 1 in 8 Canadian women will develop invasive BC at some point in their lifetime, and 1 in 33 will die from it[2]. The primary risk factors of developing BC are being a woman and age[3]**.** Nulliparity is also a major risk factor, which is defined as a woman who has never given birth to a child. Nulliparous women have a 20-40% higher risk of postmenopausal BC than parous women who gave birth before the age of 25[4]. In addition, a third trimester pregnancy can act as a prophylactic measure against BC.

The most common invasive BC histology is ductal carcinoma (IDC), which spreads from the duct of the ipsilateral breast into the parenchyma. IDC accounts for 50-75% of BCs, whereas invasive lobular carcinoma represents 5-15% of BCs, and other rarer histologies comprise the remaining patients[5]. Other related diseases include ductal carcinoma in situ (DCIS), the most common type of non-invasive BC. DCIS is limited to the lumen of the duct, as it has not breached the basement membrane[6]. DCIS is not life threatening, but is a non-obligate precursor of IDC that can increase the likelihood of developing invasive BCs.

It is long established that invasive BC is a heterogeneous disease, where individual tumors differ in many dimensions including their morphology, pathology, and molecular profiles[7]. For over half a century, most tumours have been classified into subtypes defined by the expression of just two or three proteins: Estrogen Receptor α (ERα), Progesterone

Receptor (PGR), and the Human Epidermal Growth Factor 2 (HER2) encoded by genes *ESR1*, *PGR* and *ERBB2* respectively. *ERBB2* is an acronym for Erb-B2 Receptor Tyrosine Kinase 2 and is the official gene name for HER2[8]. The name stems from *NEU*, a rat homologue of *ERBB2*, which was originally identified from chemically induced neuroblastomas and shown to be homologous to a retroviral oncogene (*v-ERBB*)[8]. Tumour expression of ERα and amplification of the *ERBB2* gene are denoted as estrogen receptor (ER) positive (ER+) and HER2-positive (HER2+). These are used to define the four so-called clinical or molecular subtypes representing distinct forms of the disease based on their status; ER+/HER-, ER+/HER2+, ER-/HER2+, ER-/HER-[9,10].

Prognosis and treatment modalities for BC have traditionally been determined using standardized clinicopathologic criteria. Clinicopathological parameters such as size of the tumour, histological grade, stage, and lymph node infiltrate can be used to classify BCs into biologically and clinically meaningful subgroups[11]. Cancers are classified by stage, which describes the extent of the spread of the cancer and the amount of tumour. The most common staging system applied to most solid tumours is the so-called Tumour, Node and Metastasis (TNM) system[12]. Tumour describes the size of the main tumour and if it has grown into other parts of the organ. Node describes if, and how much, the tumour has spread to lymph nodes. Metastasis describes if the tumour has spread to other parts of the body through the blood or lymphatic system. Early-stage cancers are localized and non-invasive whereas later stages have metastasized to other organs of the body[13]. Estimates of both patient prognosis and treatment regiments are influenced by many factors, including the stage of their disease. Generally, early-stage BC is treated with surgery and/or radiation and has a more favourable prognosis, mid-stage invasive BC is treated with surgery (mastectomy or lumpectomy) and/or chemotherapy and/or directed therapy, and late-stage BC that has metastasized is often

associated with a less favourable prognosis as the disease is treatable with hormone therapy/ chemotherapy, but incurable[14].

Molecular predictive markers are characteristics that are objectively measured and evaluated as indicators of normal biologic processes, pathogenic processes, or pharmacologic responses to therapeutic intervention[15]. These predictive markers can be genes or gene products that are able to measure whether an individual will benefit from a specific therapy. An example of this is the relationship between ERα and tumours of the luminal subtype. These tumours are defined as IDCs that are ER positive; at least initially these tumours are believed to be driven by the high expression of the ERα[16]. Such tumours are treated using anti-aromatase therapy to stop the production of estrogen or endocrine therapy agents, such as Tamoxifen™, which compete with estrogen in binding with ERα binding at the cell surface[17]. ER- tumours with amplification of *ERBB2* are categorized as HER2+ and treated with anti-*ERBB2* targeted therapies such as Herceptin®. Triple-negative tumours are characterized by the absence of ER, HER2, and PR and lack targeted therapies. They are primarily treated with chemotherapy although several directed therapies are in development. Some of these new therapies target the formation of new blood vessels (angiogenesis) such as Avastin®[18]. Lastly, the PR status indicates if tumour growth is influenced by the presence of the hormone progesterone. If present, PR can act as a target for endocrine therapies.

BC genomics and informatics have generated more refined subtype classifications. High throughput profiling assays refined the notion that multiple forms of the disease can exist. Microarray-based profiling introduced a new paradigm in which a diverse subset of BC subtypes can be identified through their different genetic signatures[10]. These differences reinforced the notion that BC is a group of disease by highlighting the biological heterogeneity found within and facilitated the development of new subtyping schemes. There now exist

several different BC subtyping schemes in the literature with varying levels of prognostic and predictive statistical capacity[19]. This includes the intrinsic subtype classification scheme based on microarray-based profiling from Perou[10], Sorlie[20], the PAM50 subtypes[16], the Cartes d'identité des tumeurs (CIT) subtypes[21]; a triple-negative specific scheme[22], and others. Each subtyping scheme partitions tumours into a set of distinct subtypes. For example, the subtypes for the PAM50 subtyping scheme are luminal A, luminal B, normal-like, basal-like; which lack ER, PR expression and HER2 amplification, similarly to the triple-negative subtype, but have changes in other proteins that the triple-negative subtype does not have[16]. For the CIT scheme, the subtypes are luminal A, luminal B, luminal C, normal-like, basal-like, and molecular-apocrine[21]. In this scheme, there is no subtype defined by HER2 amplification status, although most tumours with HER2 amplification are in the molecular-apocrine subtype. While there are some differences between the schemes, they all generally identify one or more classes that correspond to HER2+ tumours as well as one or more subtypes that correspond to the luminal (ER+) tumours. In addition, statistical and machine learning algorithms have been developed to classify tumours by each such subtyping scheme. Given a molecular profile (e.g. RNA-sequencing) of a tumour at time of diagnosis, the computational challenge is to classify each tumour with high statistical confidence into one subtype using specific marker genes and gene products. The PAM50 subtyping scheme (**Figure 1**), which is widely used, examines the expression of 50 genes that poll different biological processes[23].

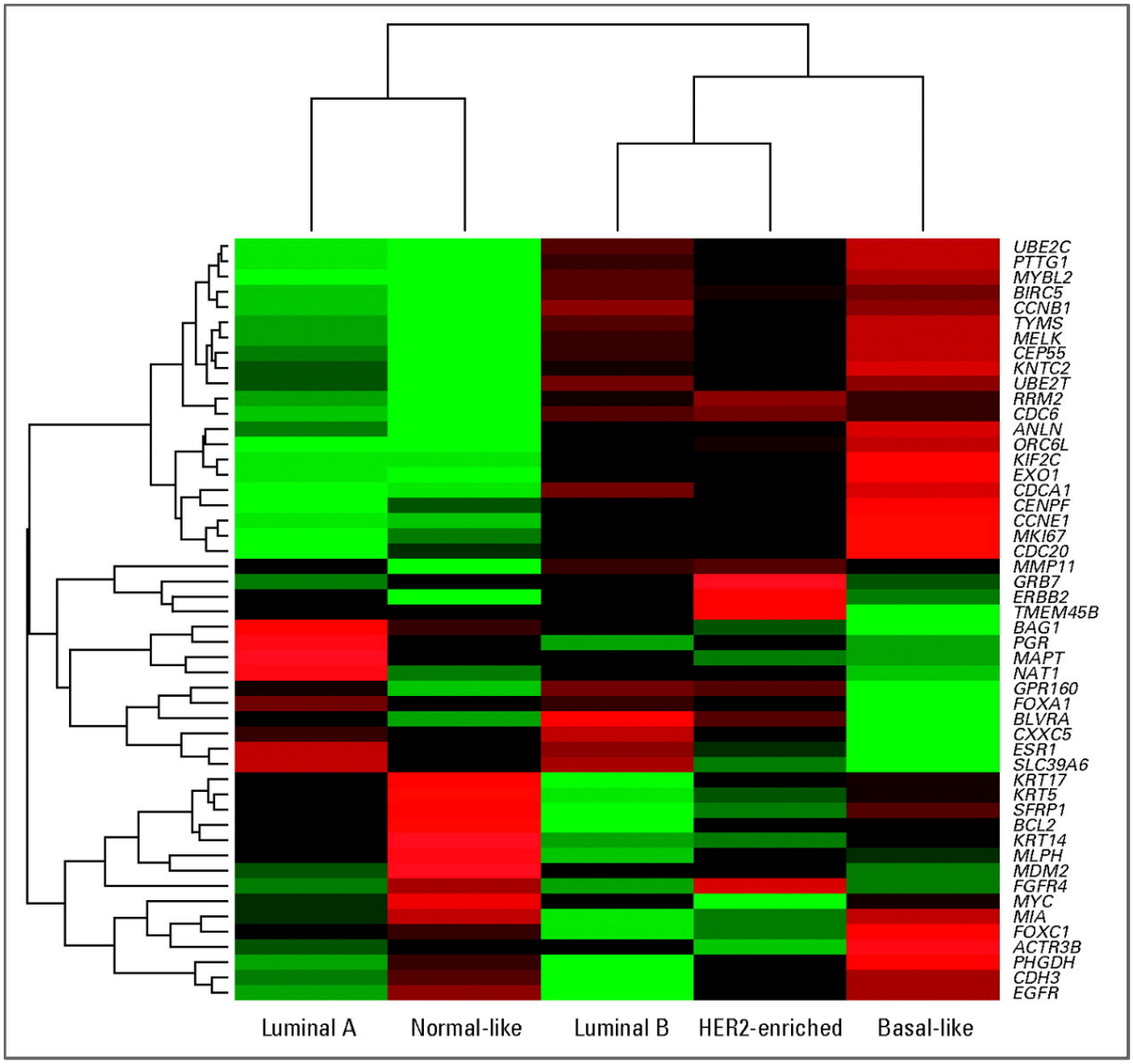


Figure 1. Original heatmap of PAM50 gene set. The heatmap classifies the genetic signatures of 50 genes into distinct BC subtypes. Transcriptomes generated using microarrays from bulk tissue. Adapted from Parker et al. (2009)[16].

The clinical utility of subtyping tumours in this manner arises from their ability to provide both an estimate of patient prognosis and sometimes an estimate of patient benefit to a specific therapy. Individual patients with the same tumour subtype can have a different prognosis. The prognostic endpoint is the ability to distinguish between those tumours with a molecular composition likely to drive progression to metastatic state (poor patient outcome) and those tumours that are inherently more indolent at the molecular level. The clinical utility of classifiers with predictive insight is that they are able to identify chemo-, endocrine, and

targeted therapies the patient will likely benefit from by estimating whether a given tumour will respond or not to a specific treatment[24]. In this case, knowledge of subtype provides insight into what drugs or other treatment modalities that will likely ablate tumour progression. For example, a patient diagnosed with luminal A will most likely be treated with Tamoxifen(™) because a 5-year Tamoxifen therapy regimen has shown to reduce BC mortality rate by one third throughout the first 15 years[25]. The benefit of Tamoxifen therapy differs from patients diagnosed with luminal B, where Tamoxifen confers a short-term benefit that attenuates over time[109]. Therefore, using genomic-based subtyping schemes provides clinical utility beyond the classic clinical subtypes.

Genomics and informatics regarding existing BC subtypes have been correlative. This thesis is predicated in part on the observation that BC subtyping schemes to date are correlative in nature: the individual genes that are used in informatics classifiers are markers of tumour subtype. Specifically, the genes that are chosen to classify a tumour into one of several subtypes within a specific subtyping scheme have expression patterns that tightly co-vary with the subtype[20]. Informally, marker genes will have strong differential expression between tumours that belong to the subtype and those tumours that do not belong to the subtype. This statement must be amended to address the fact that modern subtyping schemes do not rely on a single gene but a gene set, meaning that it is their joint distribution of expression that is used to classify a tumour. Mathematically this means that the joint distribution of the gene set must co-vary strongly with the tumour subtype, although it is possible that no individual member of this set has such strong correlation. However, in practice, the individual genes in the gene set act as independent noisy indicators and the multivariate statistics help to stabilize the consensus predictions. This is analogous to classic examples of noisy fire alarms used in machine learning[26]. In summary, the genes that are used to classify subtype are correlated with subtype and cannot be said to cause the subtype. This

makes correlation between subtype and gene expression informative but limited as it does not provide insight into the underlying biology taking place within the system. However, by applying a causal approach, we can better understand the dynamics of these gene sets. For example, if we believe that the expression of a set of genes plays a significant role in tumour subtype, we can modify the expression of that set to observe the possible change in the genes they regulate. This can be used to better understand the relationship between them and the gene set they are a part of with respect to their influence on the tumour subtype. This thesis focuses on the development of a genomic-information technology that identifies genes that are causal in the development of BC subtypes.

Our understanding of the causative agents controlling BC subtype differentiation remains incomplete. Many researchers have speculated that tumour subtypes may represent transformation of stem cells with arrest at specific stages of development or alternatively, direct transformation of various mature cell types. The works of Prat, Perou and colleagues focus on links between normal mammary development and the BC subtype differentiation process[27] (**Figure 2**). Normal mammary development follows a hierarchy from least to most differentiated. The general mechanisms involved in differentiation are not clearly understood as the underlying genes and gene products controlling cell fate remain largely undetermined. With respect to this hierarchy, this model suggests that undifferentiated mammary stem cells (MaSCs) which incur a genomic insult (damage) begin to generate cellular diversity. Such a model supports the cancer stem cell (CSC) hypothesis that cancer can arise from transformation of a normal stem cell or progenitor cell, thus giving rise to a heterogeneous population of cells. In this case, the bulk of the tumour is composed of differentiated cells with limited proliferative potential, where the CSCs maintains the tumour[27].

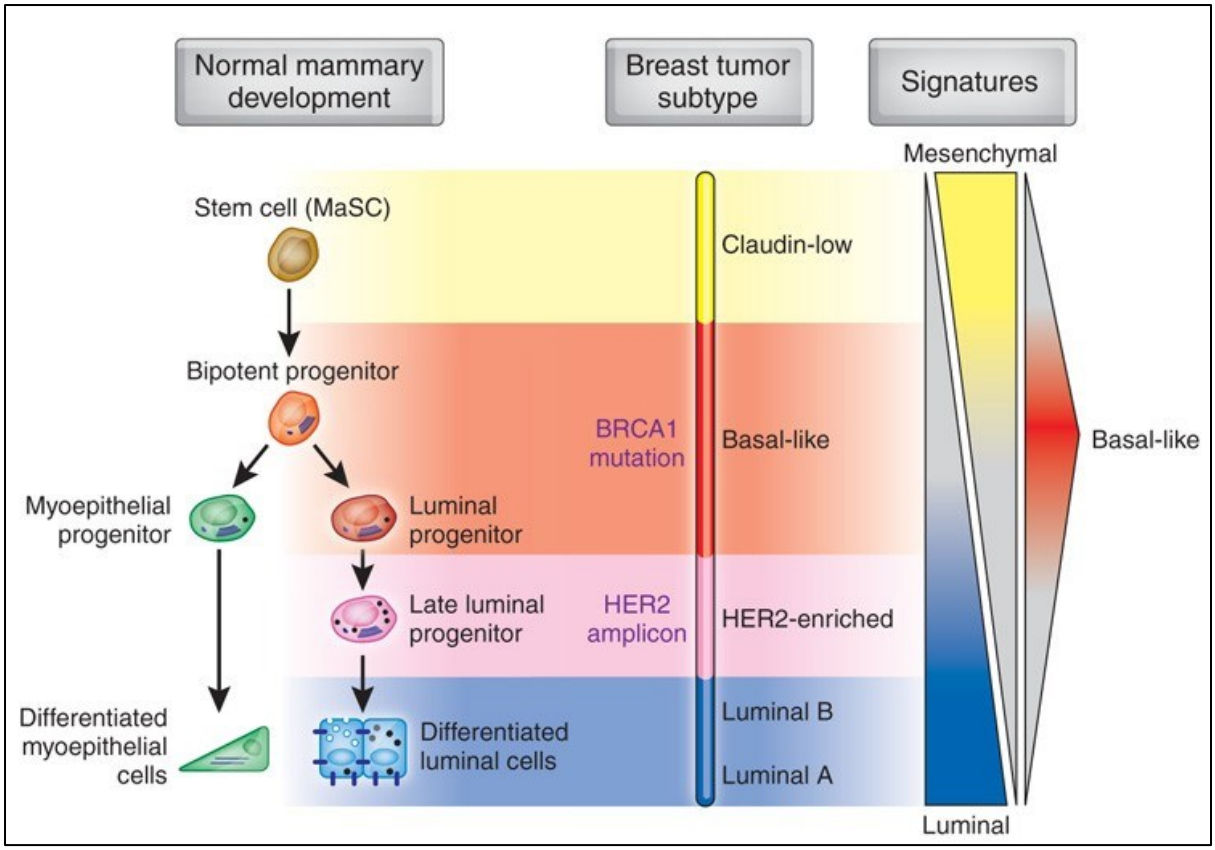


Figure 2. Putative relationships between stage of epithelial cell differentiation and BC subtype. Subpopulations of normal breast tissue and potential cells of origin for intrinsic BC subtypes. Adapted from Prat and Perou (2009)[27].

The underlying genes and gene products controlling subtype fate remain undetermined. There has been a sustained and large-scale effort to identify genes, gene products and other genomic elements that are involved in breast tumourigenesis and subtype differentiation[28]. This includes genes highlighted in **Figure 2** depicting the PAM50 gene set whose gene expression is correlated with different stages of differentiation[21,29,30]. There has been a sustained interest in the identification of the proteins involved in subtype differentiation with emphasis on the transcription factors (TFs) controlling luminal subtype differentiation[31]. More specifically, the focus has been on a set of six TFs, *ESR1*, *FOXA1*, *GATA3*, *SPDEF*, *AR*, and *XBP1*. In this thesis, we develop a system for identifying molecular determinants of BC subtype for three of these TFs, which we have termed "*the canonical three*" (*ESR1*, *FOXA1*, *GATA3*).

Traditional experimental approaches to understand causality in BC subtype differentiation are limited. Causal experimental approaches are a type of genetic screen, where a genetic perturbation/intervention is introduced to a biological system to elicit a response and the resulting gene signature is the response. There exist many genetic perturbation methodologies, but the general goal among them is that the perturbation is introduced to a biological system and influences the expression of a target gene. The expression of the gene can be altered through a targeted deletion or knockout (KO) in which the gene would not be expressed at all. Alternatively, the gene can be knocked down (KD) or over-expressed (OE), where the expression of the gene is lower or higher, respectively, than its basal expression level due to the influence of the perturbation. In previous works, the Mader lab has subjected *the canonical three* TFs to traditional univariate siRNA knock-down and bulk RNA-sequencing (RNA-seq). The resultant single transcript knock-down data was used to infer a putative causal network that explains the relationship between the perturbation and the resulting gene signature[31].

This thesis is motivated by the following challenges regarding the state of the art with respect to understanding complex biological processes:

1. As described above, profiling of biological systems without perturbations provides correlative data that cannot be directly used to break cause-correlation boundaries, a fundamental observation underlying almost all of science;

2. Current approaches typically perturb biological systems one gene (or gene product) at a time. Genes generally function as part of a network where the change in one affects the others in some way. For example, when a gene is deleted because of a perturbation, the

resulting signature would give insight into how that gene affected others through changes in their expression. More specifically, if we perturb gene X, and profile downstream survivors, we can call the observed expression pattern the signature for the perturbation of X. The signature functions as an observable phenotype and will have some genes that are over-expressed and under-expressed as a consequence of the gene modification caused by the introduction perturbation X. In addition, we can perturb genes in different ways. For example, we could KO, KD, or OE the gene X by changing the type of perturbation, which would lead to different signatures. Suppose we perturb two distinct genes (X and Y) and consider their signatures S_X and S_Y. If these perturbations occurred individually in separate cells, the signatures would represent the observable independent phenotype associated to each perturbation. However, if these perturbations occurred in the same cell, we would be able to observe molecular epistasis. Epistasis refers to when the resulting gene signature of two or more perturbations is different from the sum of the perturbations. In this example, this would mean that we do not expect to observe an additive change in differentially expressed genes for signatures S_X and S_Y in the cell. Instead, we would observe a different signature as a result of the combinatorial effects of those perturbations. For example, we could observe gene synergy, where the resulting signature is greater than the additive effect of signatures S_X and S_Y. These perturbations may also lead to synthetic lethality, where the cells challenged by modification of either gene X or Y alone does not result in cell death but when challenged by both modifications simultaneously leads to cell death. Therefore, there is a need to be able to distinguish these situations, which cannot be done using single gene deletions alone.

3. Time and resources constrain the total number of such single gene studies. Typical gene studies involve introducing perturbations to biological systems in a low throughput manner and labour, time, and resources are natural bottlenecks. For example, to explore the individual and combinatorial effects of six distinct gene knockouts, we would have $2^6 = 64$

possible combinations. (**Figure 3)**. To generate these combinations, this translates to 64 separate experiments, a labour-intensive effort. In the context of traditional bulk RNA-seq, a comparison between the combinations would require the 64 experiments are sequenced individually, which implies a significant cost.

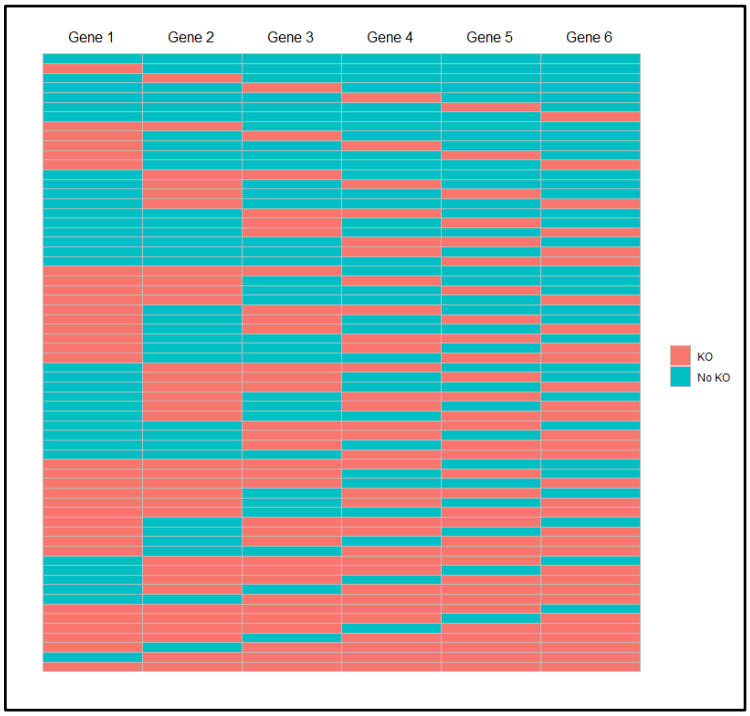


Figure 3. Possible KO combinations of six genes. Visualization of all possible KO combinations and permutations when targeting 6 genes. There are a total of 2^6 = 64 possible combinations.

Genetic screens help infer gene function, but face difficulty when used to assay complex gene interactions. Recently, several groups developed approaches to address issues 1-3 discussed above. The back-to-back papers from Dixit and colleagues and Adamson and colleagues introduced Perturb-seq, a novel high throughput method of performing pooled genetic perturbation screens[32,33]. Perturb-seq allows the user to deliver multiplexed gene inactivation perturbations to a target population of cells. This is done by first designing clustered regularly interspaced short palindromic repeats (CRISPR) based perturbations, that are packaged into lentivirus particles used to infect mammalian cells. Then, the particles are delivered to a population of cells. The delivery of these particles can be modulated by tuning

the amount of lentiviral particles being delivered, allowing for stochastic control over the expected number of times cells are infected by different perturbations. This means that the user can introduce the perturbations in a way that more than one perturbation can be present in a cell. The gene expression of these cells are then profiled using droplet-based single cell RNA sequencing, resulting in a unique signature for each individual and combination of perturbations in a single experiment. Therefore, Perturb-seq offers a solution to the aforementioned issues. 1. The use of genetic perturbations allows for the exploration of the cause-correlation boundary. 2. The observation of molecular epistasis is enabled through the delivery of combinatorial perturbations. 3. The natural bottlenecks involved in perturbation-based studies are removed with a well-designed Perturb-seq experiment that can result in the molecular profiles of every desired perturbation combination.

Furthermore, gene intervention techniques have evolved beyond Perturb-seq to improve or modify its mode of function. For example, CRISP-seq is a method functionally similar to Perturb-seq in which the perturbations are prepared and introduced using a similar method. However, instead of using a droplet-based single cell RNA sequencing, CRISP-seq uses a lower throughput FACS-sorted microwell-plate strategy[105]. Additionally, CROP-seq is a newer method that improves the CRISPR-based perturbation vector used in Perturb-seq to facilitate cloning, which decreases time and complexity of the method and facilitates larger screens[106]. There are also additional techniques such as Direct capture Perturb-seq and DART-seq, that we explore in more detail in chapters 3-5.

To address these limitations, we develop COmbinatorial INtervention sequencing (COIN-seq). The aim of COIN-seq is to build upon Perturb-seq to allow the simultaneous delivery of KO and OE intervention types to a population of cells[34,35]. Similarly to Perturb-seq, COIN-seq attempts to modify the expression of multiple target genes using the modularity of

CRISPR based genetic screening system through lentiviral delivery. More specifically, this system is designed so that the KO and OE interventions can be performed in the same system against multiple targets. This means that in a single COIN-seq experiment, the user can perturb multiple gene targets and observe their individual and combinatorial effects. Lastly, COIN-seq is designed to use single cell RNA-seq. This modern approach tackles the limitations imposed by bulk RNA-seq, as it allows us to examine the transcriptomes from individual cells in a cell population. In relation to the aforementioned issue 3, COIN-seq would allow the user to generate the profiles of all possible KO combinations using multiplexed CRISPR delivery and single cell sequencing in a single experiment, minimizing costs, labour, and time. In the context of our domain of interest, we apply COIN-seq to the luminal BC cell line MCF7 to target the canonical luminal TFs. In particular, here we have focused on the *canonical three* TFs (*ESR1*, *FOXA1* and *GATA3*) that we believe are causal to subtype differentiation and subject them to a combination of KO and OE interventions concomitantly. This design highlights the state of the art of the COIN-seq methodology in regard to the techniques used and relevant biology.

Analysis of the result data presents a significant computational challenge. The data generated using single cell RNA-seq is noisier and more complex to interpret than other methods. Traditional univariate models are simpler to implement as they characterize one variable (e.g. univariate linear regressions). However, the data generated using our intervention method contains a multivariate set of measurements from a large number of single cells. Due to the size and complexity of these data sets, they require additional computational effort to analyze. The development of bioinformatic and data science pipelines as well as novel deep learning approaches for inferring gene causal networks based on multivariate COIN-seq data is the central focus of the Hallett lab. This enables the observation

of high-order epistatic interactions caused by each set of interventions by building multivariate models to explain these interactions.

This work focuses on establishing the technical efficacy of the system. A working system could then be used to infer a more refined model of luminal subtype differentiation. This thesis highlights my contributions to the design of the COIN-seq system as well as performing the first iteration of experiments, which were unsuccessful. Chapter 2 provides an overview of breast biology before providing a review of the molecular basis of breast cancer subtypes. This includes a more detailed description of several TFs of direct relevance to the development of COIN-seq. Chapter 3 provides a literature review on the genomic technologies that COIN-seq is predicated upon. Chapter 4 introduces the enabling technologies and how they work. Chapter 5 and 6 highlight the design choices made and the results of the first COIN-seq experiments. Lastly, in Chapter 8 we discuss the results and what improvements are to be made in future iterations of COIN-seq. We also include a supplemental materials and methods section containing in detail the methods used throughout these experiments and acts as a reference for future experimentation.

Chapter 2. Background: The breast and breast cancer

Breasts are epidermal appendages that form at the base of the pectoralis major muscles and are anchored by various flexible ligaments[37]. In mammals, the breasts are located on the upper ventral region of the torso. In females, they serve as the mammary glands which produce and secrete milk for the infant.

2.1 Breast development and cell structure

The breast is a dynamic environment, where the morphology changes as a function of multiple factors, such as time, age, puberty, and pregnancy status. On the exterior of the breast lies the nipple areola complex (NAC), composed of the areola and the nipple. The areola has a round shape and varies in size. The nipple is a papillar cylindrically shaped protrusion that lies in the center of the areola. Beneath the skin, they are composed of three tissue compartments: fibrous, glandular, and connective tissues. The fibrous framework of the breast is composed of suspensory ligaments called Cooper's ligaments, providing shape and structural integrity (**Figure 4, a**).

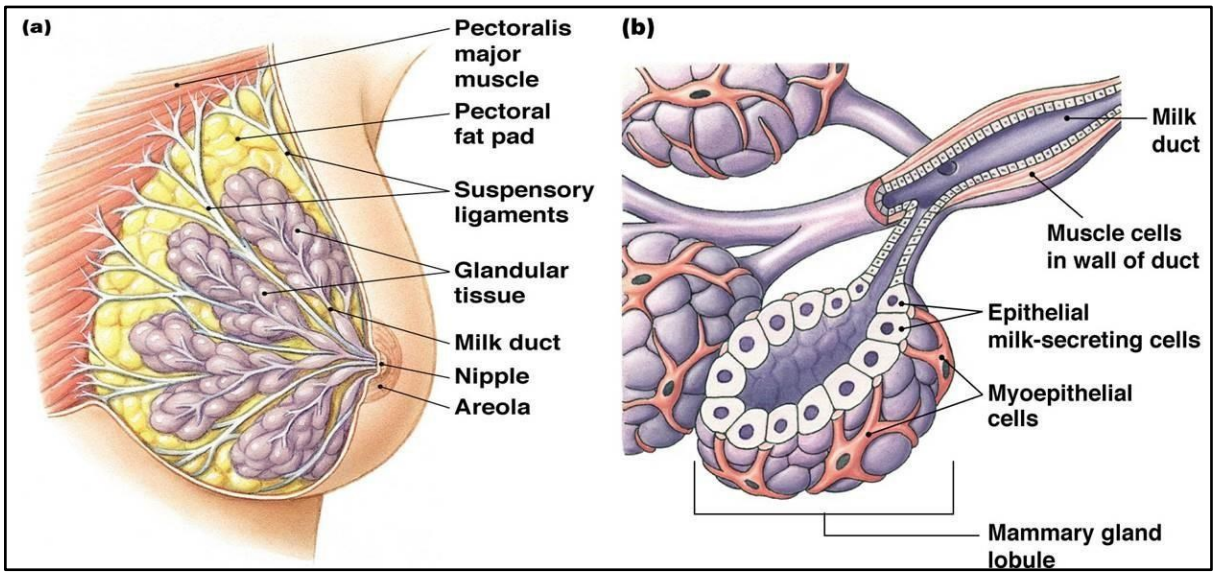


Figure 4. Mammary gland histology. a) Anatomy of the mammary gland. b) Anatomy of a Mammary lobule. Adapted from Human biology online lab[38].

The glandular tissue is composed of milk producing lobules consisting of clusters of alveoli[37]. These lobules are connected to small ducts, which connect to larger milk ducts that connect to the nipple. Histologically, the glandular tissue contains mammary secretory epithelial cells. They are composed of two main cell lineages: luminal secretory cells, which line the lumen of the alveoli and secrete milk, and basal myoepithelial cells, which form a layer around the luminal cells and mediate milk let-down through oxytocin-mediated contractions of the present smooth muscle actin[39] (**Figure 4, b**). The glandular tissue is surrounded by a basement membrane (BM), which physically separates it from the connective tissue compartment (**Figure 5**).

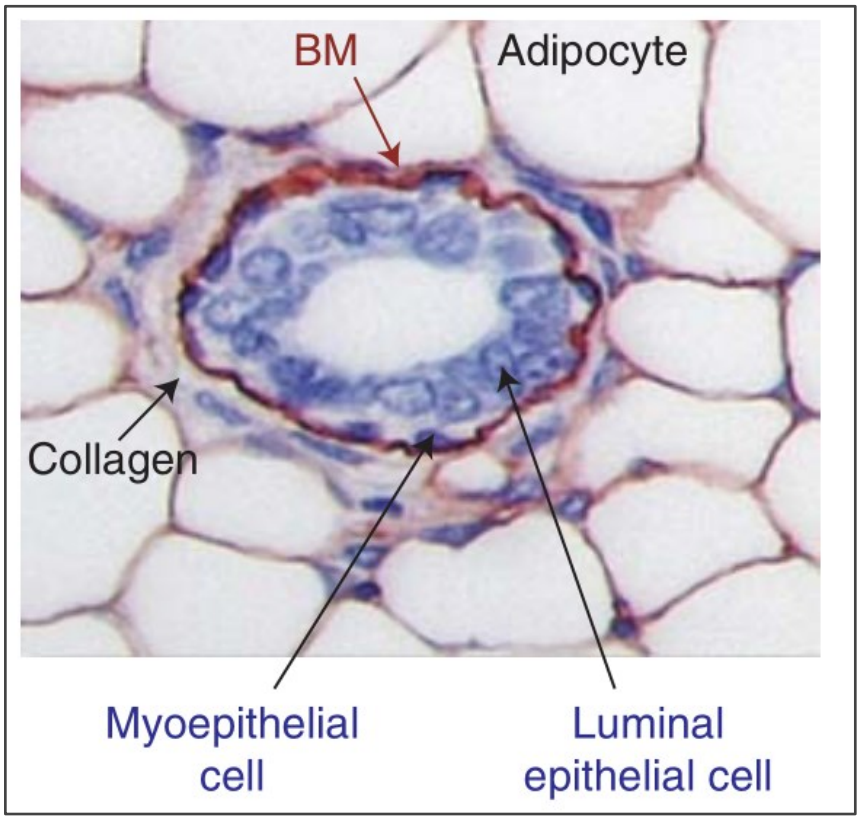


Figure 5. Architecture of breast epithelia. BM surrounds the glandular tissue and separates it from the connective tissue and adipocytes. Adapted from Muschler and Streuli (2010)[40].

The connective tissue compartment is called the stroma or fat pad. The stroma is mainly composed of adipocytes, fibroblasts, and the extracellular matrix (ECM), composed of collagens, laminins, fibronectin, glycoprotein, and proteoglycans. Adipocytes are fat cells that

play a major role in breast development and maturation. The adipose tissue is a major endocrine system that secretes many growth factors, such as estrogen, and plays a pivotal role in epithelial cell differentiation [41]. Fibroblasts are cells that maintain ECM homeostasis and regulate the morphogenesis of normal and tumourigenic mammary glands[42]. Together with adipocytes, they represent the major cellular component of the mammary stroma. The ECM plays an important role in the generation and maintenance of adult tissue by acting as a structural scaffold toward tissue integrity and sustainability[43]. The stroma acts as a support network for the epithelial cells by providing structure, nutrients, blood supply, and immune defenses.

Mouse studies have provided much of our knowledge of mammary gland development because the glands of both species are very similar to each other in structure and function. At birth, the mammary gland exists as a small rudimentary ductal tree. The onset of puberty introduces hormones such as estrogen and growth hormone (GH) to promote cell division and formation of terminal end buds (TEBs). The TEB is a bulbous shaped structure that is unique to the pubertal mammary gland and serves to guide the growth of the ducts through the stroma. While ductal elongation is taking place, the TEBs undergo regular bifurcation events to produce the mammary tree. When the TEBs reach the edge of the stroma, they regress. At this stage, the mammary gland follows a cycle triggered by pregnancy. During pregnancy, hormones such as progesterone and prolactin induce the formation of alveolar buds. The cells of the alveoli differentiate into secretory alveoli which produce milk for the child throughout lactation. When weaning, the majority of the epithelial cells of the secretory alveoli undergo apoptosis in a process called involution. Throughout this process, the gland is remodelled back to a virgin-like state, where the cycle can begin anew, making the mammary gland one of the most regenerative organs in the body (**Figure 6**)[44].

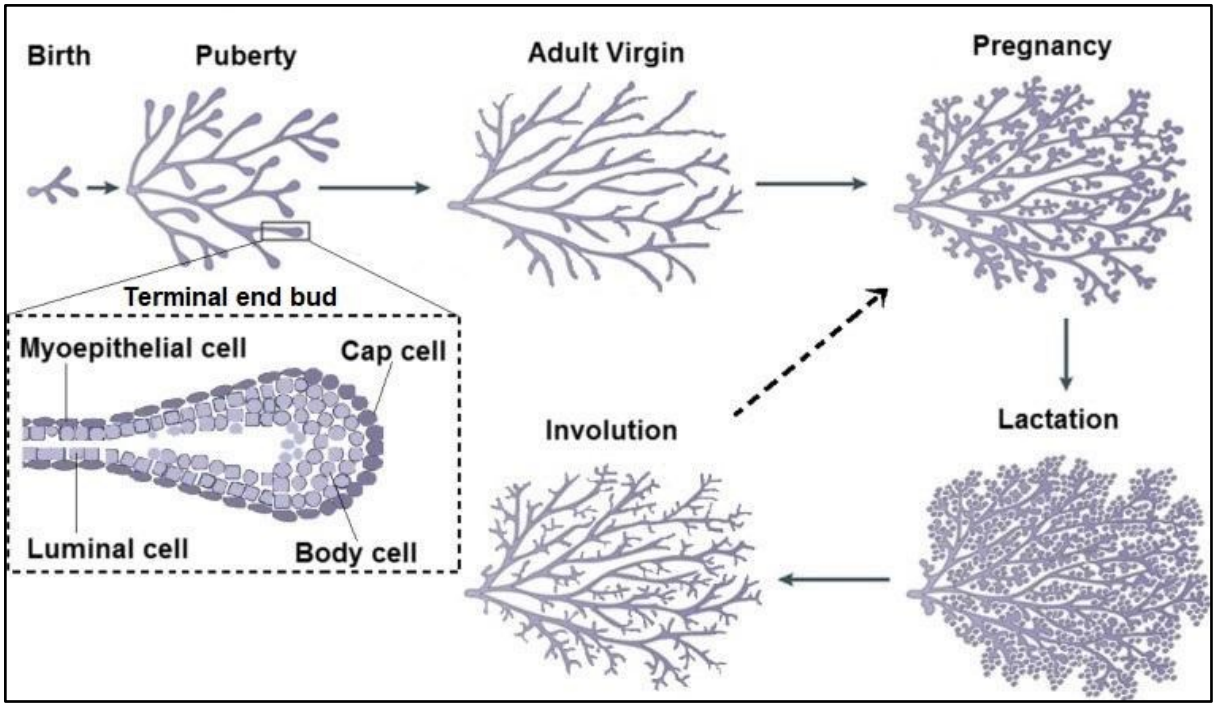


Figure 6. Overview of mammary development. Progression of mammary development from the rudimentary ductal tree at birth to alveolar budding and differentiation at pregnancy. Adapted from Paine and Lewis (2017)[44].

2.2 Breast cancer subtypes

The subtypes of epithelial BCs suggest the existence of multiple cells of origin, where subtype is a consequence of tumour development at different cell stages of development[45]. The human mammary epithelial hierarchy is used as a framework to describe the breast tissue subpopulation that was characterized using *in vitro* differentiation assays (**Figure 2**)[46]. This hierarchy begins with an undifferentiated Multipotent Adult Stem cell (MaSc) that differentiates into committed cell type progenitors. These progenitors differentiate into two distinct lineages: the luminal epithelial cell lineage in which ductal and alveolar cells line the lumen of the mammary gland and the mature myoepithelial cell lineage that surrounds the luminal epithelium and contacts the BM[27]. This framework provides a link between mammary development and tumour profiles. Although we are able to characterize the molecular

signature of what makes up a tumour subtype, there is a lack of understanding of the dynamics of cell type differentiation.

Molecular profiling of BCs stretching back to the turn of the century has refined classification, identifying new subtypes and better characterizing their underlying molecular processes. There are now several subtyping schemes in the literature including PAM50, a subtyping scheme refined from the original works of Therese Sorlie and Chuck Perou[10,20]. However, our understanding of the causative agents controlling BC subtype differentiation remains enigmatic. Several models relate normal mammary development with the subtype differentiation process, where the state of differentiation of the cell where tumourigenesis takes place determines which subtype the tumour will develop (**Figure 2**).

One of the most common subtypes is the luminal subtype, which is characterized by the expression of ER and is dependent on estrogens for growth. Our hypothesis is that subtype differentiation is determined by the *canonical three* genes, *ESR1*, *FOXA1*, *GATA3*. More specifically, we believe that there exists a tightly controlled regulatory network consisting of higher-order multivariate positive and negative interactions associated to these genes that determine subtype. Therefore, we aim to better understand the causal roles of these TFs that play an important role as master regulators of the luminal subtype.

2.3 ERα

ERα is a protein encoded by the *ESR1* gene on chromosome 6q25.1 and has 23 exons, primarily expressed in the uterus, ovary, prostate, testes, and breast, however, it is also present at lower levels in other tissues. The 66.2 kDA protein is 595 amino acids long. *ESR1* is a member of the nuclear hormone receptor superfamily of transcriptional regulators.

There are four principal function domains termed A/B, C, D, E/F (**Figure 7**). Together, the A/B and C domains represent the estrogen response element (ERE)[47,48]. The A/B domain represents the N-amino-terminal domain (NTD) involved in transcriptional activation and contains a zinc-finger that mediates binding and the C domain represents the DNA binding domain (DBD), contributing to estrogen receptor dimerization and binding to specific sequences in the chromatin. The D domain serves as a hinge that connects the C domain to the E domain and serves as a binding site for chaperone protein. In addition, it contains a nuclear localization signal (NLS) that activates when estrogen is bound and translocates the receptor complex into the nucleus[49]. The E/F domain is the ligand binding domain (LBD) that binds estrogen and contains binding sites for coactivators and corepressors. Embedded into the NTD and the DBD are the activation function domains AF-1 and AF-2 respectively. These sites act as additional regulators of estrogen receptor transcriptional activity[50].

There are four known isoforms [51]. Some of the shorter isoforms serve a different function. For example, some of these isoforms do not have an NTD and consequently lack the AF-1 domain, preventing transcriptional activation and instead form heterodimers with the full length ERα.

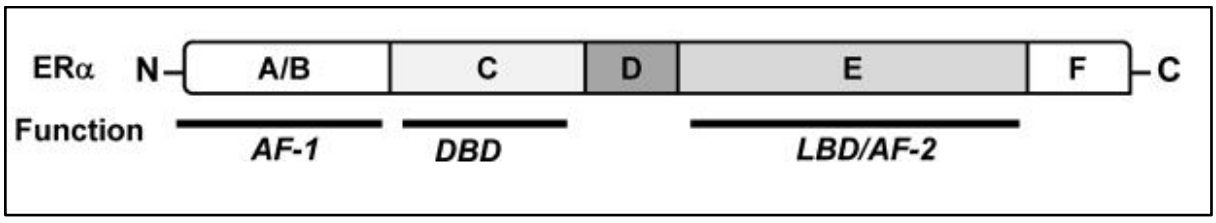


Figure 7. *ESR1* gene structure. Adapted from Cui et al. (2013)[52]

The major transcriptional effects of estrogen are mediated by interactions with nuclear receptors, such as ERα. There are three primary ER signaling mechanism pathways. The first pathway involves ER binding to E2, allowing the complex to translocate into the nucleus. This complex can then regulate the transcription of its target gene either through the direct binding

of the ERE or by binding additional TFs. In the second pathway, ER may act in a ligand-independent manner. ER can be activated by a variety of other factors, such as protein kinases, which phosphorylate the nuclear receptors and activate their transcriptional activity. The last pathway is the membrane-initiated pathway in which ER functions in a ligand-dependent manner. The complex either interacts with other membrane receptors or activates a variety of cytoplasmic signaling cascades to influence the transcription of ERE independent genes (**Figure 8**)[52].

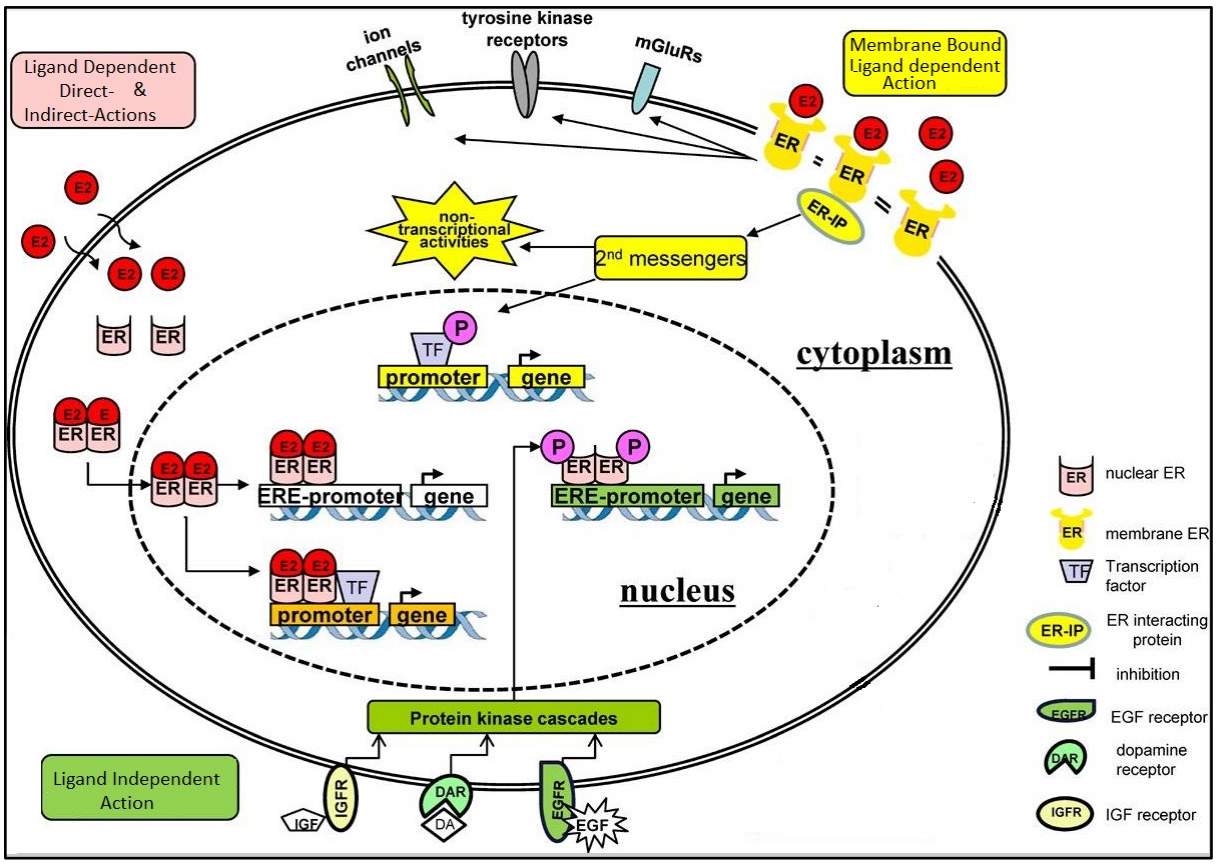


Figure 8. Estrogen Receptor Signalling Mechanisms. Visual representation of three ER signalling mechanism pathways. Adapted from Cui et al. (2013)[52]**.**

2.4 FOXA1

The forkhead box (FOX) protein family consists of transcriptional factors involved in a wide range of developmental and regulatory roles. The FOX family is characterized by an

evolutionary conserved so-called Fork-Head DNA-binding domain (FHD). In humans, fifty fork-head proteins have been identified[53]. These TFs have been partitioned into 19 groups (denoted FOXA to FOXS) based on the sequence similarity both within and outside of the forkhead domain. In general, each group has a distinct regulatory role in different cellular processes[53]. We focus specifically on *FOXA1* because it is a key regulator of steroid receptor function in cancer [54].

FOXA1 is encoded by the *FOXA1* gene on Chromosome 14q21.1. The *FOXA1* gene has three exons (**Figure 9**) and two known transcripts produced by alternative splicing. The 49.1kDA protein is 472 amino acids long with a FHD flanked by two DNA TransActivation Domains (TADs) near the N- and C- terminus[55]. The FHD is flanked by a NLS that directs the protein into the nucleus[56].

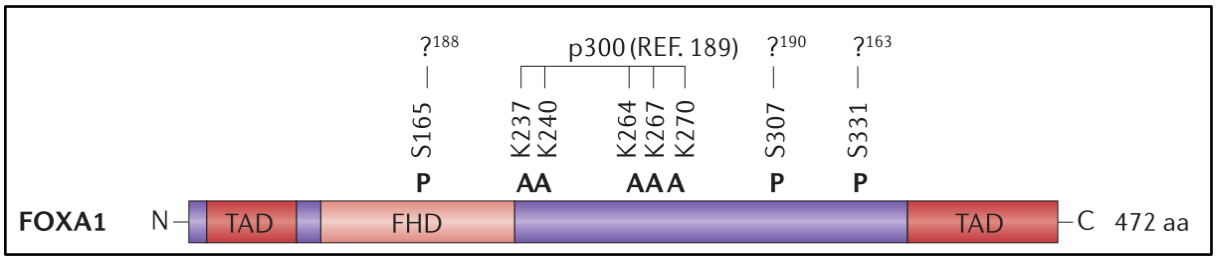


Figure 9. *FOXA1* gene structure. Adapted from Lam et al. (2013)[53].

Pioneer factors, such as FOXA1, are TFs which bind to condensed chromatin, can induce local changes to chromatin structure exposing DNA with the assistance of other cofactors and enzymes, and regulate expression of newly accessible genes along genomic locus[57]. These factors are also able to continue to access DNA prior to the time of transcriptional activation[57]. FOXA1 can bind to both DNA and core histones, resulting in the abrogation of the internucleosomal interactions mediated by histones H3 and H4 and decompacting the nucleosomal array, making it accessible to other proteins[58]. The butterfly-like structure of the FOXA1 FHD consists of three α-helices, three β-sheets and two loops (wings), arranged in winged helix structure (α1-β1-α2-α3-β2-W1-β3-W2), where the α3 helix

and W2 domain primarily bind the DNA major and minor grooves respectively[59]. The FHD binds DNA with the consensus sequence 5'-(AC)A(AT)T(AG)TT(GT)(AG)(CT)T(CT)-3'. The winged helix structure of FOXA1 resembles linker histones H1 and H5, enabling it to make contact with the sides of the nucleosome cores[60]. The H1 linker histone requires four amino acids (K40, K52, R42, and R94) to compact DNA on the nucleosome[61]. Unlike histone proteins, FOXA1 does not contain the amino acid composition necessary to condense chromatin[62].

FOXA1 appears to be a master regulator of tissue specific differentiation and function, mainly associated with sex hormone dependant tissues, such as the breast and prostate glands. In the breast, coexpression of ERα and FOXA1 is found in the luminal epithelial cells of the developing mammary gland with strong expression in the TEBs[63]. In relation to ductal morphogenesis in the mammary, FOXA1 appears to be necessary in the acquisition of ERα expression. This effect is observed with the deletion of FOXA1, which results in the loss of ERα expression in the luminal progenitor cells that give rise to the ductal lineage[63]. In addition, mammary glands with homozygous deletion of the gene are only able to grow rudimentary trees as the development of TEBs is disrupted. Heterozygous deletion of FOXA1 leads to increased alveolar development when compared to the wild type after exposure to pregnancy hormones. These findings suggest that FOXA1 may also be involved in sustaining the luminal epithelium in an undifferentiated state by pushing cells towards a luminal fate[63].

The transcriptional activation of FOXA1-dependent genes is driven through various mechanisms. In addition to being a pioneer factor, FOXA1 binding can be regulated by other parameters. For example, although there are a large number of genomic regions that contain the Forkhead motif, only a small fraction are actually bound by FOXA1[64]. DNA methylation is believed to prevent FOXA1 from accessing the chromatin. This is due to most binding events occurring in hypomethylated regions and not the Foxhead motif, which lacks the classic GC

sequence[65,66]. Furthermore, there is a positive correlation with histone modifications and FOXA1 binding. More specifically, histone 3 Lys 4 mono and dimethylation (H3K4me1 and H3K4me2) are enriched at cis-regulatory domains that are bound by both ER and FOXA1[64]. These epigenetic modifications are biased towards the enhancer regions[67]. FOXA1 binding these enhancers promotes demethylation, stabilizing the binding of the protein and promotes the recruitment of transcriptional regulatory effectors[66].

The FOXA1 protein does not act alone. Instead, it cooperates with other transcription factors and chromatin remodelling proteins. This enables it to contribute to the expression of a wide array of genes. For example, it interacts with other pioneer factors, such as activating enhancer-binding protein 2γ (AP2γ) and GATA3 to regulate expression of ERα- and AR-targeted genes by making these regions more accessible to ERα and AR[63,68]. FOXA1 is also able to form a transcriptional enhanceosome with ERα and GATA3. The TFs in the enhanceosome cooperatively bind to adjacent sides, resulting in regulation of estrogen-dependent gene expression[69].

In a manner similar to *ESR1*, *FOXA1* is highly expressed in luminal subtype tumours[10]. In ER+ MCF7 BC, FOXA1 is constitutively bound to chromatin regions that are also bound by ER, suggesting that direct ER binding requires the presence of FOXA1 binding in close proximity[70].

2.5 GATA3

In humans, the GATA gene family comprises six TFs (GATA1-6). These TFs are found on different chromosomes and are characterized by their ability to bind the “GATA” motif. GATA proteins consist of two N-terminal transactivation domains (TA1 and TA2) and a dual

zinc finger DBD. These zinc finger domains are highly conserved in the GATA family with over 70% and recognize 5'-(A/T)GATA(A/G)-3' sequences[71].

GATA3 is a protein encoded by the *GATA3* gene on Chromosome 10p14. The *GATA3* gene contains six exons and encodes for two similar transcripts. Exons 2 to 5 remain identical between the two isoforms and exons 1 and 6 vary slightly, producing two 47.9/48 kDa proteins consisting of 443 and 444 amino acids respectively[72]. The functional difference between these two variants has yet to be determined.

Members of the GATA family are expressed in different tissues. GATA1 and GATA2 are primarily expressed in hematopoietic cells, whereas GATA4, GATA5, and GATA6 are expressed in mesoderm and endoderm derived tissues. However, GATA3 is expressed in both hematopoietic and non-hematopoietic tissues[73].

GATA3 is an important gene for the establishment and maintenance of luminal cell identity. In the differentiated luminal epithelial cells lining the breast ductal structures of the mammary gland, *GATA3* is the most highly expressed TF[74]. It plays a pivotal role in mammary development, where conditional deletion around puberty prevents formation of TEBs in gland morphogenesis[75]. Furthermore, GATA3 deficiency in MaSc has led to an impairment of lactogenesis, where milk protein genes *Wap* and *Csnb* expression is decreased[75]. This indicates that GATA3 is an essential driver in the differentiation along the alveolar lineage.

Chapter 3. Genomic technologies

COIN-seq exploits CRISPR-based interventions delivered through lentiviral infections. The effect of these interventions is captured using single cell sequencing. In this chapter, we describe the previous bodies of work and related technologies that COIN-seq is predicated upon.

3.1 CRISPR

Genome editing is a term used to describe methods that allow the user to make changes in a target genome. These techniques generally utilize highly specific nucleases that induce site-specific changes in the target organisms' genome. A relatively new leader in these technologies is the CRISPR/Cas9 system. The Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) protein 9 system is a robust and multiplexable genome editing tool that allows the user to modify targeted genomic elements in a precise manner[76].

The CRISPR/Cas system was natively discovered in prokaryotes as an acquired immune system against viruses and phages. There are over 40 known Cas protein families categorized into three types, but the most used today is the Type II Cas9 protein derived from *S. pyogenes*. Natively, this system consists of a nonspecific Cas9 nuclease and a set of programmable sequence-specific crispr-RNAs (crRNAs) which interact together to form the CRISPR-Cas9 complex that cleaves the target DNA. The CRISPR locus contains Cas9 genes, a conserved AT-rich leader sequence, and an array of spacer and repeat sequences[77]. A spacer sequence is a unique sequence acquired by mobile genetic elements, such as invasive foreign DNA. These unique sequences are adapted into the CRISPR array and separated by short repetitive elements (**Figure 10**).

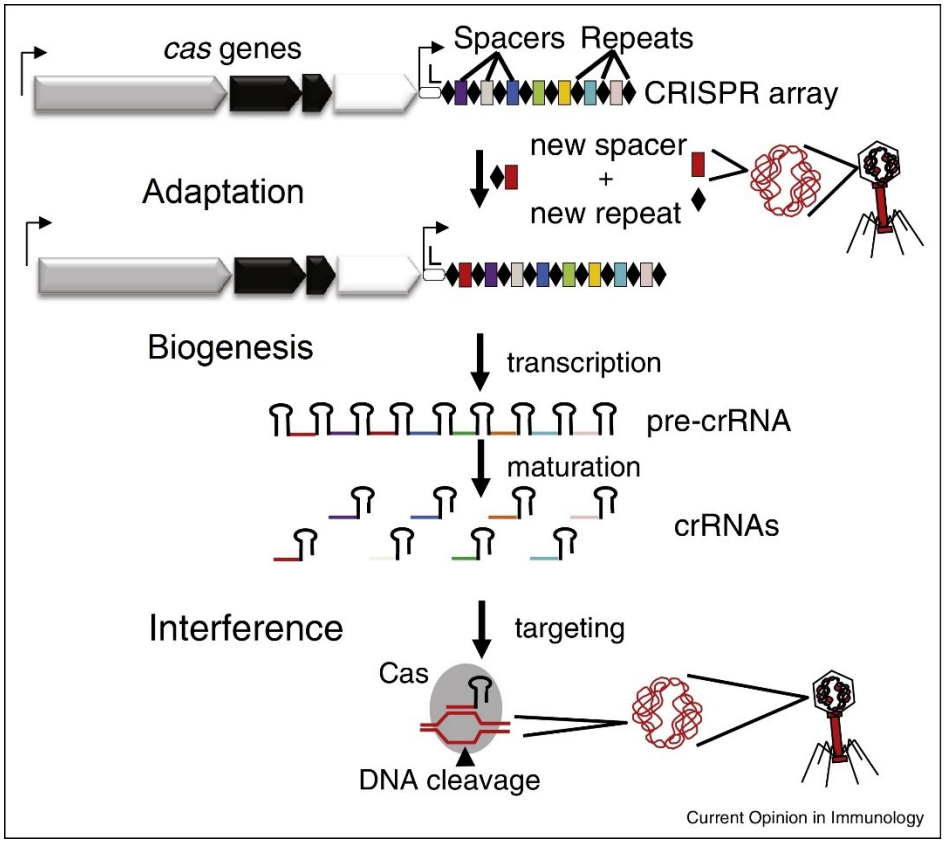


Figure 30. CRISPR-encoded immunization and interference. During the adaptation stage, pieces of DNA are sampled from the invader and are integrated as a new spacer in the array. When expressed, the CRISPR array is transcribed into crRNAs to guide Cas toward complementary DNA. Adapted from Barrangou (2015)[77].

The molecular mechanism of CRISPR-Cas9 in *S. pyogenes* can be divided into three stages: adaptation, biogenesis, and target interference. During adaptation, a partial distinct sequence of the invading foreign DNA known as a protospacer is incorporated into the CRISPR array, allowing the bacteria to memorize the genetic material. The biogenesis step enables immunity by transcribing the CRISPR array into precursor crRNA (pre-crRNA). In a Type II system, the repeat sequences of the pre-crRNA bind to transactivating crRNAs (tracrRNAs). This RBA duplex is stabilized by Cas9 and is cleaved by the host RNAse III to produce mature guide RNAs (gRNAs)[78]. The Cas9-gRNA complex provides the bacteria with immunity in the target interference stage. The Cas9 protein contains a RuvC and HNH endonuclease responsible for cleaving double stranded DNA (dsDNA)[79]. The functional activity of this

complex is highly dependent on the presence of a protospacer adjacent motif (PAM) sequence directly adjacent to the protospacer. In *S. pyogenes*, the PAM sequence is NGG[80]. The protospacer acts as a guide and binds the complementary foreign DNA, where the Cas9 protein introduces a double stranded cut in the DNA.

CRISPR is a powerful tool that has been adapted into genome editing. Natively, the presence of the crRNA/tracrRNA complex is necessary for the functional activity of this system. However, the community has designed a single guide RNA (sgRNA) containing all the essential crRNA and tracrRNA components[81]. With this, the native CRISPR system can be described as two components: the Cas9 protein and the sgRNA. These components have been incorporated into plasmids, allowing users to implement their own protospacer sequence near the sgRNA sequence through restriction cloning. This system can be introduced into a foreign organism and KO targeted genes. The double stranded breaks introduced by Cas9 are repaired through either homology-directed repair (HDR) or, more commonly, nonhomologous end-joining (NHEJ)[82]. HDR requires the user to introduce a homologous DNA template to precisely repair the cleaved. NHEJ-mediated repairs occur more frequently as there is no need for an intermediary component, making this method of repair error prone and introducing indels. When used in genome editing, designing a guide RNA to target a coding region can result in a loss of function (LOF) mutation of that gene as a result of NHEJ repair [83].

Although the role of the CRISPR system is to cleave dsDNA, it has been extended to perform a variety of tasks. Modifications to Cas9 can allow the protein to affect transcription without performing genome sequence modification. By introducing the two-point mutations D10A and H840A in the RuvC and HNH endonucleases, the protein loses its functional ability to cleave DNA and becomes a catalytically dead Cas9 (dCas9)[36]. However, the complex is still

able to bind its target based on the sgRNA sequence. Researchers have been able to adapt dCas9 to perform transcriptional regulation without genetic alteration. Transcriptional downregulation using the system is known as CRISPR interference (CRISPRi) and upregulation is CRISPR activation (CRISPRa). To perform a KD of a target gene, the sgRNA is designed to target the promoter or exon, where the dCas9-sgRNA complex binds and blocks RNA polymerase, stopping transcript elongation[36]. OE utilizes a dCas9 modified to include activator proteins, such as VP64, and sgRNAs that target the promoter. Binding of this complex functions as a transcriptional activator, resulting in increased expression of the target gene[35].

3.2 Lentivirus

Plasmid delivery in mammalian cells is performed using lentiviral transduction. Lentiviruses are a subset of retroviruses that are capable of transducing dividing and non-dividing mammalian cells without eliciting a significant immune response. Once transduced, the viruses integrate stably into the host genome and produce long term transgene expression. The components necessary to construct lentiviral particles are separated into three plasmids: the lentiviral expression plasmid which is integrated and expressed in the target cells, a packaging plasmid (psPAX2), and an envelope plasmid (VSVg). These plasmids are transfected into the φnx cell line, a cell line based on HEK 293T used for lentivirus production, to create lentiviral particles.

3.3 Existing CRISPR-based systems

There are many variations of these CRISPR systems, but we have chosen to adopt those from the KO and OE systems designed by the Zhang Lab of the Broad Institute. The KO system is a two-vector system where the Cas9 components are in the lentiCas9-Blast vector (**Figure 11, left**) and the sgRNA components are in the lentiGuide-Puro vector (**Figure 11,**

right). For this system to function, both must be transduced into the target organism. There exists a version of this system that combines the two into one vector: delivering both the Cas9 and sgRNA together. However, our design requires that multiple sgRNAs are introduced into individual cells. This poses a problem as each cell would also receive multiple Cas9 cassettes and lead to the over-production of Cas9, which is toxic to the cell. Therefore, the benefit to having these separated is that Cas9 toxicity can be avoided by generating a Cas9 expressing cell line, where one copy of the Cas9 cassette is put into the cell[34].

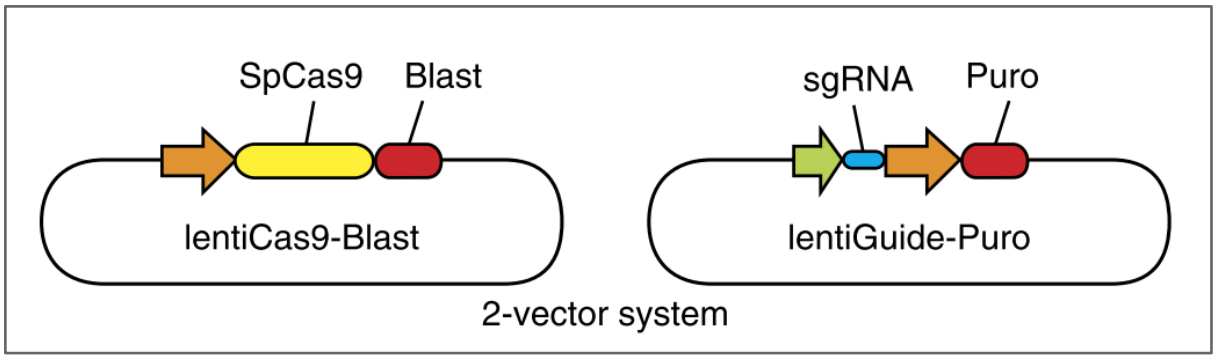


Figure 11. Two vector KO system. Left: Plasmid with Cas9 containing a blasticidin resistance cassette. **Right:** Plasmid with sgRNAs used for KO interventions containing a puromycin resistance cassette. Adapted from Joung et al. (2017)[84]**.**

The OE system uses a three-vector system entitled Synergistic Activation Mediator (SAM). The first vector, lenti dCas-VP64_Blast, contains a modified dCas9 and a VP64 transcriptional activator fusion protein (**Figure 12, top left**). The second vector, lenti sgRNA(MS2)_puro backbone contains a modified sgRNA scaffold (**Figure 12, bottom**). In a normal Cas9-sgRNA system, they interact to form a complex in which the tetraloop and stem loop 2 structure of sgRNA scaffold protrudes from the Cas9 protein (**Figure 13, a**). The modified sgRNA in the SAM system includes a minimal hairpin aptamer which protrudes the Cas9 protein as well as selectively binds dimerized MS2 bacteriophage coat proteins (**Figure 13, b**). The third vector, lentiMPH (**Figure 12, top right**), provides an MS2-p65-HSF1 (MPH) fusion protein composed of three proteins: a bacteriophage MS2 coat protein that binds the MS2 loops of the OE sgRNA scaffolds, the C-terminal activation domain from the human heat

shock transcription factor HSF1, and the C-terminal portion of the p65 subunit of mouse NF-κB. This acts as a transcriptional activator when bound to the MS2-binding loops of the modified sgRNAs (**Figure 13, c**)[35]. Together, these three vectors interact to increase the transcription of target genes.

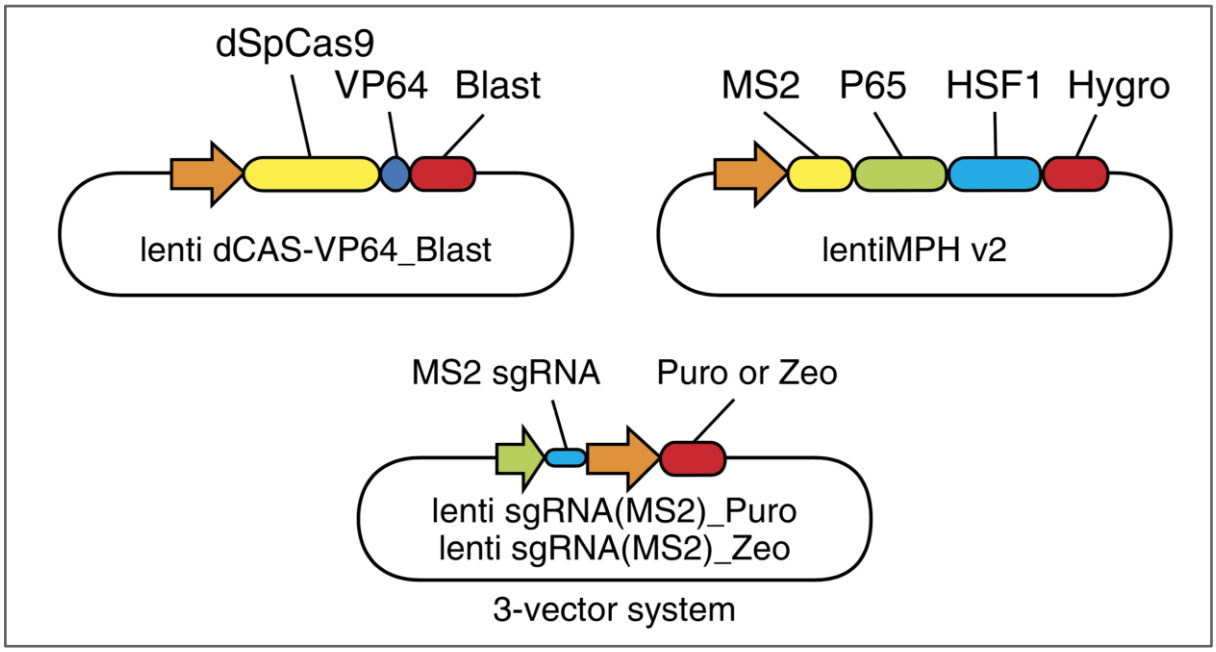


Figure 12. Three vector OE system. Left: Plasmid with catalytically dead Cas9 vector fused to transcriptional activator VP64 containing a blasticidin resistance cassette. **Right:** MS2-p65-HSF1 (MPH) fusion protein used for OE intervention containing a hygromycin resistance cassette. **Bottom:** Plasmid with modified sgRNAs used for OE interventions containing a puromycin resistance cassette. Adapted from Joung et al. (2017)[84].

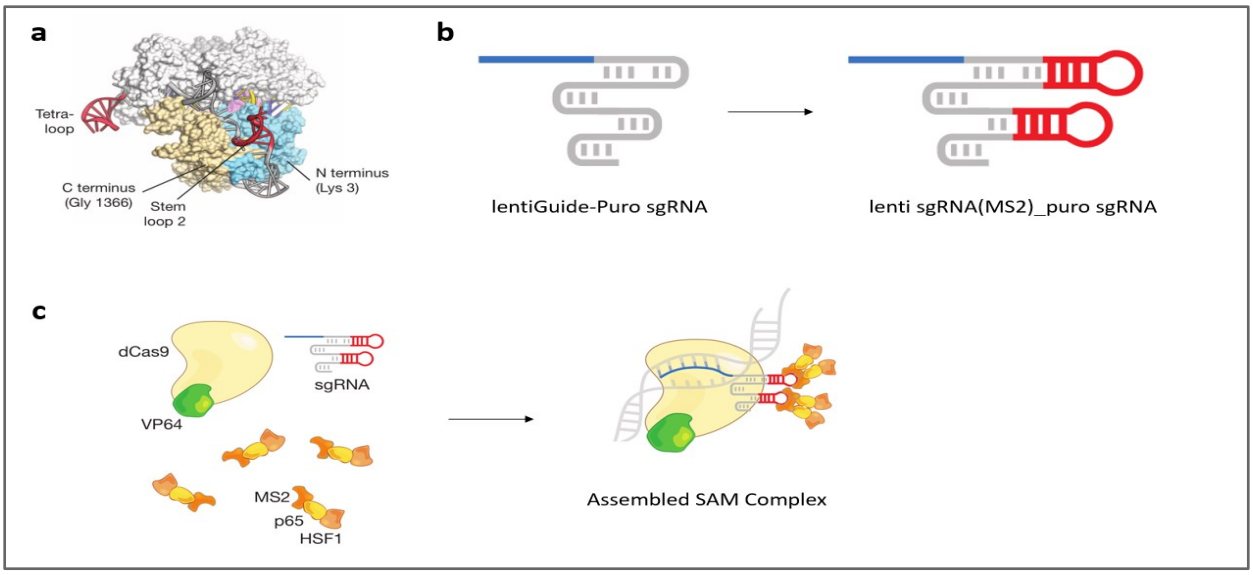


Figure 13. Synergistic Activation Mediator (SAM). a. Crystal structure of the Cas9–sgRNA target DNA ternary complex. **b.** OE sgRNA scaffold modification involving the addition of two MS2 loops. **c.** Assembled SAM Complex overexpressing gene target. Adapted from Konermann et al. (2015)[35].

Both systems have been extended to allow for orthogonal gene KO and OE gene perturbations within the same target cells. In general, the complementary DNA sequence found in sgRNAs is around 20 nucleotides long because shorter and longer lengths can negatively affect CRISPR efficiency[85]. However, when reducing the length of the RNA targeting sequence to 14-15 nucleotides and pairing with catalytically active Cas9, the complex is unable to induce a double-stranded break. These shortened RNAs are dubbed dead RNAs (dRNA). Furthermore, when using the modified sgRNA scaffold containing MS2 binding loops with the shorter RNA guide sequences, the complex can function as a transcriptional activator in the presence of MPH (**Figure 14)**[86].

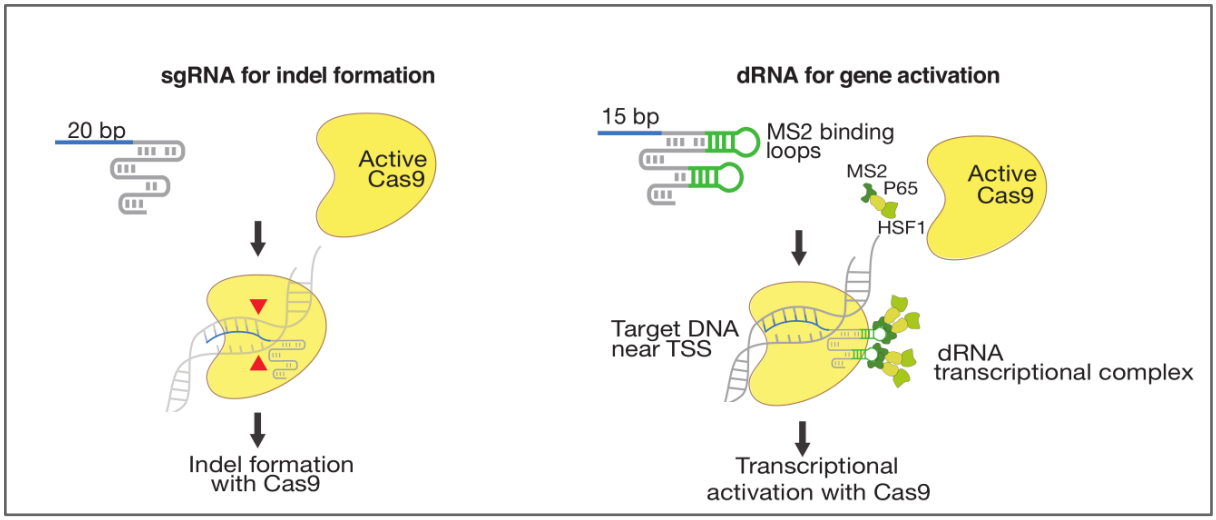


Figure 14. Modulation of intervention by modification of sgRNA length. CRISPR construct for KO (left) and OE using modified dRNA (right). Adapted from Konermann et al. (2015)[86].

3.4 Single cell transcriptomics

RNA sequencing (RNA-seq) provides a quantitative metric in which we can characterize a biological sample. It is a sequencing technique that uses next-generation sequencing (NGS) to count the individual RNA transcripts in a biological sample. There are many different sequencing technologies available, but the most commonly used is the short-read method from Ilumina. This method involves extracting the RNA from the sample, processing it into cDNA, and ligating specific adapter oligonucleotides to both ends, creating a sequencing library. These cDNA fragments are usually under 200 bp in length. The library is then sequenced, and the data is computationally processed to return raw reads, which are aligned to a reference genome, resulting in a quantification of reads associated to genes[87].

Limitations of bulk RNA-sequencing approaches.

The traditional bulk RNA-seq technology has been a limiting factor in our understanding of how a complex disease such as BC functions. Part of that complexity stems from the fact that it is a heterogeneous disease. Cell heterogeneity is a problem that bulk

profiling is unable to solve by design. The majority of genomic BC studies have been limited by bulk profiling, which reflects an averaged gene expression of the samples. These averaging signals from multiple individuals can cause misleading effects, a phenomenon known as Simpson's Paradox (**Figure 15, a**)[88]. The biology of a tumour can be better understood by studying the specialized cell types found within it and the microenvironment surrounding it[89]. Analysis of tumour profiles that have averaged out gene expression ablates the contribution of tumour heterogeneity. This effect has been shown to negatively affect the prognostic capacity of both clinical and intrinsic subtyping schemes[19]. Furthermore, bulk profiling is unable to differentiate the changes in cell populations caused by gene regulation or shifts in cell type composition. Consider a population of BC composed of two different cell types. If you wish to measure their gene expression pre- and post-drug treatment using bulk profiling, it would become impossible to distinguish if the main cause of change in gene expression is due to a change in gene regulation or cell composition (**Figure 15, b**)[88]. In time series experiments, cells in a population may differentiate asynchronously and gene expression profiles can differ among cell stages. Bulk sequencing would average the gene expression of each cell stage and is unable to differentiate expression levels affected by these stages at different time points (**Figure 16, c**).

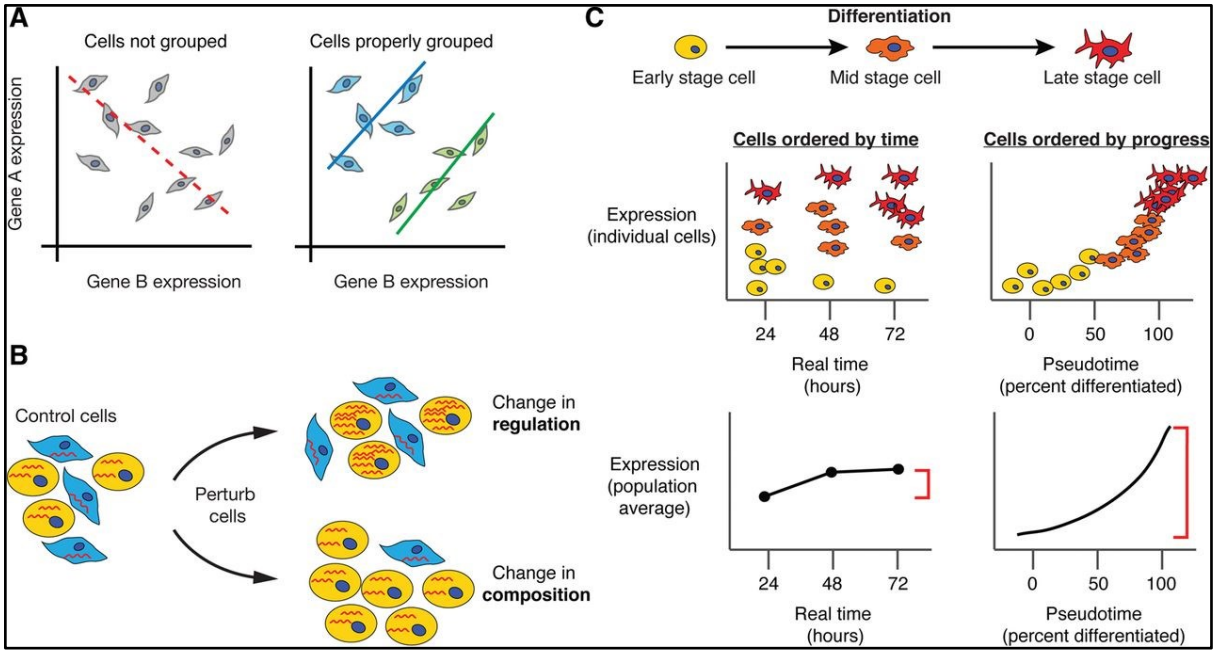


Figure 15. Single cell measurements preserve information lost bulk RNA-seq. A. Simpson's Paradox. Technologies such as bulk RNA-seq create an averaging effect that may hide important biology. **B.** The cause of change in gene expression post perturbation can be due to a change in regulation or composition. **C.** Reordering cells in pseudotime can ablate the effects of averaging caused by cells in unsynchronized different biological progression states. Adapted from Trapnell (2015)[88].

Single Cell Sequencing

The development of single cell sequencing addressed some of the shortcomings of conventional RNA-seq. Unlike classic bulk RNA-seq, single cell sequencing examines the RNA content of every individual cell. This allows us to address the three problems discussed above. There are different single cell sequencing technologies being developed. However, we use Drop-seq.

Nanoliter droplet-based single cell capture with RNA-sequencing (Drop-seq)

Single cell sequencing is a modern field with new technologies and modifications to existing technologies are being introduced regularly[90–92]. We follow the so-called Drop-seq

approach of Macosko et al.[90]. At time of profiling, the target population of cells are in aqueous suspension, and syringe pumps push three separate solutions through the inflow channels of a polydimethylsiloxane microfluidic device (PDMS). The three channels correspond to the suspended cells, a buffer containing specialized microparticles (or beads) with a lysis agent, and oil. Aqueous droplets form at the confluence of these three channels. Following a super-Poissonian distribution, some droplets contain exactly one cell and exactly one microparticle. Here, the cells lyse in the droplet, and the transcriptome of that single cell attaches to the microparticle.

Each microparticle has a 30 μM diameter and contains 10^8 extruding oligonucleotides; each such oligonucleotide contains a specialized PCR primer, a unique cell barcode (CBC), a unique molecular identifier (UMI) and a 30 bp long polyT tail (**Figure 16**). After lysis, mRNAs are captured by their polyA tails. Captured mRNAs are processed through reverse transcription using the specialized PCR primers to integrate the CBC and UMI into the mRNA. In this way, each mRNA/cDNA from the cell is labelled with the same CBC. This results in a set of microparticles called single-cell transcriptomes attached to microparticles (STAMPs). The STAMPs are then amplified into cDNA libraries and sequenced using Illumina technologies. Each paired-end read captures the barcode and a fragment of the transcript. The barcode thus identifies all the captured transcripts from the same cell (**Figure 17**).

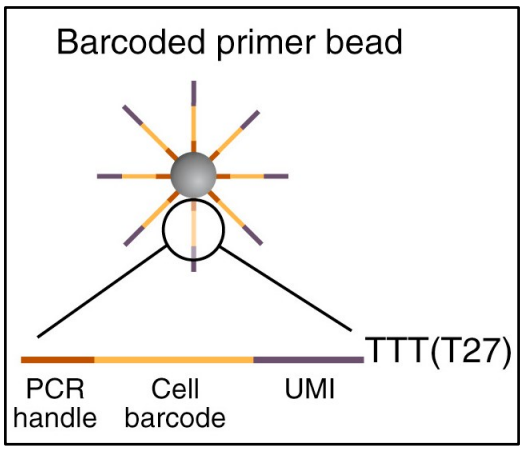


Figure 46. Composition of barcoded primer bead. Each bead contains 10^8 extruding oligonucleotides with a PCR handle, a CBC, a UMI, and a long polyT tail. The polyT tail allows for mRNA capture by binding to the polytA tails. Adapted from Macosko et al. (2015)[90].

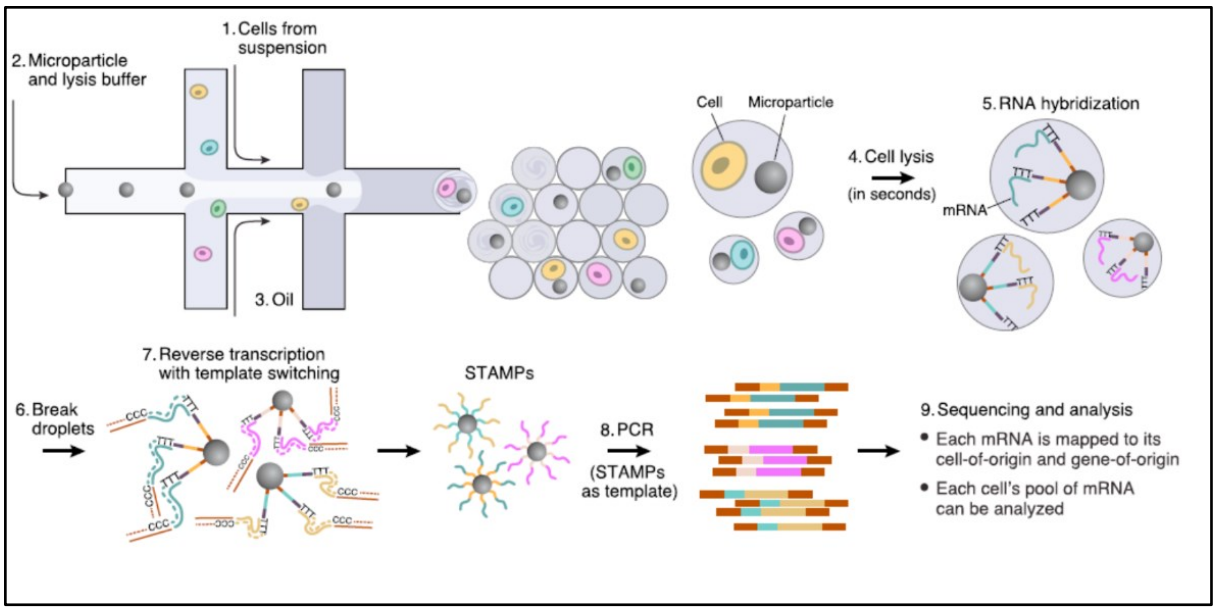


Figure 57. Extraction and Processing of Single-Cell Transcriptomes by Drop-seq. Cells, droplet forming oil, and microparticles suspended in lysis buffer are channelled through the Drop-seq device. Within each droplet containing a cell and a microparticle, the cell is lysed and the mRNA hybridizes to the polyT tails of the microparticles. Droplets are broken and mRNA is reverse transcribed to create STAMPs. STAMPs are processed to generate cDNA libraries. Adapted from Macosko et al. (2015)[90].

3.5 Perturbation screening technologies

Genetic screens help infer gene function in mammalian cells, but assaying complex phenotypes such as transcriptional profiles, remains a challenge at scale. These problems are addressed with the advent of Perturb-seq, a methodology that combines CRISPR/Cas9 technology to perform multi-locus gene perturbations with the scale of single cell RNA-seq[32]. First, protospacers are designed to KO a set of genes. These are ligated into the sgRNA scaffold of the Perturb-seq plasmid (**Figure 18B**) and are packaged into lentiviral particles

(**Figure 18A**). The particles are pooled and transduced into Cas9 expressing target cells at varying multiplicity of infections (MOI). A low MOI increases the occurrence of single perturbations, where a cell receives one sgRNA to study single gene effects, and a high MOI increases the occurrence of one or more sgRNA in one cell to study epistatic effects. When a cell receives a perturbation, the lentivirus integrates the Perturb-seq plasmid into the genome. The plasmid is composed of two cassettes: an sgRNA cassette and a selection cassette (**Figure 18B**). When integrated into the genome, the U6 promoter drives expression of the sgRNA cassette and the resulting sgRNA interacts with the Cas9 already present in the cell to perform the gene KO. Similarly, the selection cassette is integrated simultaneously, but is instead driven by an EF1α promoter. This cassette encodes the puromycin resistance gene, a blue fluorescent protein (BFP) gene, a guide barcode (GBC), and a polyA tail (**Figure 18B**). The GBC is a set of distinct nucleotide sequences designed to encode the identity of each perturbation. For example, perturbation sets A to E in **Figure 18A**, would have five unique GBCs encoding the perturbation identity. The cells are then exposed to selective pressure using the puromycin antibiotic in which only cells that received a perturbation will survive. The surviving cells are then single cell sequenced. When performing single cell sequencing, the microparticles are able to capture the cell transcripts as well as the transcripts from the Perturb-seq selection cassette. This means that for each cell, we would observe a CBC, GBC, and UMI which allows for the identification of the transcriptional profile of every unique cell and the genetic perturbation delivered (**Figure 18A**)[32]. We note that there exists a second implementation of Perturb-seq from Adamson et al. (2016) which focuses on performing KD perturbations in a similar way[33]. We note that this work is also primarily built upon the same machinery.

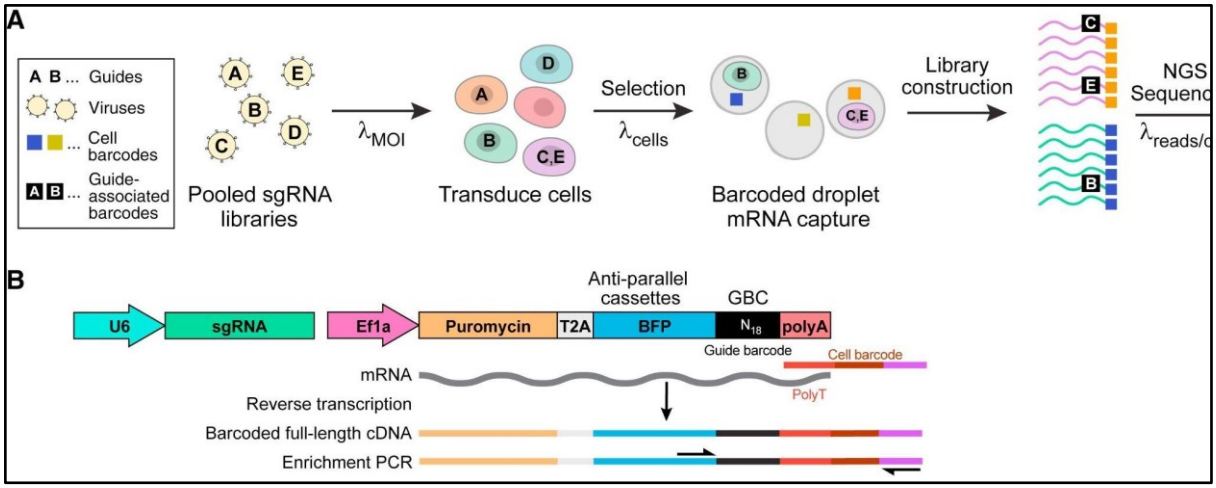


Figure 68. Perturb-seq methodology. A. Pooled CRISPR screen and single cell sequencing. Perturbations are introduced by transducing sgRNA libraries into target cells at varying MOIs to study either single gene or epistatic effects of perturbations. These cells are processed using single cell sequencing methods, such as Drop-seq. **B.** Perturb-seq vector. Adapted from Dixit et al. (2016)[32].

Chapter 4. Developments of enabling technologies

The development of COIN-seq required the adoption of existing technologies. I contributed to the integration of these technologies with Sanny Khurdia and Abdelrahman Ahmed and several other members of Dr Hallett's lab.

4.1 Development of an efficient DIY Drop-seq system for COIN-seq

One of my contributions was to assist with the construction of the Drop-seq device that enabled single cell sequencing of our samples. We follow the design of the Drop-seq system from Macosko et al.[90] (**Chapter 3.4**, see also **Sup. Met. 6**). However, there were several differences in our approach. We utilize the same PDMS device but constructed our own syringe pump single cell isolation device derived out of cheap and accessible components. The design was adopted from Booeshaghi et al. (2019)[94], where the syringe system is built from easy to access off-the-shelf components for a fraction of a cost[94]. Instead of using a confocal microscope, we opted for a smaller and cheaper digital microscope. This provided a comparable image while enabling the Drop-seq device to be small, cheap, and portable. The system has been used in several projects and biological domains including the yeast *Candida albicans,* mouse and human studies.

The cost of our system is approximately $400. By preparing reagents from scratch, the per-cell cost of a Drop-seq run is ~$0.10. This method is more affordable when compared to the leading commercial single cell RNA-seq platforms such as in 10X Genomics and Fluidigm. The cost of the 10X Chromium instrument is approximately 1000x that of our system with a per-cell cost of $0.50, making it the most expensive single cell solution available[95]. Fluidigm uses two different integrated microfluidic chips, which have the capacity to capture 96 and 800 cells, making it a low throughput solution compared to Drop-seq. Therefore, our system represents a reasonable cost-effective alternative to commercial systems yet is capable of

profiling a sufficient number of cells in order to implement the COIN-seq system within current financial constraints.

We construct the Poseidon Drop-seq device following the approach described in Booeshagi et al. (2019)[96] (**Figure 19**). The Poseidon device is composed of two major parts: the microscope station and three identical syringe pumps. First, we 3D printed the plastic frame for the microscope station and pumps using the associated CAD designs. These plastics were then assembled with cheap and accessible components to create the full station and pumps. Each pump was equipped with a stepper motor driving a lead screw. On this screw, we attached a sled mounted on linear bearings on which the plunger of a syringe can rest. This allowed the stepper motor to precisely control the aspiration or infusion of the syringe. The microscope station was equipped with a cheap microscope and a small touch screen monitor, where the user can control the device's parameters and allow the user to observe the flow of liquids through the PDMS. The components of the microscope station were controlled by a Raspberry Pi attached under the monitor using the Poseidon open-source software (https://github.com/pachterlab/poseidon). The Pi was connected to a small Arduino used to relay commands to drive the stepper motors. We made some minor modifications to the software to squash the bugs that affected microscope-monitor functioning. Lastly, we built two additional components to assure that the flow of beads was unhindered: an upright metal stand which held one syringe pump and a rotating magnet mixer controlled by the Raspberry Pi. The stand was built to hold the syringe pump with the beads vertically at a height that allows for the syringe, needle, and tubing to not touch the work surface. The mixer was designed to sit near the beads in the syringe, mixing the beads and thus allowing them to flow without clumping. The other two pumps held the droplet generation oil and samples. Together, the three pumps were connected to the PDMS chip using silicon tubing.

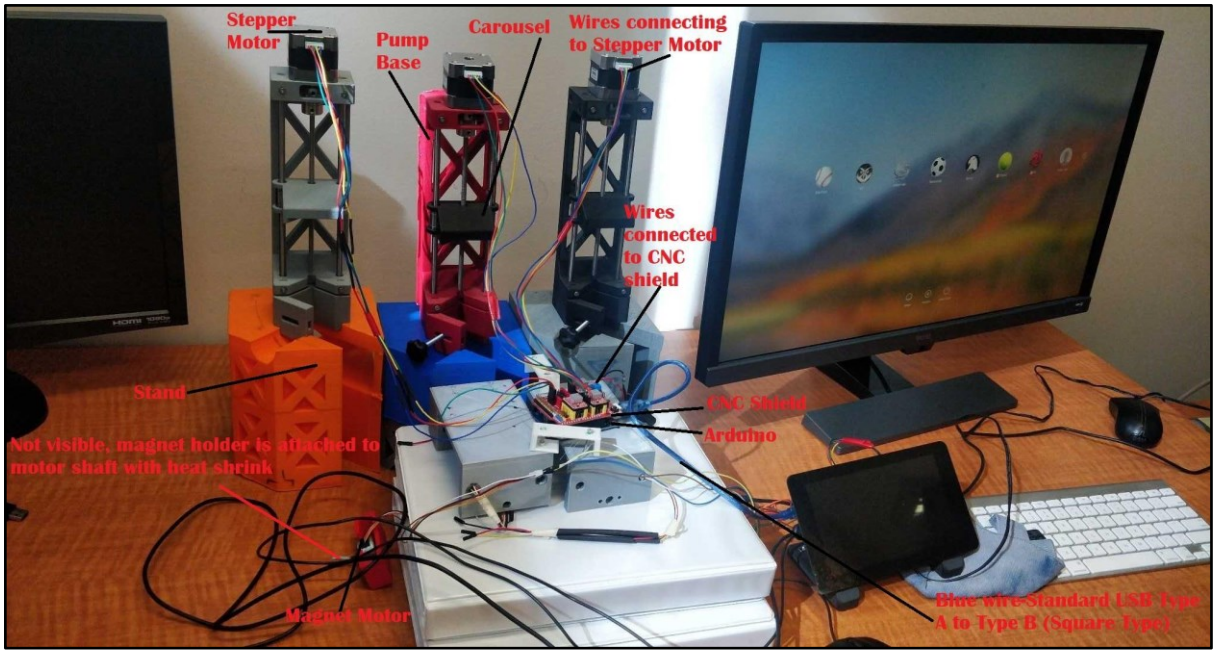


Figure 19. The Poseidon device. Assembled homemade Poseidon Drop-seq device adapted from Booeshaghi et al. (2019)[96].

Construction of the device was straightforward as we followed the protocol from Booeshaghi et al. (2019). However, there were technical problems involved with droplet formation. Using the associated open-source software and recommended pump parameters, droplet formation was inconsistent. Size of droplets varied from too small to capture microparticles and too large, where the frequency of double and triple microparticles were present. We circumvent this issue by manually tuning the pressures of each syringe until we confirm uniform and consistent droplet formation. Using the new pump parameters, droplets were uniform and the correct size, comparable to higher end medical grade syringe pumps (**Figure 20**).

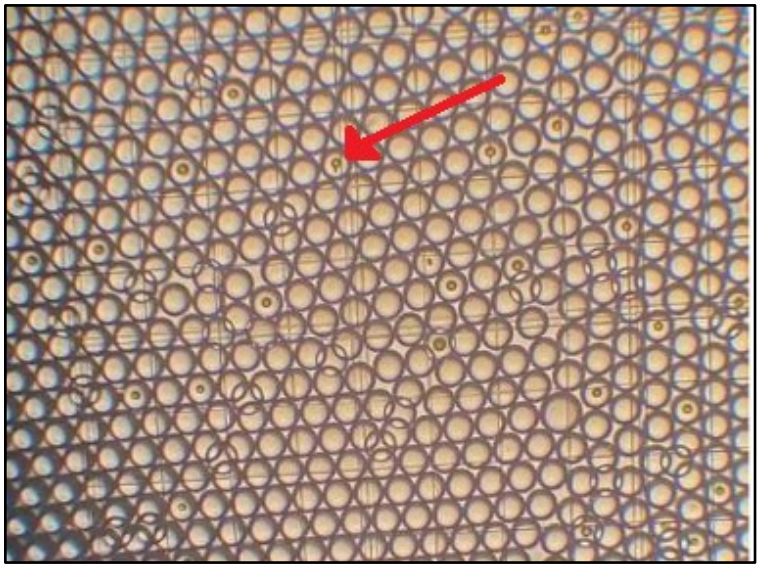


Figure 70. Optimal droplet formation. Picture of Cytometer showing droplets containing lysed cells and beads (40X). Droplets are uniform and of the correct size to minimize the capture occurrences of multiple microparticles. Red arrow denotes a droplet which has encapsulated a bead. Cells have been lysed and cannot be visualised.

4.2 DART-seq

The Perturb-seq vector (**Figure 18B**) contains a selection cassette, a unique GBC used for vector identification, and an sgRNA cassette only capable of inducing gene deletions. The implementation of COIN-seq follows the same philosophy of Perturb-seq, however instead of the Perturb-seq vector, it utilizes the Zhang lab vector systems described in **Chapter 3.3**[34,35], allowing for simultaneous KO and OE CRISPR interventions. This is done by generating clonal lines of cells transduced with both the Cas9 and MPH machinery, which we have named COIN cells. COIN cells are transduced with a pool of lentiviral constructs that encode a set of sgRNAs that target distinct genes in the pathway or process of interest. By modulating the MOI, we can stochastically control the expected number of times cells are infected by different sgRNAs. In this way, some cells are affected by single gene modulations, while in others multiple infections allow investigating epistatic effects. In principle, with the *canonical three* genes, we have 8 possible combinations excluding multiple infections with the same sgRNA

across ~100,000 sequenced cells. However, by removing the Perturb-seq vector, the GBCs are no longer present and cannot be used to identify the intervention identity. We circumvent this issue using a method called direct capture Perturb-seq.

Direct capture Perturb-seq from Replogle et al. (2020)[99] is a methodology that enables the capture of sgRNA transcripts directly, in which the captured sgRNA protospacers act as a barcode that identifies which set of interventions are present in the cell. This method was developed in the 10X Genomics single cell platform by delivering target-specific barcoded primers designed to target the 3' region of the sgRNA scaffold termed the capture sequence (**Figure 21**). These primers anneal to the capture sequence and enable reverse transcription (RT) of the sgRNAs, leading to efficient recording of the protospacer sequences, which functions as a unique identifier. The capture sequence is selected to target the 3' constant region of the sgRNA scaffold, enabling the capture of any sgRNAs regardless of the protospacer sequence. However, this modification is not directly applicable to Drop-seq because microparticle-based capture is mechanically different from the capture method used by 10X Genomics. Instead, we modify a portion of the extruding oligonucleotides to capture sgRNAs directly using a method called DART-seq.

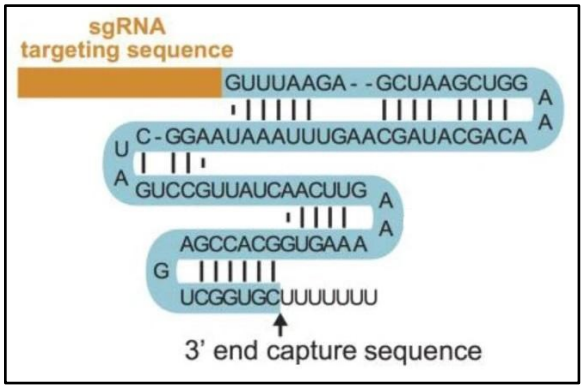


Figure 21. sgRNA structure. Consists of a variable region represented by the protospacer sequence (orange) and the constant region of the scaffold (blue). Adapted from Replogle et al (2020)[99].

DART-seq[93] is a technology that allows for the modification of up to 40% of the oligo (dT) primers at the end of the oligonucleotide tethered to the Drop-seq microparticles (**Figure 17, panel 5**). The development of DART-seq was mainly performed by one of my colleagues, Sanny Khurdia, who adapted and modified the original protocol to fit the needs of COIN-seq. In this approach, two oligonucleotides are designed: a splint oligo, which consists of a polyA sequence followed by an 'unique' region; and a custom oligo, which contains a sequence complementary to the unique region of the splint oligo followed by a specific sequence which is complementary to the intended target sequence. These oligos are designed such that they form a so-called 'toehold probe', which can then be ligated onto the extruding oligo. This enables the co-capture of the intended target sequence of the custom primer and mRNA transcripts from the unmodified oligonucleotides (**Figure 22**). We design this custom oligo to hybridize with the sgRNA molecules directly. As there is only a small fraction of the oligonucleotides on the microparticles that are modified, we enable the microparticles to co-capture cell mRNA and expressed sgRNAs.

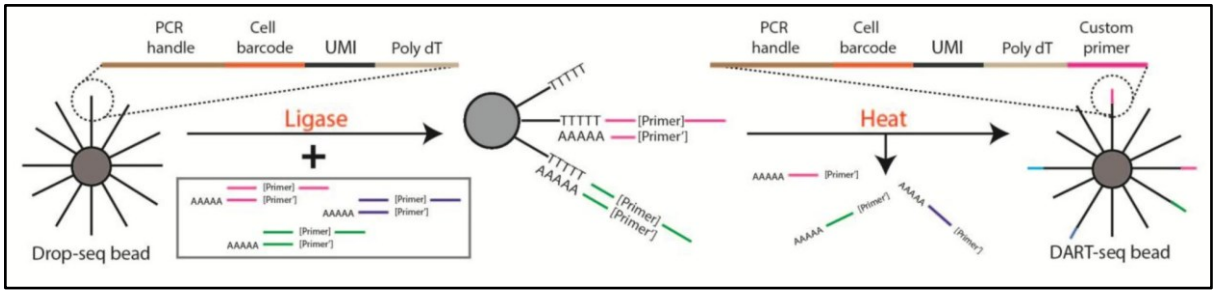


Figure 82. DART-seq protocol used to extend microparticle oligonucleotides. Custom primer sequence is designed to capture the 3' end of the sgRNA scaffold. Adapted from Saikia et al. (2019)[93].

Chapter 5. The Design of Combinatorial Intervention Sequencing: COIN-seq

Here, we build upon Perturb-seq to extend its functionality to include KO and OE interventions. Before proceeding with the technical description of these components, we highlight the fundamental criteria that we seek to optimize in this project:

1. First, the system should be high throughput, eventually allowing many concomitant interventions. This should require minimal human intervention.
2. Second, the system should be cost effective, ablating the need for hundreds of individual samples to be profiled.
3. Third, the system should be time efficient to implement and robust to human error.

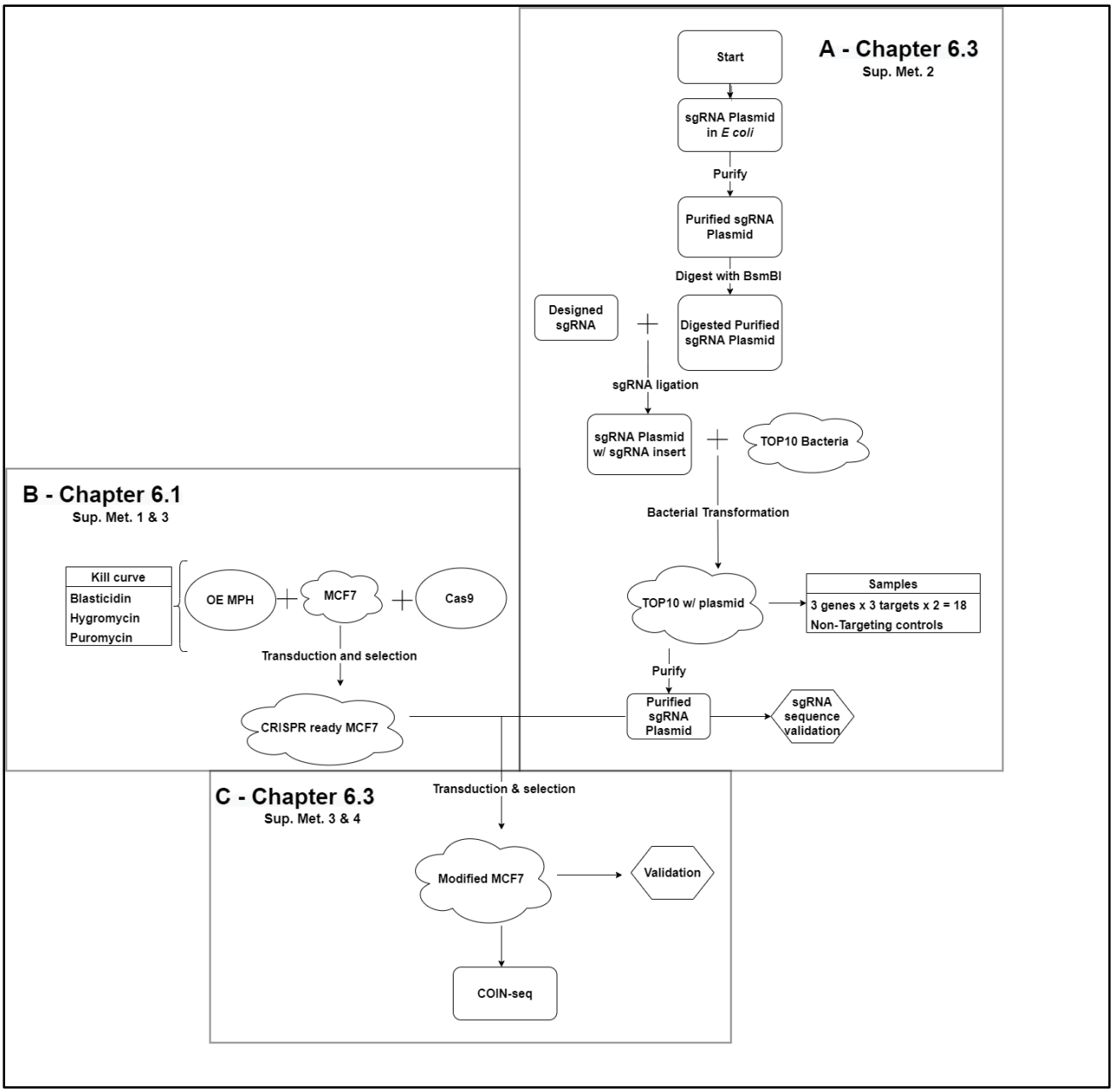


Figure 93. Implementation flow chart. Flow chart depicting the steps needed to implement COIN-seq. The individual steps in the figure are labelled by the subsection where they are discussed in the results chapter. **A.** Ligation of sgRNA expression plasmids with protospacer sequences. **B.** Clonal cell line generation. **C.** Transduction and characterization of sgRNAs effects.

5.1 The choice of MCF7 as our cell line of interest and GATA3, ESR1 and FOXA1 as our genes of interest

We use the MCF7 cell line from the American Type Culture Collection (ATCC). We chose MCF7 because it is an ER+ luminal epithelial invasive breast ductal carcinoma cell line, making it an acceptable model for ER+ tumours. In addition, the expression levels of *ESR1, GATA3,* and *FOXA1*, the *canonical three* genes believed to play a key role in luminal differentiation, are relatively higher in MCF7 than most other cell lines (**Figure 24**). It is also simpler to implement COIN-seq and related optimizations, as MCF7 is a very well characterized cell line, which improves reproducibility and simplifies "debugging".

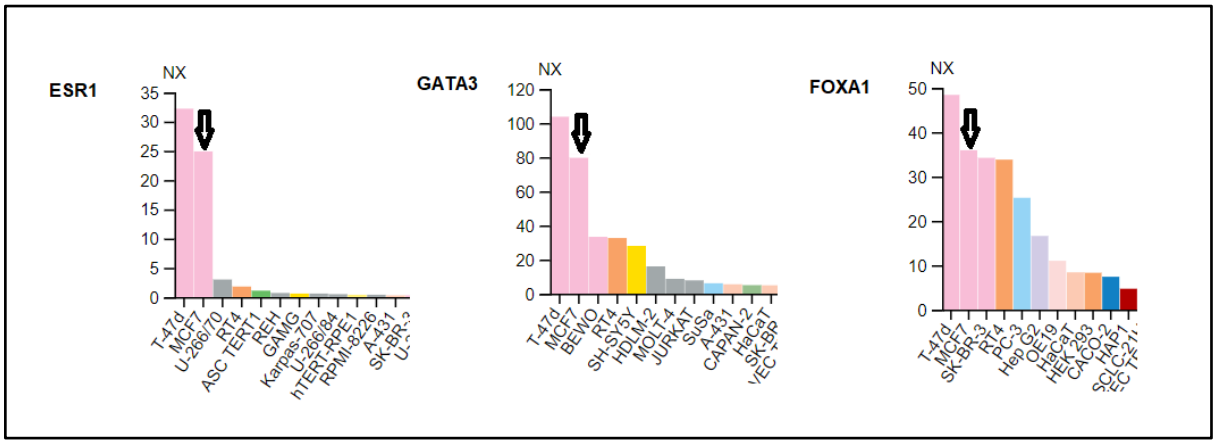


Figure 104. Bulk RNA-seq expression in cell lines. Bulk RNA-seq normalized expression profiles of each *canonical three* genes among commonly used cell lines. MCF7 has relatively high gene expression of each TF. Data obtained from Protein Atlas[107].

For technical development of the COIN-seq system, we chose these three genes from the set of six studied by the Mader lab that we believe to best satisfy a range of biological and technical criteria: *ESR1*, *FOXA1*, and *GATA3*. The decision process involved the analysis of a public MCF7 RNA-seq data datasets obtained from Project Achilles[108] as well as single cell data generated from our lab. The Project Achilles data contains results of CRISPR KO screens for 18,333 genes in 625 cell lines. The dependency score is a measurement that indicates the

likelihood of how essential a gene in question is in a cell line after a KO, where a lower score is more essential. Scores were calculated for every gene using the CERES method, a computational method that considers copy-number to estimate gene dependency[97] (**Table 2**).

The Hallett lab had previously single cell profiled the transcriptomes of a set of nine BC cell lines that span the major BC subtypes, including MCF7-ATCC. We performed an unsupervised clustering of the MCF7 single cell data for each of the genes of interest (**Figure 25**). The single cell data generated by our lab was sequenced using Drop-seq and consists of 2400 MCF7 cells. We performed unsupervised clustering of the single cell expression profile for each of the genes of interest, represented as a UMAP (**Figure 25**). UMAP is a type of clustering that maps expression profiles of cells to a manifold, a multidimensional nonlinear surface. The idea is that cells with similar expression profiles will reside close to each other on this manifold and less similar cells will be more distant. Then to visualize, the UMAP projects this multidimensional nonlinear surface to two dimensions. The axes are unitless. As we are transforming high dimensional data (the expression of thousands of genes over thousands of cells) into two dimensions, it is dangerous to place too much meaning in large versus small gaps between disconnected clusters[98]. These UMAPs highlight that gene expression levels of *GATA3*, *FOXA1*, and *ESR1* are consistently highly expressed among cells, coinciding with bulk RNA-seq expression levels (**Figure 24**). Similarly, the dependency scores of these three TFs are the lowest among the set, suggesting that knocking them out would lead to stronger molecular signatures in response (**Table 2**). We also determined that the other three luminal differentiation controlling TFs mentioned in **Chapter 1** (*AR*, *SPDEF*, and *XBP1*) would be good candidate OE targets as efficiency of target gene activation is a function of baseline expression levels where lower expression allows for the OE CRISPR system to have a larger effect[35]. We chose to keep *ESR1* as a common TF for OE and KO to compare its effect in both scenarios, as we are using an ER+ cell line. Based on these results, we decided that

GATA3, *FOXA1* and *ESR1* would be suitable KO targets, as they consistently have high expression in MCF7 cells.

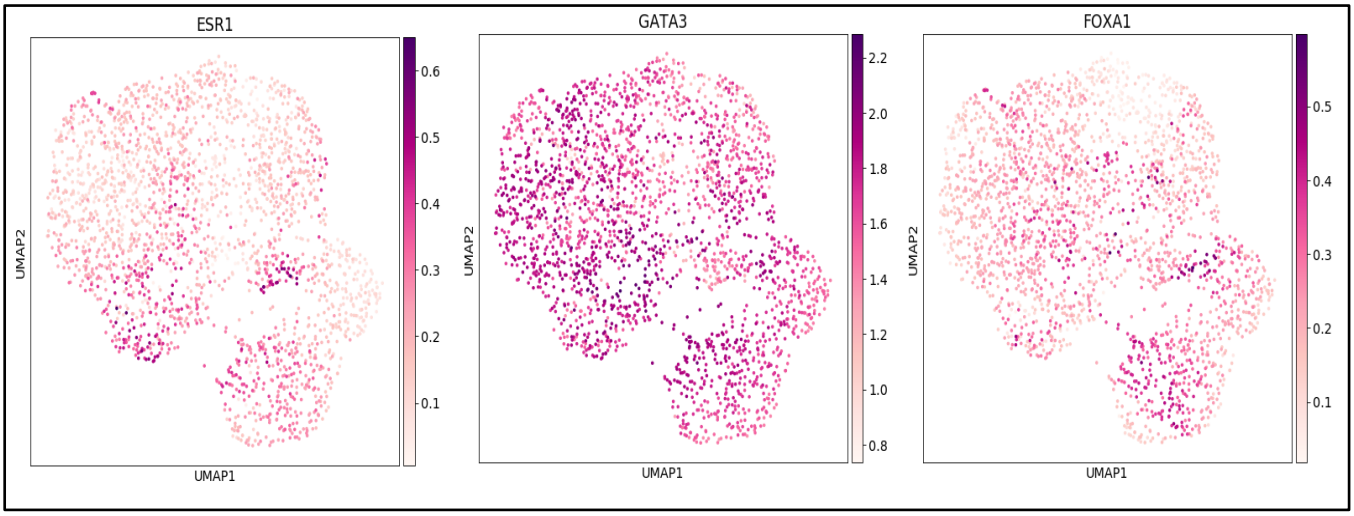


Figure 115. UMAP representation of the single cell expression of three important luminal TFs in MCF7-ATCC. Each dot represents one MCF7 cell. Gene expression levels are represented by the pre-imputed log counts and illustrated by the colour gradient. MCF7-ATCC cell line has a high baseline expression of the *canonical three* genes.

5.2 Implementation of CRISPR system for simultaneous KO and OE interventions

The COIN-seq system is designed to introduce KO and OE interventions concomitantly. Therefore, there are two types of Cas9 that would need to be introduced into the cells to perform both interventions. This is not feasible as we hypothesized that Cas9 toxicity would kill the cells and because either Cas9 would not be able to differentiate sgRNA targets as they can interact with either sgRNA. For example, if there are two sgRNAs (one targeting the promoter of a gene for OE and one targeting the exon of the same gene for KO), neither Cas9 would be able to differentiate which sgRNA is for which intervention type, leading to a loss of intervention specificity.

My primary contribution was in the design of our CRISPR implementation with the help of Sanny Khurdia and Abdelrahman Ahmed. Our intervention system follows the CRISPR/Cas9

system design from Dahlman et. al.[86] described in **Chapter 3.3 .** This design follows the OE dRNA approach of shortening the length of the protospacer which allows for the simultaneous KO and OE intervention of genes using the same CRISPR-based machinery. More specifically, the nature of the intervention is dictated by which sgRNA is incorporated into the cell.

We generated two clonal MCF7 cell lines (termed COIN MCF7): the first containing only Cas9 used for KO experiments and the second containing both Cas9 and MPH which can be used for KO and OE experiments (**Figure 23B**). We hypothesized that this lack of variability ensures that a change caused by a genetic intervention, such as a KO, will be consistent among cells in the same population, ablating any inconsistencies caused by either varying levels of endogenous protein expression or component expression. Transduced cells underwent antibiotic selection using blasticidin and hygromycin. In addition to the original protocol, we FACS sorted the survivors into single wells and cultured to make uniform clonal populations. We then transduce the Cas9 containing cells with the KO sgRNA plasmids and the Cas9 and MPH containing cells with the OE sgRNA plasmids separately.

5.4 Designing sgRNAs and protospacer sequences

For KO and OE interventions, we use lentiGuide-Puro and lenti sgRNA(MS2)_puro plasmids respectively (**Figure 23A**). The cells then express the sgRNA molecule, which interacts with the CRISPR machinery to KO or OE the target genes. Both sgRNAs use puromycin as a selection marker, which enables the survival of cells with either or both KO and OE sgRNAs.

The protospacer sequences for each of the three target genes in both KO and OE experiments were designed to target different regions or promoters of the gene as they perform their function based on location (**Sup. Met. 2**). The KO sgRNAs were designed to target the first 2-4 exons of each gene and the OE sgRNAs target the -200 to -1 region of the

promoter. These regions have been found to be the most effective at increasing the KO and OE efficiency[34,35]. In total, we designed 18 different sgRNAs using the recommended software from Joung and colleagues (**Sup. Met. 2**): three sgRNAs per gene for both KO and OE intervention type (**Table 1**)[84]. Our goal is not to validate many sgRNAs but to make best "computational guesses"; good experimental design and sophisticated computational biology/data science should be able to overcome noise and error introduced by malfunctioning biological components or human error.

5.4 Transcriptional barcoding system identifying specific interventions

To capture sgRNAs in addition to mRNA transcripts, we adapt the direct capture Perturb-seq methodology by modifying Drop-seq microparticles following the DART-seq protocol from Saikia et al.[93] (**Chapter 4.2**). We design the newly ligated custom oligo (**Table 3**) sequence to be complementary to the sequence we intend to capture, the 3' capture sequence of the sgRNA scaffold (**Figure 21**). This enables capture of sgRNA molecules and allows them to be captured, barcoded, and further processed in parallel to the similarly barcoded mRNA transcripts (**Figure 22**) (**Sup. Met. 5**).

To examine whether the modified microparticles are capturing sgRNA molecules, we used a population of characterized T47D cells expressing KO sgRNAs targeting the *ESR1* gene provided by the Mader lab. T47D is similar to MCF7 as they are both Luminal A BC cell lines. Furthermore, these experiments were performed in parallel to the creation of clonal COIN cells, which reduced time of development. These cells were transduced with the lentiCRISPRv2 (Addgene #52961) plasmid vector. LentiCRISPRv2 is the single-vector version of the two-vector system we use for KO interventions containing both the Cas9 and sgRNA expressing cassettes. Therefore, we hypothesized that using this system to test sgRNA capture would function as an accurate representation of sgRNA capture in the context of COIN-seq as the sgRNA scaffolds are identical.

Sanny Khurdia designed and performed an “in solution” experiment consisting of combining the DART-modified microparticles suspended in lysis buffer with the aforementioned T47D cells (**Chapter 4.2**) in a microfuge tube while omitting the addition of droplet generation oil. This resulted in a bulk RNA-seq-like assay, where the microparticles capture the cell transcriptomes in a non-isolated environment. The sample was then processed through the standard Drop-seq library construction protocol (**Figure 17, panels 7-9**). In a typical single cell sequencing assay, we would clean out the contaminants from the constructed cDNA library using AMPureXP beads and assess its DNA size and quality using the Agilent Tapestation 4150. However, the co-capture of sgRNAs generated two cDNA libraries: the original mRNA cDNA library and the guide library containing captured sgRNAs. Therefore, to isolate these libraries following PCR amplification of STAMPs, we processed the resultant cDNA through two rounds of AMPureXP purifications, allowing for size selective separation of the cDNA product. The first was conducted at a 0.6X ratio and then the supernatant was re-purified using a 1.2X ratio. This results in two libraries: a mRNA cDNA library and a guide library containing the captured sgRNAs (**Sup. Met. 7**).

Chapter 6 - Results

6.1 A Cas9-ready cell line for COIN-seq: MCF7 to study luminal subtype differentiation

Protein expression of Cas9 and MPH were characterized through immunoblotting assays (**Figure 23B**). To extract sample proteins, we performed a RIPA extraction using RIPA lysis buffer which enables the extraction of membrane, nuclear and cytoplasmic proteins from cultured mammalian cells. A primary antibody targeting Cas9 (**Sup. Met. 1**) was used, which resulted in a signal at 160 kDa for most of the clones (**Figure 26A**) indicating that the Cas9 protein is being expressed.

A similar validation was performed for the Cas9 + MPH clonal cell line as we used the same primary antibody for Cas9. Based on the composition of MPH described in **Chapter 3.3**, we target p65 using the anti-NF-κB p65 primary antibody to validate expression of the MPH protein. In addition, the entire NF-κB p65 protein is expressed in MCF7 at 65 kDa, which we use as a positive control. This resulted in a signal for p65 at 20 kDa, indicating that the protein is being expressed in MPH transduced cells (**Figure 26B**).

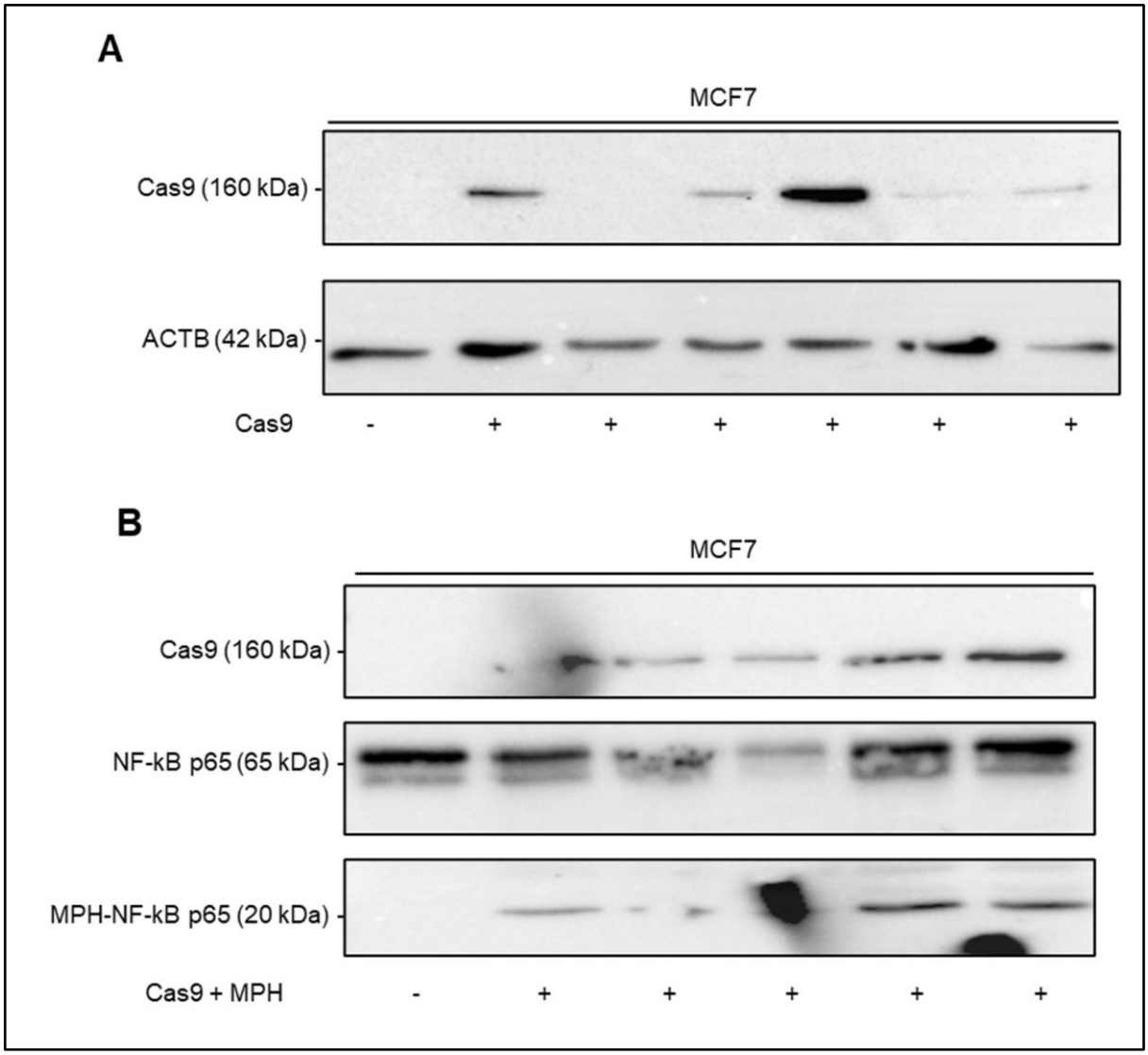


Figure 126. Western blot analysis of clonal cell lines. The first lane in each panel corresponds to the wild type MCF7-ATCC and other lanes correspond to unique clones. **A.** MCF7 clones transduced with Cas9. The bands characterize Cas9 protein expression in each clone. **B.** MCF7 clones transduced with both Cas9 and MPH. The bands validate the protein expression of both components in each clone. Successful clones were selected and expanded.

6.2 The lentivirus enabled CRISPR/Cas9 system

The protospacer sequences for each of the *canonical three* were cloned into their respective sgRNA plasmids using a Golden Gate reaction (**Chapter 5.4**). To validate that the correct sequence was inserted into the correct location, we performed Sanger sequencing

(**Figure S1**) (**Sup. Met. 2**). Almost all the sgRNAs validated apart from two; both OE-ESR1-sg1 and KO-NT1 (non-targeting), failed multiple validation attempts and were omitted (**Figure S1A**).

6.2.1 Determining antibiotic selection concentrations

We ran an antibiotic kill curve assay to determine the minimum blasticidin, hygromycin, and puromycin concentrations needed to kill all the COIN cells (**Table 4**). Cells transduced with Cas9 were selected with 7 µg/mL and maintained at 4 µg/mL blasticidin. The cells transduced with MPH were selected with 500 µg/mL and maintained at 250 µg/µL hygromycin. Lastly, the cells transduced with both sgRNA sets were then selected with 1 µg/mL puromycin.

6.2.2 KO Implementation

Validation of sgRNAs was performed via immunoblotting assays (**Figure 23C**) using methods identical to those described in **Chapter 6.1** apart from using puromycin for antibiotic selection for the sgRNAs (**Sup. Met. 4**). Intervention efficiency suggested by the immunoblots is represented by a naming scheme we created highlighting the effectiveness of the KO, where we select two out of three sgRNAs. The first and second most efficient sgRNAs are named major and minor sgRNAs, respectively. For the KO sgRNA set, we observed that most were effective at reducing the level of protein expression of the target genes, suggesting that the Cas9 in the clonal cells was functional (**Chapter 5.3**). Furthermore, none of the deletions were lethal to the population, as cell growth continued normally.

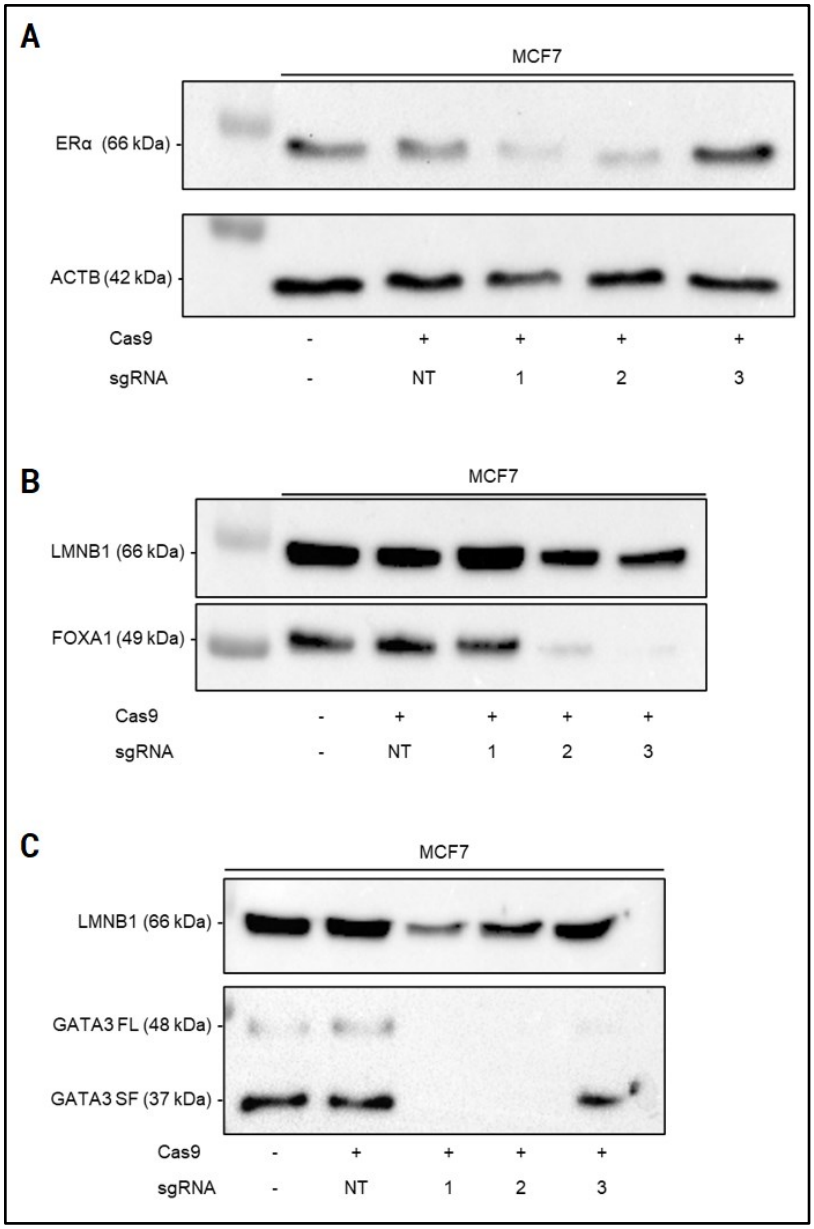


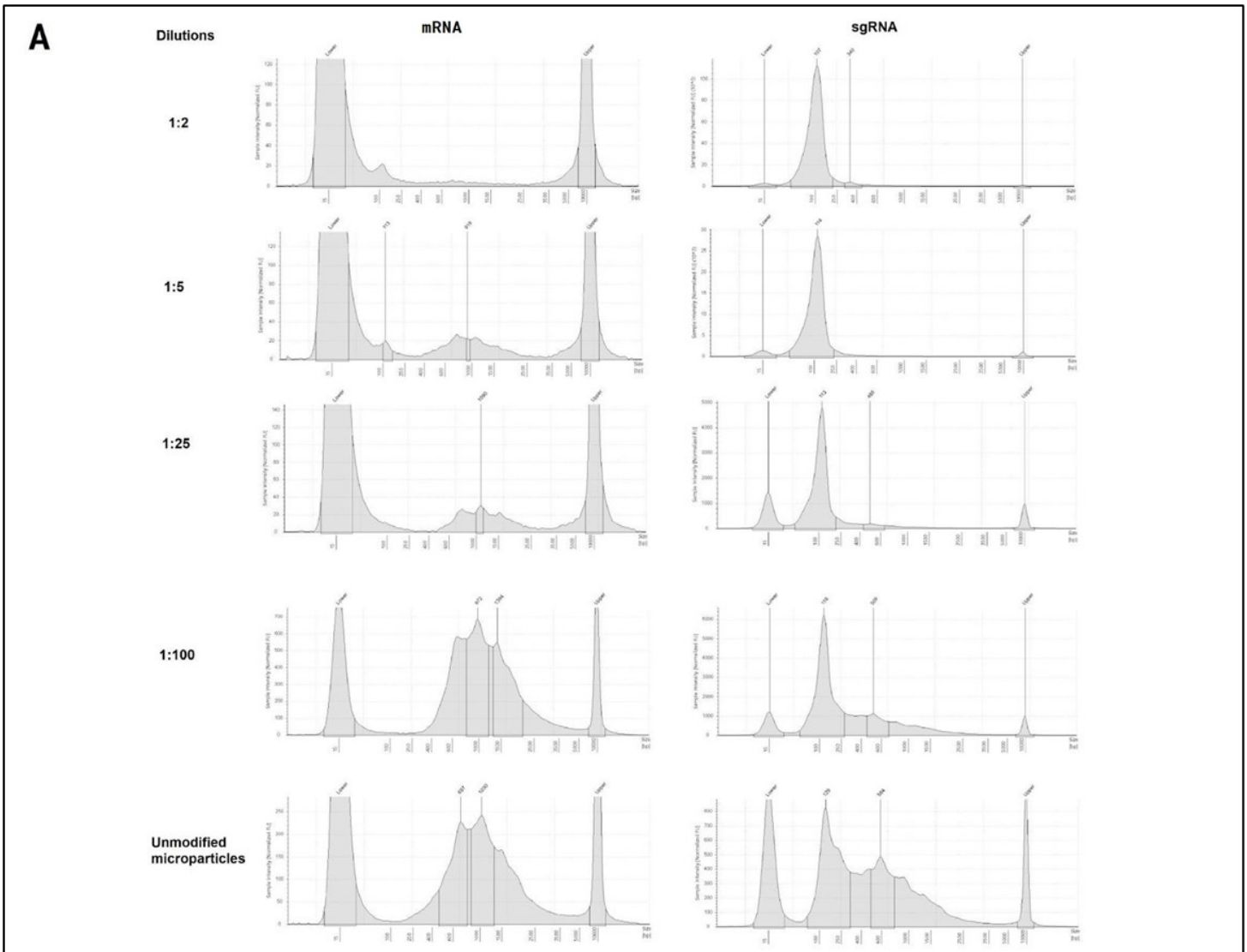
Figure 27. Protein KO validation of targeted genes using western blot. In each panel, the first lane corresponds to the wild type MCF7-ATCC. Second lane represents a negative control, consisting of cells expressing Cas9 and transduced with the non-targeting sgRNA. The remaining lanes illustrate the effects of the three sgRNAs targeting genes **A.** *ESR1*, **B.** *FOXA1*, and **C.** *GATA3* on protein expression. As expected, GATA3 protein is present as full length (FL) and splice form (SF). Protein expression of housekeeping genes ACTB and LMNB1 are used as positive loading controls.

The controls for all three immunoblots were wild type MCF7-ATCC and MCF7 with Cas9 and NT sgRNAs. We found that the *ESR1* KOs resulted in protein expression of ERα at

varying degrees (**Figure 27A**). We decided that sgRNAs one and two would be the major and minor sgRNAs based on which samples had the least protein expression. In the FOXA1 results, we observed almost no protein expression in sgRNA three (major) and slightly more in sgRNA two (minor) (**Figure 27B**). In GATA3, no protein expression was found in sgRNAs one and two (**Figure 27C**). Due to both sgRNAs completely ablating protein expression, we selected the first and second to be the major and minor sgRNAs respectively, following the order of efficiency determined *in silico*. The six sgRNAs were then used in the COIN-seq experiments (**Figure 23C**).

6.3 Capture of transcriptional barcodes

Following the experimental design outlined in **chapter 5.4**, we sought to evaluate the how much of the oligonucleotides tethered to the microparticles needed to be modified with the ligation reaction to efficiently capture sgRNAs and mRNA. We generated an mRNA cDNA library and analysed them using the Agilent Tapestation 4150 for different dilutions of the toehold probe (**Chapter 4.2**) ranging from 1:2 to 1:100 (**Figure 28A**). We observed that all dilutions (except for 1:100) resulted in lower mRNA cDNA library yield when compared to the unmodified microparticles while the sgRNA yield stayed consistent throughout. Furthermore, there appeared to be a reduction in the amount of sgRNA captured as the dilution rate increases, where the 1:100 diluted sample retains a six-fold increase of guide library capture when compared to the unmodified microparticles (**Figure 28B**). This suggests that using the 1:100 dilution allows for the capture of the guide library without sacrificing mRNA cDNA yield.



B

	mRNA			sgRNA		
Dilutions	Average Size (bp)	Conc. (pg/µL)	Region Molarity(pmol/L)	Average Size (bp)	Conc. (pg/µL)	Region Molarity(pmol/L)
1:2	-	-	-	110	74000	121000
1:5	960	35.2	64.4	113	3050	29900
1:25	1130	38.9	64.6	114	2340	36700
1:100	1080	1380	2310	135	3330	42800
Unmodified	1069	500	851	153	641	7340

Figure 28. Tapestation 4150 cDNA interrogation of mRNA and guide libraries using DART-seq microparticles at varying toehold concentrations. A. Quality control of mRNA and sgRNA libraries obtained with DART-modified microparticles at different toehold concentrations. **B.** Table showing region molarity of captured mRNA and sgRNA libraries. As the toehold concentrations decrease, we observe an increase in mRNA capture and a decrease in sgRNA capture.

6.4 Experimental design

The experimental design of our first implementation of COIN-seq consists of single cell sequencing of four experiments using our modified microparticles. We omit the use of OE interventions in our first pass of the main COIN-seq experiments due to technical issues.

Table 5A provides a summary of the experimental design explained here. **Experiment 1** is the MCF7 clonal cell line with Cas9. This experiment returns the transcriptional profile of MCF7-Cas9 cells, which is supposed to be similar to unmodified MCF7. These cells also function as the baseline cells used in the rest of the experiments. In **experiment 2**, we produce a pool of lentiviral particles containing the major and minor sgRNAs for each of the three genes and the NT-sgRNA for a total of seven sgRNAs. We then transduce the MCF7-Cas9 cells at a low MOI with the intention of inducing single sgRNA construct integrations per cell while minimizing combinations. In addition, these cells are not treated with puromycin selection. This results in eight expected gene profiles, which are defined by each individual intervention and the uninfected MCF7-Cas9 cells. **Experiment 3** builds upon to experiment 2 as it follows identical conditions with the addition of puromycin selection. We expect single sgRNAs to be introduced, but due to selection, we should not observe cells that were not transduced, generating seven expected gene profiles as there are no uninfected MCF7-Cas9 cells. In **experiment 4**, we pool four sgRNAs, the major guide for each of the three targets and the NT-sgRNA and transduce the cells with a high MOI. At this MOI, we expect that most cells will be transduced with at least one sgRNA and some will have a combination of sgRNAs present. In those events, we would be able to explore molecular epistasis among those genes. Single cell sequencing of experiment 1 is performed after blasticin selection and experiments 2-4 is performed at two time points, 7- and 14-days post transduction to observe the long-term effects of CRISPR exposure on gene expression profiles.

6.5 COIN-seq applied: Single cell sequencing and computational biology

Samples were transduced and cultured using the sgRNAs outlined in our experimental design (**Chapter 6.4**). The samples were single cell sequenced using the Drop-seq device with fresh reagents in combination with the DART modified microparticles. The captured material was sequenced with Illumina NEXT-seq in two batches of five samples each with an average of ~100M read/sample (**Table 5B**). The raw sequencing data was subjected to our bioinformatics pipeline for read processing, data normalization (**Sup. Met. 6-8**). **Table 5B** provides a summary of the sequencing related information for each sample. Post-sequencing, we discovered that there had been an error made when ordering the custom oligo that was annealed to the modified microparticles. Instead of the normal custom oligo (**Table 3**), there was an experimental error which resulted in the addition of an adenosine in the middle of the sequence (**Figure S2**). This created a non-specific oligonucleotide and resulted in the failure to capture sgRNA sequences throughout experiments 2-4.

In this thesis, we limit the analysis to a preliminary comparison of experiment 1, MCF7-ATCC cells with Cas9 against the “normal” MCF7-ATCC cell line data we had previously generated (**Chapter 5.1**) as a control. We begin by assessing the quality of the sample by observing the amount of mitochondrial transcripts found relative to gene transcripts in each cell. This is because low-quality or stressed cells often exhibit extensive mitochondrial contamination. Typically, cells are classified as such when the mitochondrial fraction is more than 5-10% in human cells[110]. We observe that a high proportion of cells exceed this mitochondrial fraction criteria in all samples which may be indicative of cell stress in the samples (**Figure 29**).

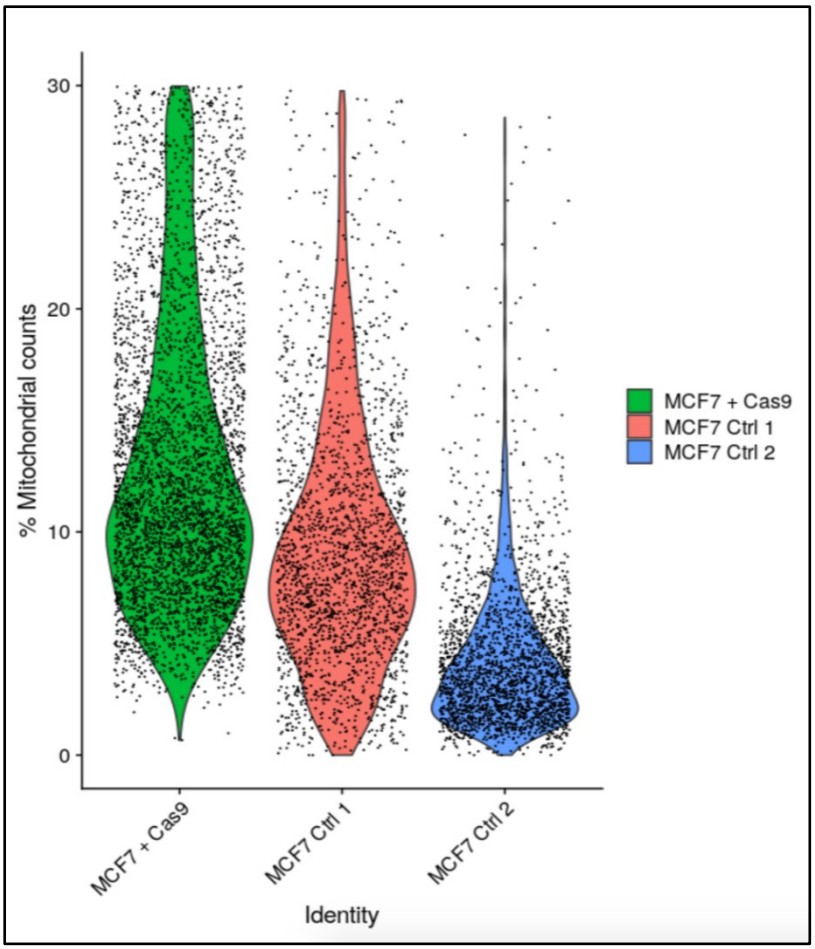


Figure 29. Mitochondrial fraction per cell between samples. Quality control examining the mitochondrial fraction of MCF7 controls and MCF7 + Cas9. Each dot represents a cell with over 200 and under 2500 genes per cell. Mitochondrial fraction threshold set to 30%.

We investigate the relationship between the two controls. These controls were generated from single cell sequencing the uninfected MCF7 cell line in two separate batches. This step aims to combine their profiles and remove any variation caused by non-biological factors related to sequencing in separate batches, known as batch effects. **Figure 30** provides an unsupervised UMAP clustering of the single cell expression profiles labelled with the population of origin. The control samples are processed using the open-source single cell toolkit Seurat. This method uses an unsupervised strategy to identify cell pairwise correspondences between single cells across datasets to generate anchor points. The

datasets are transformed using the anchor points into shared space that leads them to overlap and ablate any batch effects[100].

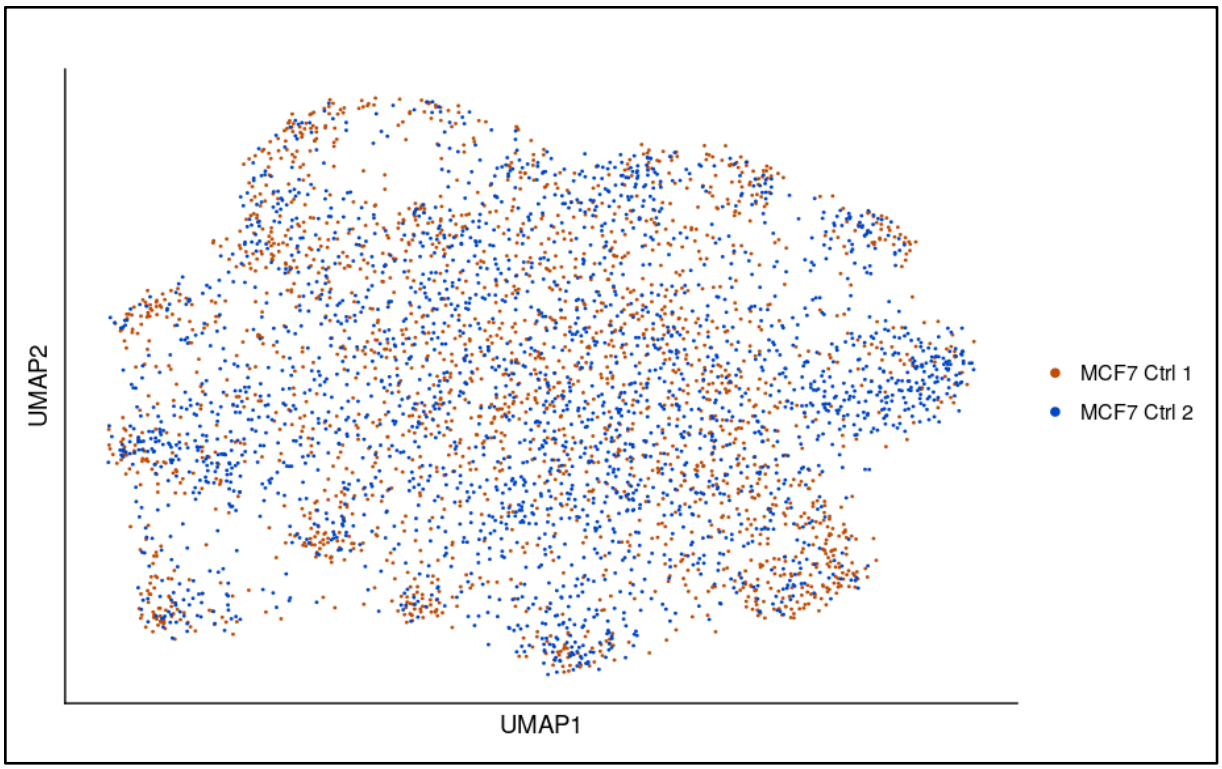


Figure 130. UMAP based visualization of the relationship between two control MCF7 samples. UMAP is constructed after integration of both control datasets following the Seurat workflow[100].

We hypothesized that the transcriptional profiles generated from experiment 1 should be similar to those of the control samples as Cas9 is not functional without sgRNAs. To test this hypothesis, we observe the relationship between our result from experiment 1 and the control MCF7 samples. In **figure 31**, we integrate the expression profile of experiment 1 into the existing dimensionality reduction of the control cell lines using Seurat. We observe that the experiment 1 samples do not significantly overlap with the control populations, although there is some intermixing. This suggests there is a significant difference in gene expression profiles between datasets. We performed differential gene expression analysis between the Cas9 group and the control groups using the Wilcoxon Rank Sum test implemented in

Seurat[100] and observed that mitochondrial and ribosomal protein genes are among the most overexpressed between conditions, indicating cellular stress in the Cas9 group (**Table 7**).

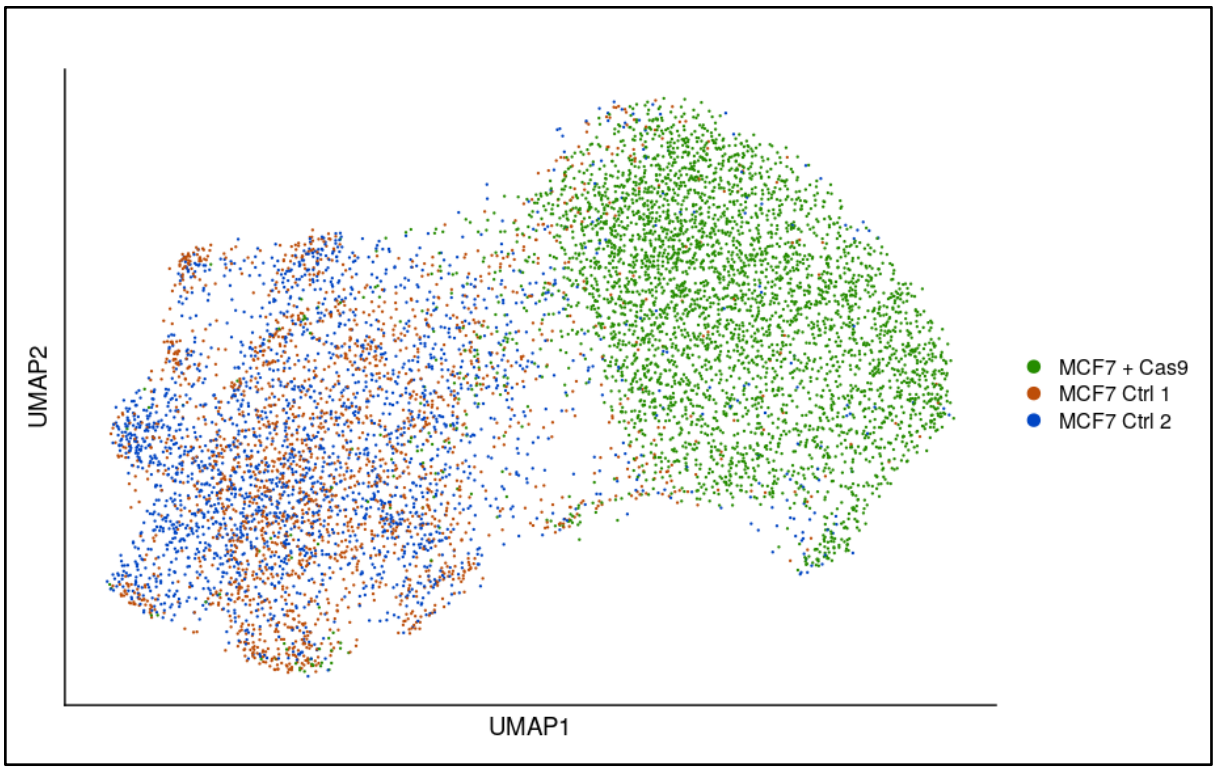


Figure 141. UMAP of the relationships between control MCF7 and MCF7 + Cas9. UMAP is constructed by integrating MCF7 + Cas9 dataset relative to control MCF7 expression profiles. There is minimal overlap between the two conditions.

In experiments 2, 3, and 4, the sequencing showed that we were able to recover RNA, however, the data produced was extremely noisy. The experiments were processed similarly to experiment 1. We performed differential expression analysis to identify distinct groups of genes expressed within each of the KO conditions. However, we were unable to determine any differences among the profiles. This suggests that the experiments failed at the transduction stage as the sgRNAs did not elicit a change in gene expression. This could be due to the MOI being too low, which would result in too few sgRNAs being transduced into the cells. Alternatively, it is possible that the MOI was sufficient, but the lentivirus particles failed to transduce. We attempted to validate the presence of sgRNAs in the cells. However, due to the ordering error of the custom oligo (**Figure S2**), the modified DART-seq microparticles were

unable to capture sgRNAs in each of the experiments which rendered us unable to confirm the presence of the sgRNAs in the cells.

Chapter 7 - Discussion & Concluding Remarks

Breast cancer is a heterogeneous disease. Understanding this heterogeneity has allowed researchers to stratify BC using the expression of a small set of proteins into molecular subtypes. Molecular profiling of BCs has further refined the classifications and led to several subtyping schemes through the identification of new subtypes and better characterization their underlying molecular processes. However, we still do not have a full understanding of the causative agents controlling BC subtype differentiation. The underlying genes and gene products controlling subtype fate remain largely undetermined as traditional experimental approaches to understand causality are limited.

We aimed to better understand BC subtype differentiation by applying genetic intervention platforms such as Perturb-seq, which provide a solution to the shortcomings found in typical genetic assays. These platforms enable the observation of higher order causal interactions through the analysis of transcriptional effects associated with genetic manipulations on genes, processes, and states. In addition, these platforms decrease the time and cost associated with assaying the complex effects of large numbers of perturbations. In this thesis, we build upon Perturb-seq by developing COIN-seq as a means of examining the effects of individual and combinations of CRISPR-mediated interventions in a high-throughput manner to infer the regulatory and causal networks underlying BC subtype differentiation. In this first implementation, we focus on the design and construction of COIN-seq with subsequent evaluation of its efficacy.

Throughout our implementation, we sought to remove technical variation. However, there are experimental components of RNA-seq which can contribute to variation. For example, a population composed of many identical cells can and will most likely exhibit some

form of variation in gene expression. We observe this phenomenon in **Figure 25**, where the expression of the *canonical three* genes varies among an MCF7-ATCC population that is supposedly isogenic. This variability is not inherently detrimental, and it can be argued that cells will always vary in expression due to different contributing factors, such as which stage of the cell cycle individual cells are currently in. In fact, COIN-seq offers the possibility to computationally filter out cells that present these sources of variation, making it more sensitive toward identifying the profiles of cells affected by CRISPR interventions. This considered, we determined that the best course of action to minimize additional variation would be to generate clonal cell lines transduced with the Cas9 and MPH components. Following this ideology, we chose to use the MCF7-ATCC cell line because it is a well characterized luminal BC cell line.

We transduced MCF7 with the aforementioned components and generated clones through single cell sorting. In this method, single cells were isolated into individual wells and allowed to expand, generating isogenic populations derived from a single cell. We selected viable clones through protein-based detection via immunoblot, which presented a set of hurdles. Most protein blots did not reveal any protein expression of Cas9 in any of the clonal samples. Although there was a possibility that the Cas9 expressing plasmid had not been transduced, numerous cells survived antibiotic selection (whereas a wild type MCF7 population would not have survived). We assumed the problem was that Cas9 is a large protein that was not being extracted correctly. Our approach to this problem was to modify different elements of the immunoblot assay in favour of larger proteins, such as longer member transfer times and more aggressive protein extraction, however, these did not resolve the issue.

We then assumed that the primary antibody against the Cas9 associated FLAG tag was non-functional and used a new one. However, this returned no results. Therefore, we switched the primary antibody to one that targets Cas9, which provided results suggesting Cas9 protein expression (**Figure 26A**). This assay was extended to the second clonal cell line transduced with both Cas9 and MPH, suggesting that both proteins were successfully being expressed (**Figure 26B**).

COIN-seq is designed in a way that intervention protospacer sequences can be designed *in silico* and delivered without the need for intermediate validation. We designed protospacer sequences targeting *ESR1*, *GATA3*, and *FOXA1* for the KO and OE interventions and ligated them into their respective plasmids, validated using Sanger sequencing (**Figure S1**). These plasmids are then transduced individually into Cas9 and Cas9 + MPH cells, leading to the expression of sgRNA molecules containing the protospacers. The gene KO efficiency of each protospacer sequence was evaluated by observing the protein expression of each target. In each of the three genes, we observed partial or complete protein reduction in two of three sgRNAs, suggesting that the sgRNAs in combination with Cas9 are functional (**Figure 27**). These interventions were designed to KO a gene. By performing immunoblots on the population of surviving cells, we create an environment in which some of the surviving cells still express the target protein to some degree. We observe some reductions in protein expression, such as sgRNA-1 against *ESR1* (**Figure 27A**). This further highlights that we do not observe 100% efficiency and emphasizes the potential utility of COIN-seq towards filtering variation.

There are several reasons this may occur. Interventions may have failed to perform a homozygous KO of the gene of interest, leading to reduced protein expression rather than completely ablating it. Furthermore, the cell cycle related temporal regulation of these genes

in an asynchronous population may result in the variability of protein expression; this effect may be responsible for slight differences in protein expression among different samples containing a heterozygous KO. These results suggest that KO interventions successfully affected protein expression via gene targeted KO. The functionality of the OE plasmids was not evaluated in this work due to transduced cells not surviving antibiotic selection. However, we diagnosed the issue to be caused by a faulty Kozak sequence in the promoter of the puromycin resistance cassette of the plasmid. We concluded that this faulty sequence may have been affecting expression of the cassette, leading to cells which were more sensitive to higher concentrations of the antibiotic. We solved this issue by tuning the concentration of puromycin to one low enough to kill wild type cells while allowing cells transduced with the OE plasmid to grow.

Modifying Drop-seq microparticles to enable co-capture of mRNA transcripts and sgRNAs was necessary to allow the identification of interventions present. We hypothesize that the sgRNAs would act as a molecular barcode represented by the unique protospacer sequences for each genetic target. We perform the microparticle modifications following the DART-seq protocol, process the T47D samples using the in solution experimental approach, and evaluate sgRNA capture via Agilent Tapestation 4150 analysis. Although the DART ligations reactions had been successful, mRNA yield was significantly reduced. We hypothesized that this observed reduction was the result of the modification (through a ligation reaction) of a large fraction of the oligonucleotides tethered to the microparticles, which interfered with mRNA capture. We tested this hypothesis by performing an additional in solution experiment using a range of diluted toehold probe concentrations (**Figure 28**). These results support that our modifications were successful at co-capturing both mRNA and guide libraries. In the future, it would be of interest to further tune the toehold dilution rate with the purpose of optimizing co-capture of both libraries.

In the first implementation of COIN-seq, we designed experiments allowing for the exploration of both single gene and epistatic effects. Experiment 1 functions as a control representing the transcriptional profile of MCF7 + Cas9. In experiment 2, cells are transduced at a low MOI with the major minor sgRNAs for each gene target and do not undergo antibiotic selection to produce transcriptional profiles for cells transduced individually with each sgRNA and non-transduced cells. Experiment 3 is a continuation of experiment 2 with the addition of puromycin and serves to produce transcriptional profiles of transduced cells. We expect a margin of error in which some cells will not have captured an associated sgRNA barcode. However, these cells can be differentiated using the expression profiles from experiment 2 which, in turn, provides an estimate of our sgRNA capture rate. Samples in experiment 4 are transduced at a high MOI which increases the likelihood of obtaining cells with multiple interventions present and should allow the exploration of epistasis.

Generation of cell populations following the design of experiments 1-4 and processing them using single cell sequencing had been successfully completed. The transcriptional profile from experiment 1 compared to wild type MCF7-ATCC cells partition with no significant overlap (**Figure 31**). However, we do observe several cells from the controls that are present between the two clusters and in MCF7-Cas9 cluster, suggesting some similarity among the populations. The quality control related to this experiment suggests that the cells were undergoing stress due to the large mitochondrial fraction, which is a specific indicator of high stress (**Figure 29**). Furthermore, we performed differential gene expression analysis and observed a significant increase in the expression of mitochondrial and ribosomal genes (**Table 6**), supporting the notion that the cells were incredibly stressed. We believe that the observed stress response is introduced by human error in sample handling before Drop-seq. Therefore, it may be necessary to either repeat experiment 1 to ablate this response. In this

thesis, we do not explore experiments 2-4 in detail, however, we observe that human error was present in the generation of the DART-seq generated microparticles and that the data generated is too noisy to interpret. To prevent this error from occurring in the future, we would implement stringent revision protocols in which more than one person would review the oligonucleotides used throughout the experiments. This would act as a safeguard against human error, as these errors can be caught early and without additional effort.

Conclusion

As a result of this research, we have prepared the foundation for COIN-seq to be further built upon. The generation and interpretation of single and combinatorial genetic interventions in BC is a technique that has not yet been performed in this domain. In addition, we believe that this design is consistent and would permit the future inclusion of KD assays[33]. This should be possible, as CRISPR-based KD assays utilize an inactivated Cas9 similarly to the OE interventions, but we have not explored this idea further. The field of pooled genetic intervention screens continues to rapidly develop, as new technologies are being designed and implemented to improve and replace existing ones. For example, there now exists a wider array of CRISPR enzyme types such as Cas13, which shows high RNA knock-down efficacy with minimal off-target activity[101,104]. In addition, single cell technologies have expanded to include methods that utilize DNA barcodes to clonally trace transcriptomes over time[102]. This development is particularly applicable in the context of this thesis, as it would enable the tracing of differentiation cause by the induced intervention.

Tables

Table 1. Table showing guide RNA sequences and the intended genetic targets. Sequence positions referenced using human genome assembly GRCh38.p13.

Gene Target	Condition	Primer Name	Sequence (5′ - 3′)	Target
ESR1	KO	sgESR1_1_FW	CACCGTACCTGGAGAACGAGCCCAG	chr6:151808303-151808322
		sgESR1_1_RV	AAACCTGGGCTCGTTCTCCAGGTAC	
		sgESR1_2_FW	CACCGCGCGGCGTTGAACTCGTAGG	chr6:151808088-151808335
		sgESR1_2_RV	AAACCCTACGAGTTCAACGCCGCGC	
		sgESR1_3_FW	CACCGTCAGATAATCGACGCCAGGG	chr6:151842604-151842623
		sgESR1_3_RV	AAACCCCTGGCGTCGATTATCTGAC	
FOXA1		sgFOXA1_1_FW	CACCGCAAGTGCGAGAAGCAGCCGG	chr14:37591967-37591986
		sgFOXA1_1_RV	AAACCCGGCTGCTTCTCGCACTTGC	
		sgFOXA1_2_FW	CACCGGGACATGTTGAAGGACGCCG	chr14:37592583-37592602
		sgFOXA1_2_RV	AAACCGGCGTCCTTCAACATGTCCC	
		sgFOXA1_3_FW	CACCGCCACAAACTAGAATGTCTGG	chr14:37591025-37591044
		sgFOXA1_3_RV	AAACCCAGACATTCTAGTTTGTGGC	
GATA3		sgGATA3_1_FW	CACCGCGGAGGGTACCTCTGCACCG	chr10:8055867-8055899
		sgGATA3_1_RV	AAACCGGTGCAGAGGTACCCTCCGC	
		sgGATA3_2_FW	CACCGGCTGCCCGTTGAGCACGGCG	chr10:8055703-8055871
		sgGATA3_2_RV	AAACCGCCGTGCTCAACGGGCAGCC	
		sgGATA3_3_FW	CACCGCCACAAACTAGAATGTCTGG	chr10:37591025-37591044

		sgGATA3_3_RV	AAACCCAGACATTCTAGTTTGTGGC	
Control 1		NT1-FW	CACCGCTGAAAAAGGAAGGAGTTGA	NA
		NT1-RV	AAACTCAACTCCTTCCTTTTTCAGC	
Control 2		NT2-FW	CACCGAAGATGAAAGGAAAGGCGTT	NA
		NT2-RV	AAACAACGCCTTTCCTTTCATCTTC	
ESR1	OE	sgESR1_1_FW	CACCGCACCACCATTTGACT	chr6:151690442-151690456
		sgESR1_1_RV	AAACAGTCAAATGGTGGTGC	
		sgESR1_2_FW	CACCGATCCTAGTCAAATGG	chr6:151690437-151690451
		sgESR1_2_RV	AAACCCATTTGACTAGGATC	
		sgESR1_3_FW	CACCGTCATTGAAAAAATAG	chr6:151690413-151690427
		sgESR1_3_RV	AAACCTATTTTTTCAATGAC	
FOXA1		sgFOXA1_1_FW	CACCGGGTGCACCTGCAAGG	chr14:37595362-37595376
		sgFOXA1_1_RV	AAACCCTTGCAGGTGCACCC	
		sgFOXA1_2_FW	CACCGTGCGGCGGACAAATG	chr14:37595319-37595333
		sgFOXA1_2_RV	AAACCATTTGTCCGCCGCAC	
		sgFOXA1_3_FW	CACCGCACCTACAAAGCCCG	chr14:-37595346-37595360
		sgFOXA1_3_RV	AAACCGGGCTTTGTAGGTGC	
GATA3		sgGATA3_1_FW	CACCGAGGATCCCCGGCACA	chr10:8054585-8054659
		sgGATA3_1_RV	AAACTGTGCCGGGGATCCTC	
		sgGATA3_2_FW	CACCGAGTTTCCTTGTGCCG	chr10:8054577-8054591
		sgGATA3_2_RV	AAACCGGCACAAGGAAACT C	

		sgGATA3_3_FW	CACCGACCCAAACCCGCTCC	chr10:8054559-8054573
		sgGATA3_3_RV	AAACGGAGCGGGTTTGGGTC	
Control 1		NT1-FW	CACCGAAAGGAAGGAGTTGA	NA
		NT1-RV	AAACTCAACTCCTTCCTTTC	
Control 2		NT2-FW	CACCGGAAAGGAAAGGCGTT	NA
		NT2-RV	AAACAACGCCTTTCCTTTCC	

Table 2. Dependency scores and gene expression of the *canonical six* in MCF7-ATCC. CERES dependency scores based on cell depletion assay data. Lower dependency scores indicate a higher likelihood that genes of interest are essential in the cell line. Data obtained from Project Achilles.

TF	Dependency Score (CERES)	Expression (TPM Log2)
FOXA1	-2.00	6.77
GATA3	-1.33	8.49
ESR1	-1.18	5.41
SPDEF	-0.25	6.44
AR	-0.16	2.84
XBP1	-0.05	9.24

Table 3. Sequences of all the oligonucleotides used for Drop-seq and DART-seq

Oligo Name	**Sequence**
Original Oligo-dt extruding from bead J's= Cell barcode N's= UMI	–Bead-Linker--- TTTTTTTAAGCAGTGGTATCAACGCAGAGTACJJJJJJJJJJJJ NNNNNNNNTTTTTTTTTTTTTTTTTTTTTTTTTTTT
Splint_Oligo	CGGTCTTCCCCAAAAAAAAAAA

Custom_Oligo	/5Phos/GGGGAAGACCGAAAAGCAACCGACTCG
New Modified oligo-dt	–Bead-Linker--- TTTTTTTAAGCAGTGGTATCAACGCAGAGTACJJJJJJJJJJJJ NNNNNNNNTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGGGA AGACCGAAAAGCAACCGACTCG
TSO	AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG
SMART_PCR	AAGCAGTGGTATCAACGCAGAGT

Table 4. Kill curve results for selection antibiotics in MCF7

Antibiotic	Concentration range	Selection time	Selection concentration	Maintenance concentration
Blasticidin	1 - 10 µg/ml	~ 5 days	7 µg/ml	4 µg/ml
Hygromycin B	0.1 - 1 mg/ml	~ 7 days	0.5 mg/ml	0.25 mg/ml
Puromycin	0.25 - 2 µg/ml	~ 3 - 4 days	1 µg/ml	-

Table 5. A. Summary table outlining experiments, experimental conditions, and sequencing parameters. **B.** Summary of alignment and mapping statistics. Reads were aligned against the GRCh38 human reference genome using STAR aligner.

A. Experimental design									B. Alignment statistics									
Experiment	EUID	SUID	Name	Batch	Replicate	Timepoint	Adapter	MOI	# reads	Avg read length	Uniquely mapped	% Unique Mapped	Avg Map Length	Number of splices	Mismatch Base %	% multi loci reads	% too short	% unmapped other
1	1	1	CAS9	2	1	1	N703	-	156,591,432	56	95,300,853	60.86	59.47	6,664,306	0.89	13.11	18.2	6.8
2	2	2	LO_NO_PURO	1	1	7	N701	Low	107,272,324	59	76,620,522	71.43	59.87	10,341,248	0.87	15.9	11	1.08
2	3	3	LO_NO_PURO	2	1	14	N701	Low	67,612,053	62	52,926,439	78.28	61.78	8,936,789	1.01	13.13	7.45	0.63
3	4	4	LO_PURO	1	1	7	N702	Low	104,814,570	60	78,422,095	74.82	60.94	10,594,183	0.91	15.38	8.3	0.96
3	5	5	LO_PURO	2	1	14	N702	Low	101,666,631	58	68,953,926	67.82	58.88	8,283,139	0.79	15.08	13.9	2.27
4	6	6	EPISTASIS	1	1	7	N705	High	147,399,069	59	102,888,570	69.8	59.48	14,239,147	0.84	15.57	12.8	1.15
4	6	7	EPISTASIS	1	2	7	N707	High	63,849,543	61	47,720,313	74.74	61.56	7,533,460	0.93	16.68	7.48	0.65
4	6	8	EPISTASIS	1	3	7	N712	High	108,758,138	60	81,264,892	74.72	60.91	12,838,604	0.88	16.01	7.53	1.09
4	7	9	EPISTASIS	2	1	14	N707	High	91,616,766	61	70,010,671	76.42	61.17	9,414,613	0.85	14.65	7.31	1.03
4	7	10	EPISTASIS	2	2	14	N712	High	95,484,113	61	73,077,034	76.53	61.37	9,634,531	0.87	13.5	8.3	1.06

Table 6. Top 20 differentially expressed genes in Cas9_2_1 (Experiment 1) compared to control MCF7 datasets.

Gene	p-value	avg_log2FC	pct.1	pct.2	p_val_adj	cluster
MT-CO2	< 0.001	2.179	0.992	0.517	< 0.001	CAS9_2_1
SP100	< 0.001	2.168	0.958	0.174	< 0.001	CAS9_2_1
MT-CO3	< 0.001	1.998	0.985	0.522	< 0.001	CAS9_2_1
MT-CYB	< 0.001	1.734	0.97	0.433	< 0.001	CAS9_2_1
MT-ATP6	< 0.001	1.685	0.961	0.45	< 0.001	CAS9_2_1
MT-RNR2	< 0.001	1.648	1	0.991	< 0.001	CAS9_2_1
MT-RNR1	< 0.001	1.560	0.911	0.329	< 0.001	CAS9_2_1
SH3KBP1	< 0.001	1.512	0.773	0.051	< 0.001	CAS9_2_1
IGF2R	< 0.001	1.508	0.82	0.113	< 0.001	CAS9_2_1
MT-ND4	< 0.001	1.449	0.972	0.496	< 0.001	CAS9_2_1
MT-ND3	< 0.001	1.399	0.811	0.208	< 0.001	CAS9_2_1
MTRNR2L12	< 0.001	1.360	0.838	0.372	< 0.001	CAS9_2_1
MT-ND1	< 0.001	1.211	0.833	0.273	< 0.001	CAS9_2_1
RPS29	< 0.001	1.205	0.905	0.464	< 0.001	CAS9_2_1
RPL37A	< 0.001	1.146	0.898	0.445	< 0.001	CAS9_2_1
HSP90AA1	< 0.001	1.120	0.933	0.574	< 0.001	CAS9_2_1
PRRG3	< 0.001	1.083	0.61	0.056	< 0.001	CAS9_2_1
KRT8	< 0.001	1.060	0.965	0.71	< 0.001	CAS9_2_1
MTCO1P12	< 0.001	1.022	0.881	0.387	< 0.001	CAS9_2_1
MTATP6P1	< 0.001	0.983	0.73	0.28	< 0.001	CAS9_2_1
ACTG1	< 0.001	0.978	0.77	0.399	< 0.001	CAS9_2_1

Table 7a. Nextera Index Sequences used for tagmentation of cDNA

Nextera Index	Sequences
N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG
N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG
N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGG
N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGG
N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGG
N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGG

Table 7b. Oligonucleotides used to ligate the i7 index onto the sgRNA construct

Sequence Name	Sequence
Nxt_Nst_701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAGTATTTCTAGCTCTAAA*A*C
Next_Nst_702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAGTATTTCTAGCTCTAAA*A*C
Next_Nst_703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAGTATTTCTAGCTCTAAA*A*C
Next_Nst_705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAGTATTTCTAGCTCTAAA*A*C
Next_Nst_707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGAT GTGTATAAGAGACAGTATTTCTAGCTCTAAA*A*C
Next_Nst_712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAGTATTTCTAGCTCTAAA*A*C

References

1. Ji P, Gong Y, Jin M-L, Hu X, Di G-H, Shao Z-M. The Burden and Trends of Breast Cancer From 1990 to 2017 at the Global, Regional, and National Levels: Results From the Global Burden of Disease Study 2017. Frontiers in oncology. 2020;10:650.

2. Breast cancer statistics - Canadian Cancer Society. [accessed 2021 Aug 10]. https://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on

3. Kamińska M, Ciszewski T, Łopacka-Szatan K, Miotła P, Starosławska E. Breast cancer risk factors. Przeglad menopauzalny. 2015;14(3):196–202.

4. Schonfeld SJ, Pfeiffer RM, Lacey JV Jr, Berrington de González A, Doody MM, Greenlee RT, Park Y, Schairer C, Schatzkin A, Sigurdson AJ, et al. Hormone-related risk factors and postmenopausal breast cancer among nulliparous versus parous women: An aggregated study. American journal of epidemiology. 2011;173(5):509–517.

5. Makki J. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. Clinical medicine insights. Pathology. 2015;8:23–31.

6. Ductal carcinoma - Canadian cancer society. [accessed 2021 Aug 12]. https://www.cancer.ca/en/cancer-information/cancer-type/breast/breast-cancer/cancerous-tumours/ductal-carcinoma/?region=on

7. Heppner GH. Tumor heterogeneity. Cancer research. 1984;44(6):2259–2265.

8. Ursini-Siegel J, Schade B, Cardiff RD, Muller WJ. Insights from transgenic mouse models of ERBB2-induced breast cancer. Nature reviews. Cancer. 2007;7(5):389–397.

9. Gruvberger S, Ringnér M, Chen Y, Panavally S, Saal LH, Borg A, Fernö M, Peterson C, Meltzer PS. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer research. 2001;61(16):5979–5984.

10. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. Nature. 8/2000 [accessed 2020 May 28];406(6797):747–752.

11. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. The Lancet. 11/2011 [accessed 2020 Jul 13];378(9805):1812–1823.

12. Rosen RD, Sapra A. TNM Classification. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2021.

13. Staging cancer - Canadian cancer society. [accessed 2021 Aug 10]. https://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/stage-and-grade/staging/?region=on

14. Greenberg PA, Hortobagyi GN, Smith TL, Ziegler LD, Frye DK, Buzdar AU. Long-term follow-up of patients with complete remission following combination chemotherapy for metastatic breast cancer. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 1996;14(8):2197–2205.

15. Mishra A, Verma M. Cancer biomarkers: are we ready for the prime time? Cancers. 2010;2(1):190–208.

16. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2009 [accessed 2020 Apr 17];27(8):1160–1167.

17. Shagufta, Ahmad I. Tamoxifen a pioneering drug: An update on the therapeutic potential of tamoxifen derivatives. European journal of medicinal chemistry. 2018;143:515–531.

18. Manso L, Moreno F, Márquez R, Castelo B, Arcediano A, Arroyo M, Ballesteros AI, Calvo I, Echarri MJ, Enrech S, et al. Use of bevacizumab as a first-line treatment for metastatic breast cancer. Current oncology . 2015;22(2):e51–60.

19. Tofigh A, Suderman M, Paquet ER, Livingstone J, Bertos N, Saleh SM, Zhao H, Souleimanova M, Cory S, Lesurf R, et al. The Prognostic Ease and Difficulty of Invasive Breast Carcinoma. Cell reports. 10/2014 [accessed 2020 May 16];9(1):129–142.

20. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001 [accessed 2020 May 16];98(19):10869–10874.

21. Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, Bibeau F, MacGrogan G, Lerebours F, Finetti P, Longy M, et al. A refined molecular taxonomy of breast cancer. Oncogene. 2012;31(9):1196–1206.

22. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of clinical investigation. 2011;121(7):2750–2767.

23. Ohnstad HO, Borgen E, Falk RS, Lien TG, Aaserud M, Sveli MAT, Kyte JA, Kristensen VN, Geitvik GA, Schlichting E, et al. Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. Breast cancer research: BCR. 2017;19(1):120.

24. Viale G. The current state of breast cancer classification. Annals of oncology: official journal of the European Society for Medical Oncology / ESMO. 2012;23 Suppl 10:x207–10.

25. Ignatiadis M, Sotiriou C. Luminal breast cancer: from biology to treatment. Nature reviews. Clinical oncology. 2013;10(9):494–506.

26. Sultan Mahmud M, Islam MS, Rahman MA. Smart Fire Detection System with Early Notifications Using Machine Learning. International journal of computational intelligence and applications. 2017;16(02):1750009.

27. Prat A, Perou CM. Mammary development meets cancer genomics. Nature medicine. 8/2009 [accessed 2020 Apr 17];15(8):842–844.

28. Visvader JE, Stingl J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. Genes & development. 2014;28(11):1143–1158.

29. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C. Biological processes associated with breast cancer clinical outcome

depend on the molecular subtypes. Clinical cancer research: an official journal of the American Association for Cancer Research. 2008;14(16):5158–5165.

30. Rakha EA, Reis-Filho JS, Ellis IO. Combinatorial biomarker expression in breast cancer. Breast cancer research and treatment. 2010;120(2):293–308.

31. Ismail H. Mechanisms Controlling Luminal Identity of Breast Tumours. 2018. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/21181

32. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 12/2016 [accessed 2019 May 4];167(7):1853–1866.e17.

33. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell. 2016;167(7):1867–1882.e21.

34. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nature methods. 2014;11(8):783–784.

35. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature. 2015;517(7536):583–588.

36. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell. 2013;152(5):1173–1183.

37. Zucca-Matthes G, Urban C, Vallejo A. Anatomy of the nipple and breast ducts. Gland Surgery. 2016;5(1):32–36.

38. Human biology online lab / organ histology of “mammary gland” by azure Qian. [accessed 2021 Aug 10]. http://humanbiologylab.pbworks.com/w/page/138522819/Organ%20Histology%C2%A0of%C2%A0%22Mammary%20Gland%22%C2%A0by%20Azure%20Qian

39. Hennighausen L, Robinson GW. Information networks in the mammary gland. Nature reviews. Molecular cell biology. 9/2005 [accessed 2020 May 11];6(9):715–725.

40. Muschler J, Streuli CH. Cell–Matrix Interactions in Mammary Gland Development and Breast Cancer. Cold Spring Harbor perspectives in biology. 2010;2(10). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2944360/. doi:10.1101/cshperspect.a003202

41. Kothari C, Diorio C, Durocher F. The Importance of Breast Adipose Tissue in Breast Cancer. International journal of molecular sciences. 2020;21(16):5760.

42. Avagliano A, Fiume G, Ruocco MR, Martucci N, Vecchio E, Insabato L, Russo D, Accurso A, Masone S, Montagnani S, et al. Influence of Fibroblasts on Mammary Gland Development, Breast Cancer Microenvironment Remodeling, and Cancer Cell Dissemination. Cancers. 2020;12(6):1697.

43. Oskarsson T. Extracellular matrix components in breast cancer progression and metastasis. Breast . 2013;22:S66–S72.

44. Paine IS, Lewis MT. The Terminal End Bud: the Little Engine that Could. Journal of mammary gland biology and neoplasia. 6/2017 [accessed 2020 May 16];22(2):93–108.

45. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat M-L, Gyorki DE, Ward T, Partanen A, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nature medicine. 2009 [accessed 2020 Jun 23];15(8):907–913.

46. Stingl J, Eaves CJ, Zandieh I, Emerman JT. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. Breast cancer research and treatment. 5/2001 [accessed 2020 Jun 29];67(2):93–109.

47. Scheidereit C, Krauter P, von der Ahe D, Janich S, Rabenau O, Cato AC, Suske G, Westphal HM, Beato M. Mechanism of gene regulation by steroid hormones. Journal of steroid biochemistry. 1986;24(1):19–24.

48. Truss M, Beato M. Steroid hormone receptors: interaction with deoxyribonucleic acid and transcription factors. Endocrine reviews. 1993;14(4):459–479.

49. Angus L, Beije N, Jager A, Martens JWM, Sleijfer S. ESR1 mutations: Moving towards guiding treatment decision-making in metastatic breast cancer patients. Cancer treatment reviews. 2017;52:33–40.

50. Kumar R, Zakharov MN, Khan SH, Miki R, Jang H, Toraldo G, Singh R, Bhasin S, Jasuja R. The Dynamic Structure of the Estrogen Receptor. Journal of amino acids. 2011;2011:812540.

51. Fuentes N, Silveyra P. Estrogen receptor signaling mechanisms. Advances in protein chemistry and structural biology. 2019;116:135–170.

52. Cui J, Shen Y, Li R. Estrogen synthesis and signaling pathways during aging: from periphery to brain. Trends in molecular medicine. 2013;19(3):197–209.

53. Lam EW-F, Brosens JJ, Gomes AR, Koo C-Y. Forkhead box proteins: tuning forks for transcriptional harmony. Nature reviews. Cancer. 2013;13(7):482–495.

54. Augello MA, Hickey TE, Knudsen KE. FOXA1: master of steroid receptor function in cancer: FOXA1: master of steroid receptor function in cancer. The EMBO journal. 2011;30(19):3885–3894.

55. Kaufmann E, Knöchel W. Five years on the wings of fork head. Mechanisms of development. 1996;57(1):3–20.

56. Kaestner KH. The Hepatocyte Nuclear Factor 3 (HNF3 or FOXA) Family in Metabolism. Trends in endocrinology and metabolism: TEM. 2000;11(7):281–285.

57. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes & development. 2011;25(21):2227–2241.

58. Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. Molecular cell. 2002;9(2):279–289.

59. Wijchers PJEC, Burbach JPH, Smidt MP. In control of biology: of mice, men and Foxes. Biochemical Journal. 2006;397(2):233–246.

60. Clark KL, Halay ED, Lai E, Burley SK. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature. 1993;364(6436):412–420.

61. Goytisolo FA, Gerchman SE, Yu X, Rees C, Graziano V, Ramakrishnan V, Thomas JO. Identification of two DNA-binding sites on the globular domain of histone H5. The EMBO journal. 1996;15(13):3421–3429.

62. Cirillo LA. Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. The EMBO journal. 1998;17(1):244–254.

63. Bernardo GM, Lozada KL, Miedler JD, Harburg G, Hewitt SC, Mosley JD, Godwin AK, Korach KS, Visvader JE, Kaestner KH, et al. FOXA1 is an essential determinant of ER expression and mammary ductal morphogenesis. Development . 2010;137(12):2045–2054.

64. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M. FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. Cell. 2008;132(6):958–970.

65. Nakshatri H, Badve S. FOXA1 in breast cancer. Expert reviews in molecular medicine. 2009;11:e8.

66. Serandour AA, Avner S, Percevault F, Demay F, Bizot M, Lucchetti-Miganeh C, Barloy-Hubler F, Brown M, Lupien M, Metivier R, et al. Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. Genome research. 2011;21(4):555–565.

67. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics. 2007;39(3):311–318.

68. Tan SK, Lin ZH, Chang CW, Varang V, Chng KR, Pan YF, Yong EL, Sung WK, Cheung E. AP-2γ regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription: Long-range gene regulation by AP-2γ. The EMBO journal. 2011 [accessed 2020 Aug 2];30(13):2569–2581.

69. Kong SL, Li G, Loh SL, Sung W-K, Liu ET. Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. Molecular systems biology. 2011;7:526.

70. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, et al. Chromosome-Wide Mapping of Estrogen Receptor Binding Reveals Long-Range Regulation Requiring the Forkhead Protein FoxA1. Cell. 2005;122(1):33–43.

71. Lentjes MH, Niessen HEC, Akiyama Y, de Bruïne AP, Melotte V, van Engeland M. The emerging role of GATA transcription factors in development and disease. Expert reviews in molecular medicine. 2016;18:e3.

72. Asnagli H, Afkarian M, Murphy KM. Cutting edge: Identification of an alternative GATA-3 promoter directing tissue-specific gene expression in mouse and human. Journal of immunology . 2002;168(9):4268–4271.

73. Chou J, Provot S, Werb Z. GATA3 in Development and Cancer Differentiation: Cells GATA Have It! Journal of cellular physiology. 2010;222(1):42–49.

74. Kouros-Mehr H, Slorach EM, Sternlicht MD, Werb Z. GATA-3 Maintains the Differentiation of the Luminal Cell Fate in the Mammary Gland. Cell. 2006;127(5):1041–1055.

75. Asselin-Labat M-L, Sutherland KD, Barker H, Thomas R, Shackleton M, Forrest NC, Hartley L, Robb L, Grosveld FG, van der Wees J, et al. Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. Nature cell biology. 2007;9(2):201–209.

76. Zhang F, Wen Y, Guo X. CRISPR/Cas9 for genome editing: progress, implications and challenges. Human molecular genetics. 2014;23(R1):R40–R46.

77. Barrangou R. The roles of CRISPR–Cas systems in adaptive immunity and beyond. Current opinion in immunology. 2015;32:36–41.

78. Deltcheva E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. 2011:8.

79. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences. 2012;109(39):E2579–E2586.

80. Karvelis T, Gasiunas G, Siksnys V. Methods for decoding Cas9 protospacer adjacent motif (PAM) sequences: A brief overview. Methods . 2017;121-122:3–8.

81. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. 2012;337:7.

82. Rouet P, Smih F, Jasin M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Molecular and cellular biology. 1994;14(12):8096–8106.

83. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-Guided Human Genome Engineering via Cas9. Science. 2013;339(6121):823–826.

84. Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, Sanjana NE, Zhang F. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nature protocols. 2017;12(4):828–863.

85. Zhang J-P, Li X-L, Neises A, Chen W, Hu L-P, Ji G-Z, Yu J-Y, Xu J, Yuan W-P, Cheng T, et al. Different Effects of sgRNA Length on CRISPR-mediated Gene Knockout Efficiency. Scientific reports. 2016;6(1):28566.

86. Dahlman JE, Abudayyeh OO, Joung J, Gootenberg JS, Zhang F, Konermann S. Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. Nature biotechnology. 2015;33(11):1159–1161.

87. 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 . 2013;29(1):15–21.

88. Trapnell C. Defining cell types and states with single-cell genomics. Genome research. 2015;25(10):1491–1498.

89. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646–674.

90. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 05/2015 [accessed 2019 May 6];161(5):1202–1214.

91. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–1201.

92. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nature communications. 2017;8:14049.

93. Saikia M, Burnham P, Keshavjee SH, Wang MFZ, Heyang M, Moral-Lopez P, Hinchman MM, Danko CG, Parker JSL, De Vlaminck I. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. Nature methods. 2019;16(1):59–62.

94. Booeshaghi AS, da Veiga Beltrame E, Bannon D, Gehring J, Pachter L. Design principles for open source bioinstrumentation: the poseidon syringe pump system as an example. bioRxiv. 2019 Jan 17. http://biorxiv.org/lookup/doi/10.1101/521096. doi:10.1101/521096

95. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, Huang Y, Wang J. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. Genomics; 2018. http://biorxiv.org/lookup/doi/10.1101/313130

96. Booeshaghi AS, Beltrame E da V, Bannon D, Gehring J, Pachter L. Principles of open source bioinstrumentation applied to the poseidon syringe pump system. Scientific reports. 2019;9(1):12385.

97. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nature genetics. 2017 [accessed 2021 Jul 23];49(12):1779–1784.

98. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology. 2019;37(1):38–44.

99. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, Meer EJ, Terry JM, Riordan DP, Srinivas N, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nature biotechnology. 2020;38(8):954–961.

100. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–3587.e29.

101. Wessels H-H, Méndez-Mancilla A, Guo X, Legut M, Daniloski Z, Sanjana NE. Massively parallel Cas13 screens reveal principles for guide RNA design. Nature Biotechnology. 2020;38(6):722–727. doi:10.1038/s41587-020-0456-9

102. Weinreb C, Rodriguez-Fraticelli A, Camargo FD, Klein AM. Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science. 2020;367(6479):eaaw3381. doi:10.1126/science.aaw3381

103. Svensson, Valentine, Eduardo da Veiga Beltrame, and Lior Pachter. “A Curated Database Reveals Trends in Single-Cell Transcriptomics.” Database 2020: baaa073. https://doi.org/10.1093/database/baaa073.

104. Wessels, Hans-Hermann, Alejandro Méndez-Mancilla, Efthymia Papalexi, William M Mauck, Lu Lu, John A. Morris, Eleni Mimitou, Peter Smibert, Neville E. Sanjana, and Rahul

Satija. “Efficient Combinatorial Targeting of RNA Transcripts in Single Cells with Cas13 RNA Perturb-Seq.” Preprint. Genomics, February 2, 2022. https://doi.org/10.1101/2022.02.02.478894.

105. Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Salame TM, Tanay A, van Oudenaarden A, Amit I. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell. 2016;167(7):1883-1896.e15. doi:10.1016/j.cell.2016.11.039

106. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. Pooled CRISPR screening with single-cell transcriptome readout. Nature Methods. 2017;14(3):297–301. doi:10.1038/nmeth.4177

107. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419. doi:10.1126/science.1260419

108. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, et al. Defining a Cancer Dependency Map. Cell. 2017;170(3):564-576.e16. doi:10.1016/j.cell.2017.06.010

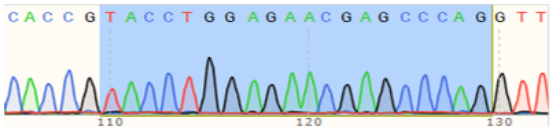
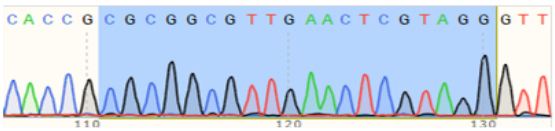
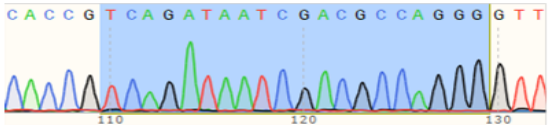
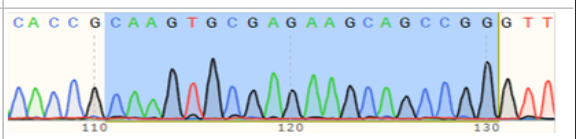
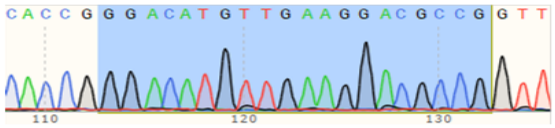
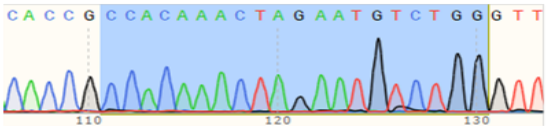
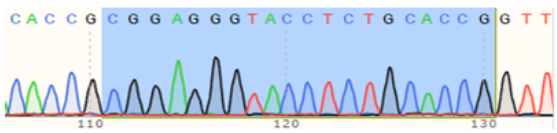
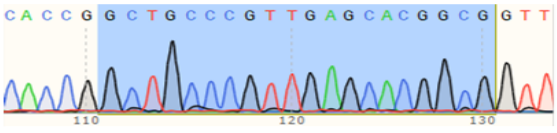
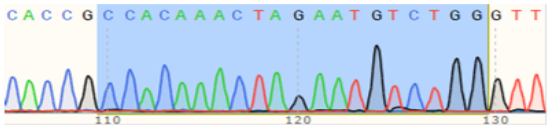
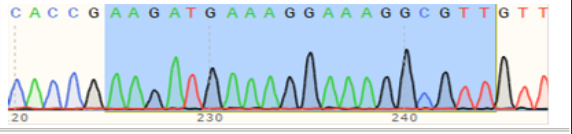
109. Yu NY, Iftimi A, Yau C, Tobin NP, van ’t Veer L, Hoadley KA, Benz CC, Nordenskjöld B, Fornander T, Stål O, et al. Assessment of Long-term Distant Recurrence-Free Survival Associated With Tamoxifen Therapy in Postmenopausal Patients With Luminal A or Luminal B Breast Cancer. JAMA Oncology. 2019;5(9):1304–1309. doi:10.1001/jamaoncol.2019.1856

110. Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Mathelier A, editor. Bioinformatics. 2021;37(7):963–967. doi:10.1093/bioinformatics/btaa751

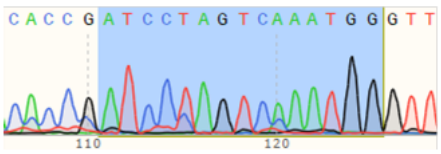
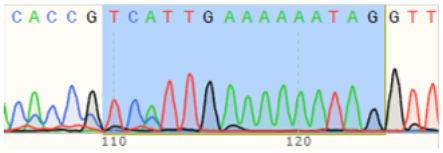
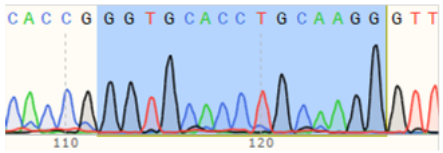
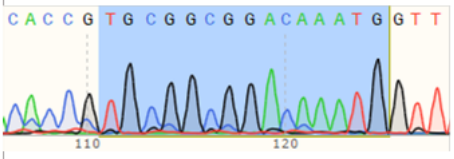
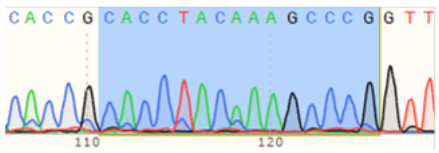
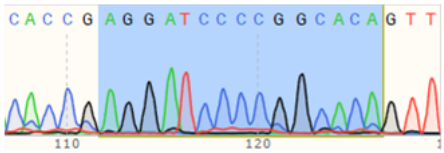
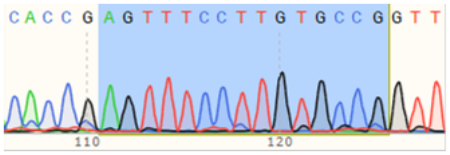
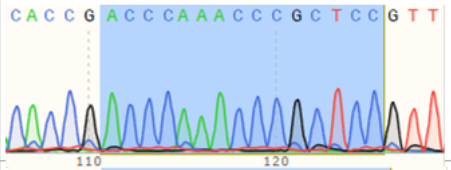
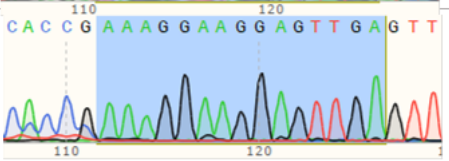
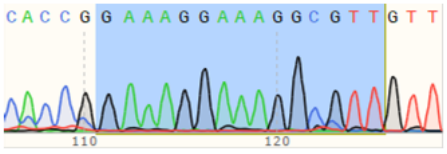
Supplemental Figures.

Supplemental figure 1. Validation of sgRNA ligations. Chromatograms confirming ligation of **A.** KO sgRNA sequences in lentiGuide-Puro and **B.** OE sgRNA sequences in lenti sgRNA(MS2)_puro. Almost all sgRNAs were successfully ligated. Unsuccessful ligations are omitted.

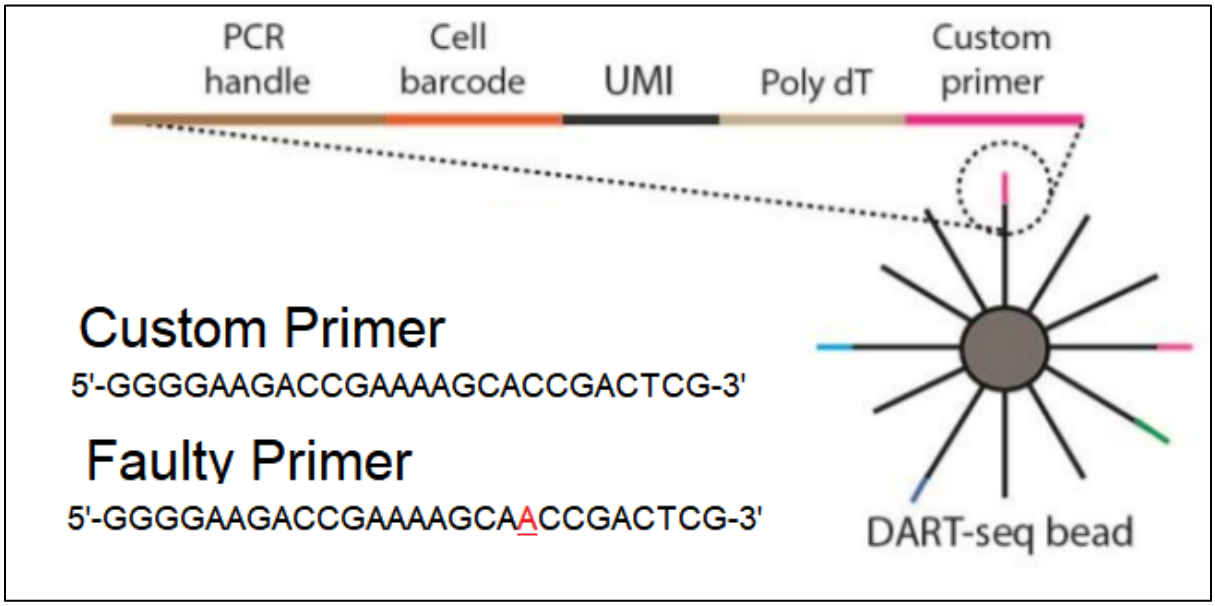
A

Target	sgRNA #	Sequence (5'-3')	Sequencing Chromatogram
ESR1	1	TACCTGGAGAACGAGCCCAG	
	2	CGCGGCGTTGAACTCGTAGG	
	3	TCAGATAATCGACGCCAGGG	
FOXA1	1	CAAGTGCGAGAAGCAGCCGG	
	2	GGACATGTTGAAGGACGCCG	
	3	CCACAAACTAGAATGTCTGG	
GATA3	1	CGGAGGGTACCTCTGCACCG	
	2	GCTGCCCGTTGAGCACGGCG	
	3	CCACAAACTAGAATGTCTGG	
Non-Targetting	2	AAGATGAAAGGAAAGGCGTT	

B

Target	sgRNA #	Sequence (5'-3')	Sequencing Chromatogram
ESR1	2	ATCCTAGTCAAATGG	
	3	TCATTGAAAAAATAG	
FOXA1	1	GGTGCACCTGCAAGG	
	2	TGCGGCGGACAAATG	
	3	CACCTACAAAGCCCG	
GATA3	1	AGGATCCCCGGCACA	
	2	AGTTTCCTTGTGCCG	
	3	ACCCAAACCCGCTCC	
Non-Targetting	1	AAAGGAAGGAGTTGA	
	2	GAAAGGAAAGGCGTT	

Supplemental figure 2. Visualization of faulty custom primer.



Supplemental figure 3. Example of primer design for sgRNA protospacer ligation.

```
PrimerFW   5' - CACCG NNNNNNNNNNNNNNNNNNN - 3'
PrimerRV   3' -     C NNNNNNNNNNNNNNNNNNN CAAA - 5'
```

Supplemental Materials and Methods

1 Cell lines and cell culturing.

MCF7-ATCC and φnx cell lines were provided from the Mader lab. MCF7 cells were cultured in RPMI 1640 supplemented with 10% FBS and 1% penicillin/streptomycin and passaged every three to four days at a 1:3 ratio. Φnx cells were cultured in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin and passaged every two days at a 1:3 ratio. I prepared the reagents fresh before passaging. I then generated and validated clonal cell lines of MCF7 transduced with Cas9 and Cas9 + MPH via lentiviral transduction.

To determine the concentration of the utilized antibiotics, I performed an antibiotic kill curve. MCF7 cells were transduced with lentiCas9-Blast (Addgene #52962) to generate Cas9 ready cells. Selection was performed using 7 µg/mL blasticidin and maintained at 4 µg/mL. Selected cells were FACS sorted into 96-well plates and cultured. Cells were expanded and confluent cells were tested for Cas9 expression via western blot using Anti-Cas9 (mouse polyclonal, 14697, Cell Signalling Technology, 1:1000 Dilution). Successfully modified cells were clonally amplified and transduced with lentiMPHv2 (Addgene #89308). Cells were selected with the addition of 500 µg/mL hygromycin and maintained at 250 µg/µL. Resistant cells were FACS sorted into 96-well plates and cultured. Cells were expanded and confluent cells were tested for Cas9 and MPH expression via western blot using Anti-Cas9 and Anti-NF-κB p65 (rabbit polyclonal, 8242, Cell Signalling Technology, 1:1000 Dilution). Validated clones were amplified to generate a COIN-seq ready cell line we have named COIN-MCF7. Generated cell lines were stored in liquid nitrogen with 5% DMS and cultured with 0.2 units/ml bovine insulin and 10% FBS to the final concentration. For preparation for droplet-sequencing, we rinsed the cell layer with 0.25% (w/v) Trypsin and 0.53 mM EDTA solution, added 2 mL of Trypsin-EDTA solution before visual examination for dispersion under an inverted microscope.

Cells were suspended following the protocol from Macosko et al. to a final concentration of 120 cells/μL in preparation for droplet-sequencing.

2 CRISPR/Cas9 Guide RNAs.

Sanny Khurdia and I designed the protospacer sequences for KO and OE experiments using the Genetic Perturbation Platform sgRNA designer from the Broad Institute (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design). **Table 1** catalogs each chosen protospacer with its intended target and sequence in descending order of *in silico* determined intervention efficiency. The sequences were integrated into a standardized primer design (**Figure S3**) to contain *BsmBI* overhangs used in golden gate ligation and were ordered from IDT.

We individually cloned both KO and OE oligonucleotides sets following the identical protocol. Firstly, the forward and reverse oligonucleotides are annealed. For each target, 1 μL the forward and reverse oligonucleotides are diluted to 100 μM were combined with 1 μL of 10X T4 ligase buffer (NEB, B0202S), 0.5 μL T4 PNK (NEB, M0201S), and 6.5 μl ddH2O. The mixture is cycled at 37°C for 30 min, incubated at 95°C for 5 min, before being ramped down to 25°C at a rate of 5°C/min, resulting in the annealed pair.

The annealed pair is diluted in a 1:10 ratio of ddH2O. Next, a Golden Gate reaction is performed using 1 μL of the diluted pair is combined with 12.5 uL 2X rapid ligase buffer (Promega, C6711), 0.125 μL of 20mg/ml BSA (Sigma, B8667), 1 μL *BsmBI* (Fermentas, FD0454), 0.125 μL T4 ligase (ThermoFisher, 15224017), 9.25 μl ddH2O, and 1 μL of the backbone vector, lentiGuide-Puro (Addgene #52963) for KO and lenti sgRNA(MS2)_puro backbone (Addgene #73795) for OE at 25 ng/μL. The mixture is cycled at 37°C for 5 min and

20°C for 5 min for a total of 15 cycles, resulting in the insertion of the guide RNA protospacer sequences into the backbone vector.

For amplification of the product, 10 µL of the plasmid was combined with 50 µL TOP10 chemically competent *E. coli*. and incubated for 2 min on ice. The sample was then heat shocked at 42°C for 1 minute and incubated on ice for 2 min. Samples were then spread onto LB + Ampicillin agar plates and incubated overnight at 37°C. Single colonies were selected and expanded. The plasmids were extracted and purified using QIAGEN Plasmid Midi Kit (Quiagen, 12143) following the manufacturer's instructions. Quantification of each plasmid was performed using the Tecan Infinite 200 PRO (Tecan, 30052730).

Inserted sequences were validated using Sanger Sequencing. Samples were diluted to a concentration of 200-300 ng/µL at a minimum of 5 µL. The U6 primer (5'-CAGCACAAAAGGAAACTCACC-3') was used for Sanger Sequencing at 5 µM. Samples were submitted at the IRIC Genomics Platform.

3 Construction of lentiviral vectors and transduction.

The Cas9, MPH, and individual sgRNAs expression plasmids were assembled by myself into lentiviral constructs using the identical following protocol. On day 1, the φnx cell line was seeded at $5x10^6$ cells. In the morning of day 2, two mixes are prepared: Mix 1 is prepared by mixing 1 µg/µL of both the psPAX2 (Addgene #12260) and VSVg (Addgene #14888) lentiviral packaging vectors with 1 µg/µL expression plasmid. Mix 2 is prepared by mixing 54 µL of 1 µg/µL PEI (Polysciences, 23966) and 446 µL HBSS (Gibco, 24020117). Mixes 1 and 2 were combined and incubated at room temperature for 15 min. The solution was then added dropwise to φnx cells, mixed, and incubated for 16 hours at 37°C. On day 3, the media was changed and treated with 200 µL of 500 mM Sodium Butyrate. At least 6 hours after

treatment, cells were rinsed with PBS and media was changed to RPMI. After a 16 hour incubation period, the viral supernatant was collected and filtered through a 45 µm PVDF filter, flash frozen using dry ice, and stored at -80°C. Cells transduced with sgRNA expression plasmids were selected using 1 µg/mL puromycin. Lentiviral titres were determined by infecting cells with 6 different volumes of lentivirus ranging from 0 to 1 mL in 6 well culture plates and observing the number of surviving cells after complete selection.

4 Validation of constructs and lentiviral infections.

I validated the efficiency of KO target sequences via western blot. Protein lysates were prepared with RIPA lysis buffer containing a protease inhibitor cocktail of pepstatin, aprotinin, leupeptin, and PMSF. Samples were standardised for protein via Lowry Assay (Bio-Rad, 5000111) and boiled at 95°C for 5 min. After denaturation, samples were separated by 7-8% via SDS-PAGE and electrotransferred onto a 0.2 µm polyvinylidene difluoride (PVDF) membrane (ThermoFIsher, 88520). Blots were blocked with 5% skim milk in PBS-T and probed with different primary antibodies anti-ERα (rabbit polyclonal, 03-820, Millipore, 1:2000 Dilution), anti-FOXA1 (rabbit polyclonal, ab23738, Abcam, 1:2000 Dilution), anti-GATA3 (rabbit polyclonal, 558686, BD Pharmingen, 1:1000 Dilution), anti-ACTB (mouse polyclonal, A5316, Sigma-Aldrich, 1:1000 Dilution), and anti-LMNB1 (rabbit polyclonal, ab16048, Abcam, 1:10000 Dilution) in 5% skim milk in PBS-T overnight at 4°C. Blots were then incubated with secondary antibody HRP-conjugated mouse/rabbit in 5% skim milk in PBS-T for 1h at room temperature. Detection was done using Clarity Western ECL Substrate (Bio-Rad, 1705060).

5 Microparticle modification for sgRNA capture

Oligo-dt beads (ChemGenes, #MACOSKO-2011-10(V+), custom) were modified by Sanny Khurdia such that a short sequence was ligated onto the 5’ ends of the oligo-dt extending from the beads. First, the toehold probes were generated by combining 20 µL of

both 500 mM Splint Oligo and 500 mM Custom Oligo (**Table 3**), with 5 µl Tris-EDTA Buffer, and 5µl 0.25M NaCl Solution. This reaction mixture was incubated up to 95°C and then cooled to 14°C at a constant rate of -0.1°C/sec. Then, 200 µL TE Buffer was combined to give a stock solution of 100 mM stock of so-called “toehold probes”. 6µL of this stock solution was combined with 12,000 beads, 432 µL of ultrapure water, 90 µL TE Buffer, 240 µL of PEG-4000 (50% w/v), 240 µL of T4 DNA ligase buffer and 12 µL of T4 DNA Ligase. This mixture was incubated for 1 hour at 37°C while shaking at 1800 rpm. The reaction mix was then heat shocked at 65°C for 3 minutes to inactivate the enzymes, before being placed on ice for at least 1 minute to quench the reaction. The bead mixture was washed once with TE-SDS (0.05%) and twice with TE-TW (0.1%) and stored at 4°C until ready for use.

6 Drop-seq runs, reverse transcription, and PCR amplification

Drop-seq runs were performed by myself, Sanny Khurdia, and Abdelrahman Ahmed. The reagents used were prepared fresh the day of the Drop-seq runs. When performing a run, we designated three stations: a cell preparation station, a Drop-seq station, and a reverse transcription station. During a batch of runs, we sought to minimize sources of variation by having the same individual perform their task at the same station following identical procedure throughout the run.

The run begins at the cell preparation station. The samples we prepared for Drop-seq were washed with PBS rinsed and detached using 2 mL of Trypsin-EDTA solution incubated at 37ºC for 3 minutes. After visual inspection under an inverted light microscope to confirm loss of cell adhesion, Cells were collected and washed with PBS-BSA and PBS. The cells were then filtered using a 40µM filter and counted using a hemocytometer. The cells were suspended at a final concentration of 120k cells/µL in PBS-BSA and put onto ice in preparation for use.

The cell suspension is then brought to the Drop-seq station. Lysis Buffer was prepared by combining 5 mL DEPC treated H2O, 3 mL 20% Ficoll PM-400, 2 mL of 1 M Tris pH 8.0, 400 µL of 0.5 M EDTA, 100 µL of 20% Sarkosyl and stored at room temperature. Once ready, 50 µL 1 M molecular biology grade DTT is added per 1 ml of lysis buffer being used and is fed into the microfluidic device alongside the Droplet oil (BioRad, 1863005), and cell solution. The outflow was collected in a 50 mL falcon tube and was combined with 30 mL 6X SSC and 1 mL Perfluoro-octanol. The tube is then sent to the reverse transcription station. The falcon tube was sealed and shaken vigorously 5 times, before being centrifuged at 1000 x g for 1 minute. After careful visual inspection to confirm a floating layer of beads at the oil-water interface, the top layer was carefully removed and discarded without disturbing the interface. Then, 30 mL of 6X SSC was combined to interrupt the floating layer of beads. After waiting for the oil to settle to the bottom, the top layer was removed and transferred into a new falcon tube. This tube is then spun at 1000 x g for 1 minute to pellet and collect the beads. The remaining supernatant is carefully removed, and then the remaining 1mL is pipette-mixed to resuspend the beads. The bead suspension is transferred to an Eppendorf tube and centrifuged at max for 30 sec. The beads are washed twice with 6X SSC, and then washed with 5X RT Buffer. The beads are then combined with Reverse Transcriptase (Invitrogen, EP0753), TSO (**Table 3**), and are incubated at room temperature for 30 mins while shaking, before being incubated at 42ºC for 90 min while rotating at 1600 RPM. After reverse transcription, wash the beads once with TE-SDS and twice with TE-TW. Next, the excess primers are cleaved off by Exonuclease Treatment by combining the beads with 10 µL *ExoI* and 190 µL 1X Exo Buffer and incubating at 37ºC for 45 min. Next the beads were washed once with TE-SDS, twice with TE-TW, and then once again with ultrapure water. Using a hemocytometer, batches of 4,000 beads were apportioned into PCR tubes. Beads were combined with 24.6 µL KAPA Hifi HotStart ReadyMix (Roche, 07958935001), 0.4 µL of 100 mM SMART oligo (**Table 3**) and 25 µL ultrapure water and cycled at 95ºC for 3 min, four cycles of 98ºC for 20 sec, 65ºC for 45 sec, and 72ºC for 3

min, nine cycles of 98ºC for 20 sec, 67ºC for 20 sec, and 72ºC for 3 min, and finally 72ºC for 5 min, finishing with a hold at 4ºC.

7 Library Separation and Tagmentation

Following PCR amplification, the two cDNA libraries (sgRNA library: ~140 bp, mRNA library: 800-1200 bp) are purified using a first round of 0.6 SPRI selection, where the supernatant (sgRNA library) is kept and then undergoes a subsequent round of 2x SPRI selection. The first product is the mRNA derived cDNA product and the second is the guide derived cDNA product. All libraries were then quantified using an Agilent Tapestation 4150 system. Following Quantification, mRNA libraries were tagemented using the NexteraXT tagmentation kit (Illumina Inc., FC-131-1024) and index sequences (**Table 7a**) whilst sgRNA libraries were prepared using PCR to add Next_nst_x indexes (**Table 7b**).

8 Next generation sequencing and bioinformatics and statistics for the single cell profiles

Samples were processed and sequenced using the NextSeq 500 following standard protocol[90] (average 100M reads/sample). In general, all computations were performed using Python version 3.67 or R version 3.6.1. Raw FastQ files were processed with the DropEST pipeline developed for estimating molecular count matrices from droplet-based single cell RNA-seq developed by the Kharchenko Lab at Harvard University (github.com/kharchenkolab/dropEst). The General steps are detailed below:

1. DropTag.sh: Script for demultiplexing raw FastQ files. Cell barcodes and UMI's are extracted. The resulting FastQ files are aligned against a reference.
2. Alignment: Reads were aligned against the GRCh38 human reference genome using STAR aligner (github.com/alexdobin/STAR).

3. DropEst: Builds count matrices from the output .bam files generated by STAR aligner. The script also generates statistics that are used for quality control.
4. DropReport: Generates HTML report for each sample with information regarding library quality and other important statistics (e.g. top genes, number of cells, reads/UMI).

Single cell analysis was performed following Seurat V4 standard workflow developed by the Satija Lab at New York University (satijalab.org/seurat/articles/multimodal_vignette.html)[100].