Single-cell approaches for deciphering complex tissue heterogeneity

Kristóf András Törkenczy

A DISSERTATION

Presented to the Department of Molecular and Medical Genetics
Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

June 2020

School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

_____

This is to certify that the PhD dissertation of
Kristóf András Törkenczy
has been approved by

_____
Andrew Adey, PhD, Mentor/Advisor

_____
Paul Spellman, PhD, Committee Chair

_____
Rosalie Sears, PhD, Member

_____
Emek Demir, PhD, Member

_____
Laura Heiser, PhD, Member

TABLE OF CONTENTS

List of Figures/Tables

Abbreviations

SNP	Single nucleotide polymorphisms
SNV	Single nucleotide variants
SV	Structural variants
CNV	Copy number variation
CNN-LOH	Copy number neutral loss of heterozygosity
CNL-LOH	Copy number loss of heterozygosity
DSB	Double strand breaks
NHEJ	Nonhomologous end-joining
GC	Gene conversion
SSA	Single-strand annealing
IHC	Immunohistochemistry
PR	Progesterone
ER	Estrogen
HER2	Human epidermal growth factor receptor 2
TNBC	Triple negative breast cancer
WGA	Whole genome amplification
MEK	Mitogen-activated protein kinase
DOP-PCR	Degenerate oligonucleotide-primed PCR
MDA	Multiple displacement amplification
MALBAC	Multiple annealing and looping-based amplification
LIANTI	Linear amplification via transposon insertion
MAPD	Mean absolute deviation of pairwise differences
CBS	Circular binary segmentation
HMM	Hidden markov model

ATAC-seq	Assay for transposase-accessible chromatin
EMT	Epithelial-to-mesenchymal transition
MASC	Mammary stem cell
DRP	Drug resistant persistor
FISH	Fluorescence in situ hybridizationSscRNA-seq Single-cell RNA sequencing
scATAC-seq	Single-cell ATAC sequencing
sci-ATAC-seq	Single-cell combinatorial indexed ATAC sequencing
sciDNA-seq	Single-cell combinatorial indexed DNA sequencing
PCA	Principal component analysis
UMAP	Uniform approximation and projection method
t-SNE	t-distributed stochastic neighbor embedding
LSI	Latent semantic indexing
TF-IDF	Term frequency-inverse document frequency transformation
SVD	Singular value decomposition
LAND	Lithium assisted nucleosome depletion
xSDS	Crosslinking with SDS treatment
FANS	Fluorescence activated nuclei sorting
PDAC	Pancreatic ductal adenocarcinoma
AST	Astrocytes
INT	Interneurons
OLI	Oligodendrocytes
MRG	Microglia
OPCs	Oligodendrocyte progenitor cells
CCAN	Cis-co-accessibility network
TAD	Topologically associated domain
BCCLs	Breast cancer cell lines

CCA	Canonical correlation analysis
TF	Transcription Factor
CLS	Cell line specific site
NOME-seq	Nucleosome occupancy and methylation sequencing
UMI	Unique molecular identifier
DE	Differentially expressed
DA	Differentially accessible
CLG	Cell line-specific genes
MNN	Mutual nearest neighbors
TRT	Treatment response trajectory
RE	Regulatory elements
Pr	Promoters

Acknowledgments

I would first like to thank Andrew for his patient mentorship. You have given me an opportunity to move into a novel field, grow, and find my own scientific interests. I have become a better scientist in the process of graduate school and I have you to thank for that. Thank you for being understanding about the traveling I needed to visit family and friends back home. The flexibility in work schedule made me feel safe and I put extra effort in whenever I could. Lastly, thank you for introducing me to Magic: The Gathering, just because the nerd in me did not know what it needed.

I would also like to thank all past and present members of my lab who have supported me throughout these five and a half years. You have provided scientific and emotional support whenever I asked for it, made me feel welcome and valued. Ryan, thank you for being an amazing friend throughout all these years and for keeping me on all ten of my scientific toes. I would also like to thank members of the O'Roak lab, your input in our shared lab meetings have been very valuable for my scientific progress. On the same note I would also like to thank Stephen Moore for organizing and presiding over a truly amazing journal club. I have learned so much about varying fields of genetics during discussions with some amazing scientists. I would also like to thank Jackie Wirz for all the help she provided to me and my fellow graduate students.

The MMG department has provided me with a home. I would like to thank all the administrative and scientific staff for making that happen. I would like to especially thank the program directors of MMG: Amanda and Mushui. I would like to thank all of my DAC members: Paul, Rosie, Emek, and Laura for their continual support and feedback. Paul has played a pivotal role in me coming to OHSU and I would like to thank him for that.

As I am writing this, we are going through some truly turbulent times. I never would have thought that anything could sever the connection I have with family and friends. This social isolation really brings forth the importance of the relationships that I have had the fortune of having.

I would like to thank all of the friends I have made inside and outside of grad school. Thank you for participating in my various experiments of group socializing (cooking competitions) and for making me do things I never thought I would be able to do (e. g.: Ragnar, Hood to Coast). I would not be here if it was not for my "foster family" in Portland: András, Anna, Áron, Nóra.

I would also like to thank the members of my other life, the friends who make me feel like I never left home (Miki, Peti, Ambró, Dani, Marci, Zoli, Dan, Andris). Thank you to my parents: Mamuska and Papuska. Nagyon szeretlek benneteket. Thank you to my grandparents. I wish I could all see you all in person. You have given so much love and support that I cannot even describe it. Finally, I would like to thank Eileen. Our journey here could not be aptly summarized in the space I have available in this dissertation, but I can summarize our future in one word: excitement.

Abstract

During healthy development and aging, somatic mutations aggregate in tissues, lending different mutational profiles to cells within these tissues. Similarly, developmental processes create the functional diversity necessary for normal tissue function. Cancer arises from the aberrant functioning of these processes. Mutational profiles and epigenetic regulation support uncontrolled growth of neoplasms with the potential ability to invade nearby tissues. The resulting intra-tumor heterogeneity contributes to evasion of drug pressures via Darwinian selection, and cell-state plasticity allows for dynamic shifts in regulation into drug-resistant persistor states. While bulk genomic assays profile averages of sampled cell populations, single-cell approaches allow for a picture of the heterogeneity of healthy and diseased complex tissues. In this dissertation, I assess the state of the single-cell field with a focus on assays characterizing whole genome copy number variation and chromatin accessibility. I show three examples of profiling heterogeneity by (i) assessing the genomes of thousands of cells in healthy and diseased tissues (ii) mapping the chromatin landscape of the murine hippocampus, and (iii) characterizing the development of Trametinib resistance through cell state plasticity across basal-like triple negative breast cancer cell lines.

Chapter 1: Introduction

1.1 Heterogeneity of complex tissues

Estimates for the total number of cells in the human body range from 10^{12} to 10^{16} cells (Bianconi et al., 2013) with the current number for a 70 kg male reference approximated at 3×10^{13} cells (Sender et al., 2016). These cells serve a variety of roles as comprising parts of larger organized structures forming functionally complex tissues and organs necessary for life. Defining and characterizing the cell types present within complex tissues is a necessary component to building a comprehensive reference map that can help us understand fundamental biological processes of healthy tissues and their diseased counterparts (Rozenblatt-Rosen et al., 2017). The functional heterogeneity presenting across cells in a complex tissue can be the result of cell-to-cell genetic variation present in cells and/or the heterogeneity of regulation on top of this across all components of the central dogma (DNA, RNA, Protein). It is important to note, that observed phenotypic differences between cells can be independent of genetic variation within a tissue and can be restricted to differences in regulation only. In addition, heterogeneity of different regulatory levels may not match in the same cell populations (Goldman et al., 2019; Hinohara & Polyak, 2019a). This necessitates the accurate genomic and regulatory characterization of complex tissues. The evolution of massively parallel sequencing and the emergence of multi-omics approaches has led to large consortium projects of healthy and diseased bulk tissues across large numbers of individuals (Auton et al., 2015; P. J. Campbell et al., 2020; Dunham et al., 2012; Hudson et al., 2010; Roadmap Epigenomics Consortium et al., 2015; Weinstein et al., 2013). However, due to technical limitations, this has not translated fully to within-tissue heterogeneity. Bulk approaches require tissue samples consisting of input material on the order of micrograms; therefore, an average profile is reported across all assayed cells (N. E. Navin, 2015). This can result in potentially missing important sources of heterogeneity as both healthy and diseased tissues have been shown to harbor genomic alterations and regulatory differences present at low frequencies (<5%) in

sampled cell populations (Carter et al., 2012; Cibulskis et al., 2013; Mo et al., 2015; Rozenblatt-Rosen et al., 2020). The emergence of single-cell methods finally made it possible to accurately capture the full breadth of represented diversity. In this dissertation I will assess how current single-cell technologies and computational analysis methods have helped unravel cellular genetic, epigenetic, and transcriptomic heterogeneity within two of the most studied complex tissues: the brain and the breast. The second focus of this dissertation is how heterogeneity, a basic property of healthy tissues, provides the foundation to evasion of treatment response in cancers on multiple regulatory levels.

1.1.2 Genomic heterogeneity in healthy tissues

Mutations arise as multicellular organisms develop from a single embryonic cell (Figure 1.1A.) The size of mutations can range from single nucleotide variants (SNVs), short insertions and deletions in the genome (Indels), to larger structural variants (SVs) including the gain or loss of entire chromosomes (aneuploidy) and translocations (equal or unequal exchange of genetic material between chromosomes, 1.1B). Different molecular mechanisms give rise to different categories of mutations. Permanent changes to the DNA can occur via errors during DNA replication, meiosis, and mitosis, or through damage via exposure to radiation (*e.g.*: formation of pyrimidine dimers upon photochemical reactions) or carcinogens (Bertram, 2000; Chatterjee & Walker, 2017). Similarly, mobile genetic elements can introduce deletions or insertions to the DNA sequence (Chénais et al., 2012). Repair processes within a cell can correct for introduced changes, but can also be quite error prone (J. Chen et al., 2014; Rodgers & Mcvey, 2016; S. Sharma et al., 2015). For example, copy number variants (gains or losses of large genomic regions, CNVs), which are the focus of chapter 2 of this dissertation, have a range of identified molecular mechanism associated with their formation, including non-allelic homologous recombination (NAHR), fork stalling and template switching (FOSTES), non-homologous end-joining (NHEJ), and mobile element insertion (MEI). The frequency of these events in genomic regions is often driven by local

genomic architecture, such as clusters of low copy repeats, repeated sequences and repetitive elements (Bickhart & Liu, 2014; Carvalho & Lupski, 2016).

When a mutation occurs in a germ cell lineage (*i.e.* sperm or egg), it can be passed onto the next generation, where the mutation will be present in all somatic cells (Milholland et al., 2017). Similarly, *de novo* mutations can arise, which are present in all cells of an offspring but cannot be detected in the parents (Freed et al., 2014). These mutations may occur early in development (*i.e.*: first few cell divisions of the zygote); however, studies applying more sensitive genetic assays have shown cases of low-level mosaicism in parents of *de novo* cases (I. M. Campbell et al., 2014; Van Der Maarel et al., 2000). When we look at SNVs, we find the rate of *de novo* germline mutations range from approximately 1.18×10^{-8} to 2.5×10^{-8} mutations per base pair per generation in humans and 4.6×10^{-9} to 6.5×10^{-9} mutations per base pair per generation in mice (Conrad et al., 2011; Milholland et al., 2017; Uchimura et al., 2015). These germline mutations contribute to the overall variation present in a species. The Thousand Genomes Project found a typical human genome differs in 4.1 to 5 million sites from the reference (depending on the studied population), with the majority (99.9%) consisting of single nucleotide polymorphisms (SNPs) and short indels. Structural variants were less frequent (2,100 to 2,500 SVs in a typical genome, ~160 CNVs) but covered a larger portion of the genome than SNPs and short indels (~20 million bases of sequence per typical genome). When taking SNPs and indels into consideration, analysis of 69 samples from each of the studied 6 populations found expression quantitative trait loci (eQTL) at 3,285 genes at 5% false discovery rate. For each of the studied populations the top 7.5–19.5% of eQTL variants overlapped transcription factor binding sites, indicating their high importance in the regulation of gene expression levels (see section 1.2.1). For larger structural variants the database of genomic variants lists approximately 7 million entries in the human genome with 984 thousand copy number variants alone overlapping with 85% of exons in the database (MacDonald et al., 2014). This highlights the importance of accurately cataloging CNVs to understand the potential effects on expression and their association with disease.

Post-zygotic, somatic mutations occur in cells after conception and can only be passed on within their cell lineages (Fig 1.1A). This results in somatic mutations often being restricted to specific tissues (C. Li & Williams, 2013). Somatic SNVs (Fig 1.1B) are approximated to have a higher mutational rate than their germline counterparts (Lynch, 2010),with a recent study showing 2.8×10^{-7} and 4.4×10^{-7} mutations per base pair per generation for human and mouse, respectively (Milholland et al., 2017). This is an average value, as somatic mutations accumulate at different rates depending on tissue type (García-Nieto et al., 2019; Lee-Six et al., 2019; Lodato et al., 2018; Iñigo Martincorena et al., 2015). This is likely due to the functional diversity of tissues, which allow for different somatic mutational loads (C. Li & Williams, 2013). A recent comprehensive study cataloging 280,000 mutations from 36 healthy tissues showed disparate mutational loads across tissues, which also correlated with age, sex, and ethnicity. Overall mutational load increased with age across all tissues but was relatively lower in brain tissues. Breast tissue specifically showed a strong sex bias, with females harboring a higher mutational load (García-Nieto et al., 2019). This study also explored selection on somatic variants via the ratio of non-synonymous (*i.e.* amino acid altering, dN) to synonymous (*i.e.* amino acid conserving, dS) mutations across genes. Values of dN/dS close to 1 represent little or no detectable selection, dN/dS > 1 suggests positive selection, and dN/dS < 1 implies purifying selection. Interestingly, missense and nonsense mutations across all tissues showed negative selection (dN/dS < 1) except for mutations previously observed in cancer samples, which exhibited positive selection across many healthy tissues (García-Nieto et al., 2019). This was especially prominent in non-cancerous skin, where mutations at cancer associated gene *NOTCH1* showed strong positive selection load (García-Nieto et al., 2019; Iñigo Martincorena et al., 2015). Similarly, in normal breast tissues there was an increased number of mutations in genes in pathways regulating proliferation, cell to cell adhesion, and cell survival such as *TGFBR2*, *CTNNB1* and *AKT1* (García-Nieto et al., 2019). This indicates underlying mechanisms pushing tissues towards a cancerous state by allowing for mutations to be positively selected for on a cellular level, which in time compromise healthy tissue function (Inigo Martincorena, 2019).

The high proportion of post-mitotic cell types in the brain makes it an interesting organ for study in the context of somatic variation (Fig 1.1A). While neurons are thought to have low replication rates outside of regions of neurogenesis (Bergmann et al., 2015), actively dividing non-neuronal (*e.g.* glial) populations aggregate mutations through processes linked to mitosis as we age. Somatic mosaicism, the phenomenon of post-zygotic mutations creating genetically distinct cell populations, is widespread in the brain. The observed rates of somatic variation range from 1.3-40% in regions of the brain, strongly scaling with age (Bushman & Chun, 2013; Kingsbury et al., 2005; Lodato et al., 2018). The effect of these mutations depends on many factors, including the developmental time of occurrence. Somatic mutations are often cited as candidate mechanisms for a wide range of psychiatric disorders such as autism spectrum disorder (Marshall et al., 2008) and schizophrenia (Stone et al., 2008). However, somatic mutations do not always lead to disease states. Neurons in healthy brains were shown to frequently harbor SNVs (~1500 mutations per neuron), and only some of which were linked to schizophrenia (Lodato et al., 2015). Similarly, neurons in healthy brains contain structural variants, such as CNVs (<0.5 somatic CNVs per neuron), that can be linked with autism spectrum disorder (Cai et al., 2014; McConnell et al., 2013). These mutations were often patient-specific and showed a low recurrence between cells. Given the highly region-specific and low replicating nature of adult neurons (Bergmann et al., 2015), brain tissue can tolerate a certain mutational load that is sequestered into smaller neuronal populations, as these mutations cannot spread further. Following the same logic, widely spread clonal mutations that are generated during early development and do not cause disease could generally be neutral or potentially beneficial (Bushman & Chun, 2013). This might also explain the dearth of extreme whole chromosome aneuploidy (defined as a gain or loss of five or more chromosomes) in the healthy brain, indicating that programmed cell death removes large-scale mutations that will likely have an extreme functional impact (Peterson et al., 2012). However, less aneuploid and mosaic euploid cells have been found in active neural circuitry (Kingsbury et al., 2005).

While the exact functionality of this neuronal diversity is still not known, in theory the mutational landscape of the mosaic brain shows an overview of mutational forces acting on early embryonic development, with a potential that early beneficial mutations were selected for in clonal cell populations for increased robustness in the brain. With aging neurons accrue somatic mutations individually though processes not related to mitosis (*e.g.* oxidative DNA damage, Lodato et al., 2018). These isolated, deleterious mutations add to the mutational load of this tissue on top of mutations acquired in early embryonic development, which can lead to eventual neurodegeneration.

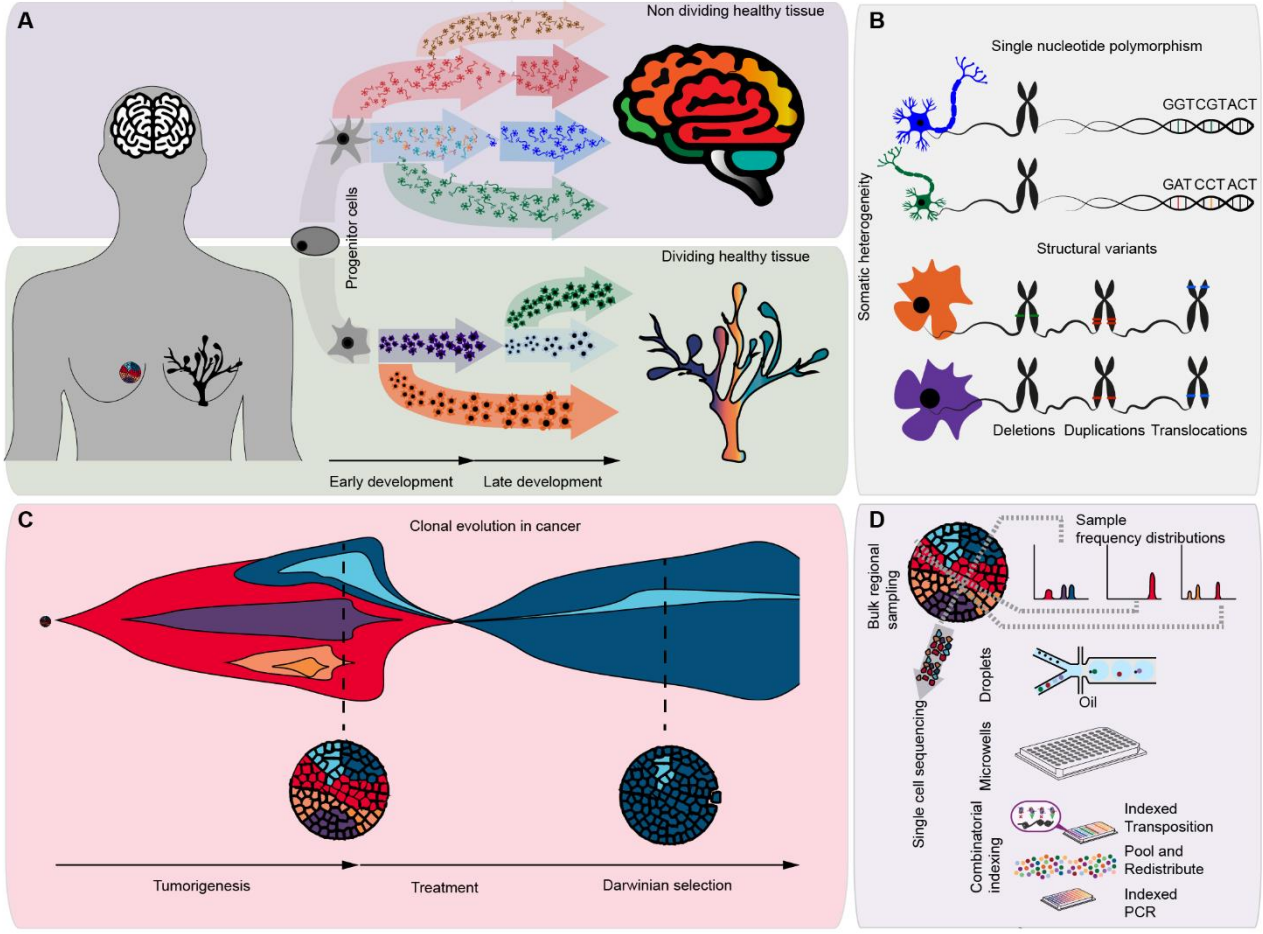

Figure 1.1 Somatic genomic mutations in healthy and diseased complex tissues. (A) Cells accumulate somatic mutations through early- and late-developmental processes. Mutations in neurons show a snapshot of early development and later accumulation of mutations in individual neurons. Lineage formation continues, however, in actively dividing tissues in the mammary gland. (B) Somatic mutations can range from single nucleotide polymorphisms to structural variants, such as deletions, duplications, and translocations. (C) Clonal evolution in a breast cancer tumor and selection for resistant genotype upon treatment. (D) Sequencing methods for detecting sub- clonal mutation can range from regional bulk sequencing approaches to different forms of single-cell sequencing.

1.1.3 Genomic heterogeneity in cancer

Cancer is a collection of related genetic diseases that are characterized by cells with abnormal growth, apoptotic dysfunction, the ability to modify their local microenvironments, and invade nearby tissues, while evading immune surveillance (Hanahan & Weinberg, 2011). Cells can attain these attributes by co-opting cellular pathways that regulate these characteristics via inherited germline and/or later acquired somatic mutation(s) (E. Y. H. P. Lee & Muller, 2010; Sanchez-Vega et al., 2018). As a result, certain types of cancer, including breast cancers, can go on to form abnormal growths called tumors (American Cancer Society, 2019b). The genes responsible for the formation of different types of cancer can vary largely from tumor to tumor (Beksac et al., 2017; P. J. Campbell et al., 2020; Cros et al., 2018; Grzywa et al., 2017; Weinstein et al., 2013), with indication of some recurrently mutated genes (Krepischi et al., 2012; E. Y. H. P. Lee & Muller, 2010; Shlien & Malkin, 2009; Lixing Yang et al., 2013). The latter were first studied in the context of heritable cancers where germline and later somatic mutations are often thought to result in the dysregulation of gene expression in important tumor suppressors and oncogenes (B. Liu et al., 2015; Mitelman et al., 2007). Tumor suppressors and proto-oncogenes are often depicted as the "brakes" and "gas pedals" of regulation within a cell, which points to the fundamental difference of their dysregulation in cancer. Oncogenes are the permanently activated forms of wild type proto-oncogenes by usually dominant, gain-of-function mutations. On the other hand, tumor suppressor genes are deactivated forms of wild type genes by recessive, loss-of-function mutations (Osborne et al., 2004).

Activation of a proto-oncogene can take many forms ranging from point mutations, which decouple a molecular switch from cellular regulation (*e.g.* Ras proteins), translocations that create fusion proteins with altered regulation of protein expression (e. g. *BCR/ABL*), and copy number alterations where the amplified expression of the proto-oncogene is the result of multiple active copies of it throughout the genome (*e.g. c-MYC*). These dominant genetic alterations result in gain of function changes, where only one effected chromosome is needed for phenotypic change, such

as cell growth, cell proliferation, and survival (Bertram, 2000; E. Y. H. P. Lee & Muller, 2010; Osborne et al., 2004; Rubin, 1998).

In contrast, most loss of function mutations occurring in tumor suppressor genes are recessive by nature. The phenomenon of the two-step inactivation a tumor suppressor in a cell was proposed by Alfred Knudson, which later became known as the two-hit hypothesis of tumorigenesis. This theory, first shown in retinoblastoma for the tumor suppressor gene *RB1*, posits that both alleles of a tumor suppressor gene need to be inactivated for phenotypic change to be observed, which can happen either through mutations or epigenetic silencing (Knudson, 1971). The first hit can be inherited as a germline mutation or can be an early-acquired somatic mutation (sporadic cancer), making the tumor suppressor gene heterozygous. A second somatic mutation in the remaining normal functioning allele can result in the inactivation of the tumor suppressor gene. This can happen via a copy number loss in the tumor suppressor gene resulting in loss of heterozygosity (CNL-LOH) or by a copy number neutral loss of heterozygosity (CNN-LOH). CNN-LOH is most often the result of gene conversion via homologous recombination or by the duplication of the chromosome containing the recessive allele before or after the LOH event (Ryland et al., 2015; Tischfield, 1997). Epigenetic deactivation of the tumor suppressor gene can also result in LOH (Ryland et al., 2015). CNL-LOH is easier to detect due to the great loss of genetic material, which can be picked up by methods such as comparative genomic hybridization (arrayCGH), fluorescence *in situ* hybridization (FISH), and karyotyping. CNN-LOH events can only be discovered by methods which can provide information on both copy number and heterozygosity across the genome, such as SNP arrays, and whole exome or genome sequencing studies (Osborne et al., 2004; Ryland et al., 2015; Tischfield, 1997).

Tumor suppressor genes can often affect apoptosis (*PTEN*) and other processes linked to cell cycle regulation, such as cell division (*TP53*), and DNA repair (*BRCA1*, *BRCA2*) (Ryland et al., 2015). The most widely studied tumor suppressor gene encodes tumor protein p53 *(TP53)*. In humans *TP53* is located on chr17p13.1. p53 inhibits the propagation of genetically mutated cells,

with 15 additional isoforms which serve different roles in genetic regulatory pathways. p53 is activated by post-translational modifications which enable the protein to function as a trans-activator of genes downstream in pathways that often get modulated by other cofactor genes to regulate apoptosis, inhibit cell cycle, angiogenesis and metastasis. MDM2 degrades p53, thus acting as a negative feedback regulator. Missense mutations in the *TP53* gene allow escape from the degradative effects of MDM2. Germline mutants of this *TP53* are associated with familial Li-Fraumeni cancer syndrome, which results in multiple tumors of different tissues, including breast and a 100% penetrance by the age of 70 (Stracquadanio et al., 2016). Similar, somatic mutations in this gene can also be found in more than 50% of all cancer genomes (Muller & Vousden, 2013; Stracquadanio et al., 2016). This shows the importance of *TP53* in cancer, which, when mutated, leads to a form of p53 which is unresponsive to a variety of stress signals. Tumor suppressors are generally loss-of-function mutations, but mutations occurring in *TP53* often result in gain-of-function alterations of its isoforms leading to a dominant negative effect over the remaining wild type p53. This arguably makes *TP53* an oncogene as well. In primary breast carcinomas *TP53* has been shown to be mutated in 18%–35% of cases, with most missense mutations occurring in the DNA binding domain, which affects the binding of pro-apoptotic cofactors such as ASPP1 and ASPP2 (Lacroix et al., 2006; P. Yang et al., 2013).

DNA repair plays an essential role in maintaining genome stability and tumor suppression. High-risk genes such as *PTEN*, *BRCA1* and *BRCA2* have functional roles in DNA damage signaling, DNA repair processes and cell cycle checkpoints. As a result, when these genes get mutated, the cell's ability to repair double strand breaks (DSB) can be impaired. The repair of these breaks can be via the more error prone nonhomologous end-joining (NHEJ) or the two forms of homology-directed repair, gene conversion (GC) and single-strand annealing (SSA). While homology-directed repair is active during phases of the cell cycle following DNA replication (S,G_2 when sister chromatids are present), NHEJ is the dominant form of DSB repair during earlier phases (G_0, G_1, early S; Gudmundsdottir & Ashworth, 2006; Minami et al., 2014) After DNA damage,

BRCA2 mobilizes and regulates the activity of RAD51, which functions as a recombinase during the GC repair pathway. When a cell is BRCA2 deficient, GC is downregulated and the more error prone (RAD51 independent) SSA pathway is upregulated (Gudmundsdottir & Ashworth, 2006).BRCA1 also co-localizes with RAD51, but is thought to be a regulator further upstream in the DNA repair pathway and consequently, BRCA1 deficiency results in the downregulation of both GC and SSA. In addition, BRCA1 also has been suspected to play a role in a more precise form of NHEJ (Ku/DNA-PKcs dependent) as opposed to the more error prone micro-homology mediated NHEJ. BRCA1 response is modulated according to the stress type by the damage response checkpoint genes ATM (*e.g.* ionizing radiation) and ATR (*e.g.* UV damage), which also regulate p53, which in parallel acts as an activator for apoptosis (Gudmundsdottir & Ashworth, 2006; Lacroix et al., 2006). As a response to DNA damage, the cell cycle is halted at the G_2-M or S phase checkpoints depending on which serine of BRCA1 is phosphorylated (Ser1387: G2/M; Ser1387:S) by ATM or ATR. Depending on the position of phosphorylation, BRCA1 stimulates the transcription of p21 or p27, which are required for the cell-cycle arrest at G2-M or Intra S phase, respectively (Gudmundsdottir & Ashworth, 2006). CHK2 also participates in modulating the response by BRCA1 by phosphorylating Ser988 after activation by ATM. In addition to DSB repair, BRCA1 has also been linked to a sub-pathway of nucleotide-excision repair (NER), a form of single strand break repair where BRCA1 in association with MSH2 and MSH6 preferentially repair base lesions from the transcribed strand (Gudmundsdottir & Ashworth, 2006; Minami et al., 2014; P. Yang et al., 2013). Phosphatase and tensin homolog on chromosome 10 tumor (*PTEN*) suppressor gene also regulates G2/M arrest and apoptosis by controlling p53 degradation by MDM2 through the regulation of PI3K-mediated receptor tyrosine kinase signaling to the survival kinase AKT. The aforementioned PI3K/AKT pathway is thought to modulate BRCA1 through phosphorylation at S694. BRCA1 may also affect the PI3K/AKT pathway by direct downregulation and by acting on upstream kinases of AKT (Gudmundsdottir & Ashworth, 2006). Due to the frequent mutations in genes associated with DNA repair pathways in breast cancer, novel

therapeutic approaches exploit the cancer cells' increased dependence for survival on alternative DNA repair pathways. This provides the opportunity to target these pathways with inhibitors, resulting in deleterious genomic instability of the cancer cells while sparing normal cells that have the original impaired DNA repair pathway intact, an approach termed "synthetic lethality". The best clinical example of this is the targeted treatment with PARP inhibitors of BRCA deficient cells (O'Connor, 2015; Van Gent & Kanaar, 2016). PARP inhibitors block single-strand break repair, which is synthetic lethal in cells with defective homologous repair pathways. By combining targeted inhibitors of DNA damage response with radiotherapy, the efficacy of treatment could potentially improve because of radiosensitization of cells. Similarly, chemotherapy can be improved by choosing complementary DNA damage inducing agents with the DNA repair mechanism targeted by the inhibitor (O'Connor, 2015).

In addition to studying frequently mutated cancer-associated genes, a contributing salient feature of cancer solid tumors is their intra-tumor heterogeneity and the clonal evolution of their constituents (Fig 1.1C). The prevailing theory of tumor evolution (Nowell, 1976) posits that this is the result of individual cells forming diverging distinct clonal subpopulations that have their genome and their clonal spread within the tumor shaped by selective pressures, thus further affecting the overall mutational composition and tumor heterogeneity (N. E. Navin, 2015; Nowell, 1976). As a result of these evolutionary pressures, tumors contain mutations that can direct tumor evolution (*i.e.* are actively selected on), namely driver mutations, and non-contributing passenger mutations that are created by the same operative mutational and DNA repair processes that create driver mutations. Passenger mutations, however, do not offer a selective advantage to the cells they are present in and therefore do not influence the clonal spread within the tumor (Nik-Zainal, Alexandrov, et al., 2012). It is important to note that selective advantage is defined in the context of other cells present within the tumor and its environment. Therefore, similarly to other instances of Darwinian evolution, a shift in evolutionary pressures can change tumor composition rapidly.

This is particularly emphasized in the case of treatment, which can confer a selective advantage to cells harboring variants promoting resistance (McGranahan & Swanton, 2017, Figure 1.1C).

Intra-tumor heterogeneity can vary largely between tumor types, even when considering differences in sampling procedure, tumor stage, and sequencing depth (Alexandrov et al., 2020; McGranahan & Swanton, 2017). For example, melanoma and lung cancer exhibit larger coding mutation burdens than other tumor types, which is likely due to the involvement of years of exposure to exogenous mutagens (McGranahan & Swanton, 2017, *e.g.* ultraviolet light and tobacco carcinogens). These mutational processes leave their patterns in the genome over a patient's life, which can then be assessed via observing recurrent mutational signatures across tumors. Individual signatures have to be decoded from the aggregate genomic patterns of multiple cancer patients by solving the blind source separation problem of pre-defined biologically relevant mutational classes via nonnegative matrix factorization. Studies using large cancer cohorts have identified unique signatures related to various etiologies, including smoking and UV exposure (Alexandrov et al., 2013, 2016, 2020; Nik-Zainal, Alexandrov, et al., 2012).

The high heterogeneity observed in different types of cancers also provides a unique challenge in identifying driver mutations located in coding (M. H. Bailey et al., 2018) and non-coding regions (P. J. Campbell et al., 2020) of the genome. Consortium projects, such as the Pan-Cancer Analysis of Whole Genomes, can help identify driver mutation via the aggregation of data across large sets of cancer genomes. Recently, an analysis on 2,658 whole-cancer genomes across 38 tumor types (with matching normal tissues) was published to identify drivers of cancer (P. J. Campbell et al., 2020). On average, 4-5 driver mutations were identified per genome in coding and non-genomic elements. Interestingly, approximately 5% of cases had no identifiable driver mutations, indicating that the full discovery of all driver mutations has not been achieved (P. J. Campbell et al., 2020). This study also observed examples of patterns of clustered point mutations and SVs related to mutational processes that can generate multiple mutations in a single catastrophic event. The three studied processes were chromoplexy (shuffled chain rearrangements

resulting from repair of co-occurring double strand breaks observed in 17.8% of all samples), kaetegis (locally clustered point mutations with a single DNA strand bias observed in 60.5% of all cancers), and chromothripsis (tens to hundreds of DNA breaks located in one or few chromosomes that are then randomly reassembled) (Berger et al., 2011; P. J. Campbell et al., 2020; Korbel & Campbell, 2013; Nik-Zainal, Alexandrov, et al., 2012).

Somatic mutations can be used to order events that occur in tumor evolution. This is based on the assumption that mutations shared by all cancer cells within the sample (*i.e.* mutations happening before the last selective sweep) happen before subclonal mutations, which occur after the emergence of the most recent common ancestor. The resulting differences in the variant allele frequencies of point mutations can therefore inform on the underlying clonal architecture of bulk tumor samples (Durinck et al., 2011; Nik-Zainal, Van Loo, et al., 2012). Similarly, copy number changes can further help with defining molecular clocks in a tumor as mutations occurring in a region with a copy number gain will be present at a differing number of chromosomal copies depending on their time of occurrence. Mutations preceding the copy number change will be duplicated, but mutations after the copy number gain will be only in one chromosome copy. This presents as differences in the ratio of heterozygous to homozygous mutations in regions of CN-LOH, providing a measure of the age of duplication (Durinck et al., 2011; Gerstung et al., 2020; Jolly & Van Loo, 2018). Based on this analysis chromothripsis presented as an early event in multiple types of cancers (*e.g.*: liposarcomas, prostate adenocarcinoma and squamous cell lung cancer) (P. J. Campbell et al., 2020). Using the same concepts of evolutionary ordering a recent study has shown the evolutionary history of 2,778 cancer samples from 2,658 unique donors across 38 cancer types. Interestingly, this study revealed a few common driver genes (*e.g. TP53, KRAS, PIK3CA*) to be frequently mutated in early tumor evolution, with an increase in the number of potential driver genes across tumors in later disease progression (Gerstung et al., 2020). This underscores the diversity of regulation of later tumor development.

Next to bulk approaches, single-cell methods offer an additional means to ascertain ordering of events in a single tumor. As opposed to bulk sequencing, the genetic makeup of a cell is directly assayed and therefore its relation to other sampled cells in the tumor can be inferred based on its mutational profile. Single-cell genomic analysis can thus elucidate subclonal heterogeneity and evolutionary history by detecting SNPs with high depth sequencing or by using low-pass sequencing to detect CNVs. High-coverage studies showed clonal architecture in muscle invasive bladder cancer (Y. Li et al., 2012), clear cell renal carcinoma (X. Xu et al., 2012), myoproliferative neoplasm (Hou et al., 2012), colon cancer (C. Yu et al., 2014), childhood lymphoblastic (Gawad et al., 2014) and secondary myeloid (Hughes et al., 2014) acute leukemias by sequencing whole exomes or multiple genomic loci. Other studies used read depth analysis on whole genome low-coverage data for CNV detection in single cells of primary tumors of breast cancers (N. Navin et al., 2011; Y. Wang et al., 2014).

1.1.4 Genomic heterogeneity in breast cancer

Breast cancer encompasses a diverse group of solid tumors originating from breast tissue (most often from the cells lining the lobules or ducts) that are the leading type of cancer in women worldwide. In 2019, there were an estimated 268,600 new cases of breast carcinomas in the United States with approximately 41,760 resulting deaths that year alone, making this the second leading cause of death by cancer (American Cancer Society, 2019a). The leading causes for these deaths are distantly located metastases, which are currently still considered difficult to cure (Fouad et al., 2015). Breast cancer outcomes are influenced by many factors, including race, ethnicity, sex of the patient, environmental factors and patient history, the anatomical region, tumor grade, and histological assessment of the tumor (Sinn & Kreipe, 2013). In this dissertation I will focus on invasive ductal carcinomas presenting in female patients, which is the most common lethal form of this disease. This group of carcinomas show high biological and mutational variability, which impedes our understanding of treatment, response, and outcome (Koren & Bentires-Alj, 2015; Martelotto et al., 2014). This creates the need for the robust characterization of heterogeneity.

Breast cancers traditionally have been classified based on marker expression profiling via immunohistochemistry (IHC) into at least four molecular subtypes that correlate with the presence or absence of progesterone (PR), estrogen (ER), and human epidermal growth factor receptor 2 (HER2) receptors present on the tumor cell surfaces (Dent et al., 2007; Morris et al., 2007). Based on this classification most tumors are hormone receptor positive (ER+, PR+/-, ~65-80% of breast cancer) followed by HER2-receptor positive (15-30%) tumors, and Triple Negative (~10-25%) tumors negative to all markers (Dent et al., 2007; Lehmann et al., 2011b; Morris et al., 2007). Additional markers such as Ki67 (marker indicating the number of actively dividing cells), p53 (marker for most frequently mutated tumor suppressor) and EGFR (proliferation marker frequently mutated in breast cancers) can improve the power of prognosis based on these IHC categories, and are actively being used in clinics (Cheang et al., 2009; Kobayashi et al., 2013).

Expression profiling of thousands of genes via DNA microarray technologies and next generation sequencing has led to the identification of recurrent molecular categories of breast cancer (Sørlie et al., 2003): Luminal A, Luminal B, HER2-enriched (HER2E), Normal-like, Basal-like, and Claudin-low (Prat et al., 2010; Yersal & Barutca, 2014). While these molecular categories and IHC-defined subtypes can largely overlap, they are not always mutually inclusive. One example being that 80% of basal-like tumors are triple negative, whereas 70% of triple negative tumors are basal-like (Goldhirsch et al., 2013). Luminal A and Luminal B subtypes constitute approximately 74% and 10% of all breast cancers. Both are ER+ and/or PR+, which makes them viable targets for hormone treatment. However, Luminal A cancers (Her2- and have low levels of Ki67 protein) have better prognoses than Luminal B subtypes (Her2-/ Her2+ high levels of Ki67 proteins) due to lower proliferation rates (Prat et al., 2015). The rarest subtypes (~4%) are Her2-enriched cancers, which lack hormone receptors but can be treated by novel Her2 targeting therapies. Triple negative breast cancers (TNBC; ~14%) have the worst prognoses due to the absence of targetable receptors (American Cancer Society, 2019b, 2019a; Koren & Bentires-Alj, 2015). In addition, while TNBC tumors respond well to initial chemotherapy, the inherent high

intra-tumor heterogeneity in this molecular increases the chance for an early relapse (Carey, 2011; Dent et al., 2007; C. Kim et al., 2018b; Liedtke et al., 2008). As a result, recurrent genomic mutations have been closely studied in TNBC tumors to develop targeted therapies. SNPs and CNVs were found in tumor suppressor genes *TP53*, *RB1*, *PTEN*, *BRCA1*, *INPP4B* and oncogenes at *PI3K*, *AKT3*, *EGFR*, all of which modulate proliferation and survival pathways (Gonzalez-Angulo et al., 2011; Herschkowitz et al., 2012; Park et al., 2014; Shah et al., 2012). Specifically, studies showed the aberrant activation of PI3K/AKT/mTOR and MAPK/MEK/ERK pathways, normally tasked with regulating cell cycle entry and proliferation necessary maintaining normal human physiology (Gonzalez-Angulo et al., 2011; Saini et al., 2013). Existing MEK inhibitors, including the MEK1 and MEK2 inhibitor, Trametinib (Zeiser, 2014), have shown antiproliferative effect *in vitro* and *in vivo* (Heiser et al., 2012; Hoeflich et al., 2009; Lehmann et al., 2011a; Pratilas et al., 2009; Saini et al., 2013), but despite having evidence for TNBC cancer cells strongly relying on these proliferation pathways, patient drug trials often showed inherent or acquired resistance after initial treatment success (Adjei et al., 2008; Rinehart et al., 2004; Zawistowski et al., 2017). While combination treatments with other drugs show promise (Ramaswamy et al., 2016), the inherent intra-tumor heterogeneity of TNBC tumors is problematic due to the presence or evolution of potentially resistant subclones in response to treatment. This mechanism of Darwinian selection (Fig 1.1C) on subclones upon treatment have been recently shown using regional bulk sequencing (Yates et al., 2015) and single-cell sequencing (C. Kim et al., 2018a). These studies show that drug resistant clonal populations follow branched evolution where resistant subclones become fixed in the population (*i.e.* increase in frequency) during a selective sweep and less resistant ones are extinguished during neoadjuvant chemotherapy.

1.1.5 Single-cell methods for genomic variant detection

Characterization of the frequencies of cells with somatic mutations poses technological challenges for the detection of variants present in a smaller fraction of investigated cell populations,

such as complex tissues and tumors. This is due to bulk sequencing methods having limited sensitivity in detecting somatic variation in less than 5% of the cell population (Cibulskis et al., 2013), thus filtering out potentially important variants that have arisen later developmental or tumor evolution (Gawad et al., 2016; N. E. Navin, 2015). Modifications to bulk sequencing methods attempt to solve this by deconvolving the combination of signal based on relative clonal contributions of regional or temporal samples of a tissue or tumor (Fig 1.1D). However, for these methods to work, clonal origin of cells has to be established; therefore rare sporadic mutations, such as the ones in the brain, cannot be identified (Dou et al., 2018; X. Li et al., 2018; Williams et al., 2018).

The true breakthrough came with the advancement of single-cell genome sequencing methods, which permit cells to be sequenced individually. This has allowed for the accurate profiling of SNPs (Lodato et al., 2015, 2018) and CNVs (Cai et al., 2014; McConnell et al., 2013) in the brain, and across multiple types of cancer (Casasent et al., 2018; Gao et al., 2016; C. Kim et al., 2018b; Y. Li et al., 2012; N. Navin et al., 2011; Ni et al., 2013; Y. Wang et al., 2014; C. Yu et al., 2014). Inherent biases and challenges of applied methods still remain however. Currently, there is no "golden standard" for single-cell genome sequencing, with multiple existing technologies and computational methods providing different advantages depending on the studied biological question. One inherent challenge lies in the low amount of DNA that can be isolated from individual cells, which, depending on cell type and ploidy, can range from 6 pg in diploid cells to 12 pg of DNA in cancerous aneuploid cells (De Bourcy et al., 2014; N. E. Navin, 2015). To address this problem, different whole genome amplification (WGA) strategies have been developed to address the need for larger quantity of DNA from the entire genome for sequencing, while avoiding introducing amplification-based biases. The two most widely used WGA methods are degenerate oligonucleotide-primed PCR (DOP-PCR; Telenius et al., 1992) and multiple displacement amplification (MDA; Dean et al., 2001). DOP-PCR uses degenerate oligonucleotide priming across the genome followed by PCR amplification. While this method has high uniformity of

amplification, it preferentially amplifies selected sites due to variable PCR efficiency. Even using tetraploid cancerous cells led to only covering approximately 10% of the human genome (N. Navin et al., 2011). The low noise from the uniform amplification of this method makes DOP-PCR a good candidate for detecting CNVs, but falls short in SNV related studies due to the higher error rate of thermolabile polymerases (Deleye et al., 2017; Gawad et al., 2016). Multiple displacement amplification (MDA) methods use the high-fidelity phi29 polymerase, a thermostable enzyme, which has low error rates, high processivity, and genomic coverage (~70% for human genome for diploid and ~96% for tetraploid cells), but introduces amplification bias during the initial strand displacement and exponential amplification of the enzyme (Fu et al., 2019; Gawad et al., 2016; Leung et al., 2015). Multiple approaches aim to improve the amplification bias of this method either by creating water-oil emulsions of input genomic DNA (Fu et al., 2019) or by modifying the pre-amplification steps to be quasi-linear (Zong et al., 2012) or linear (C. Chen et al., 2017). Hybrid methods such as multiple annealing and looping-based amplification (MALBAC) aim to combine the advantages of PCR and MDA-based methods. MALBAC starts with a pre-amplification step where random priming of thermostable enzymes introduces common sequences and temperature cycling creates loops of the isothermal amplicons. These loops decrease noise by stopping runaway exponential amplification before the following PCR amplification steps. This quasi-linear amplification method results in an intermediate coverage of the human genome with low noise and intermediate amounts of introduced errors relative to the two previous methods (Gawad et al., 2016; Zong et al., 2012). Another method aims to decrease amplification bias via excluding random priming and exponential pre-amplification steps. Linear Amplification via Transposon Insertion (LIANTI) achieves this by first using a hyperactive form of Tn5 transposase, now commonly used for library preparation methods, to fragment accessible double stranded DNA and ligate loaded synthetic oligonucleotides on both ends (Adey et al., 2010). A T7 promoter is inserted via the transposition event followed by *in vitro* transcription and reverse-transcription of the second strand to form amplicons. This method achieves low error and high genome coverage (97% of genome

covered, Chen et al., 2017). Decreasing signal to noise ratio of reads and allelic drop out has allowed for better CNV and SNP calling in individual cells.

There are a wide range of computational methods for analyzing single-cell somatic variation. Initial single-cell studies applied software developed for bulk data, and more recently, single-cell specific approaches have been established. The primary computational challenge for analysis is the low coverage of individual cells, the inherent allelic dropout, and the amplification bias of genome amplification methods (Fan et al., 2019). As this dissertation puts emphasis on profiling high number of cells at low coverage, I will focus on CNV calling methods. Methods used in bulk data, such as break point calling based on mismatches of aligned read pairs are not feasible in single-cell data due to low coverage (Fan et al., 2019; Knouse et al., 2016; X. Wang et al., 2018). Therefore, most CNV profiling algorithms rely on read aggregation within large, often megabase pair scale binned windows within individual cells, which then can be compared to identify gain or loss of chromosomal material. In order to have representative genomic windows, copy number calling algorithms correct for the varying mappability and GC content of the genome by normalizing bin values and using varying bin sizes (Fan et al., 2019; X. Wang et al., 2018). Quality of individual cells can then be assessed to remove outliers based on measures such as the mean absolute deviation of pairwise differences (MAPD) of adjacent windows (Garvin et al., 2015). Similarly, regions with highly repetitive elements such as centromeres and telomeres are commonly removed from the data. This is then followed by the segmentation of genomic windows into contiguous regions, defining breakpoints, and estimating absolute copy numbers (Fan et al., 2019). An exception to existing methods is SCNV, which uses a bin free approach to infer copy number changes (X. Wang et al., 2018). While available methods differ in window calling, normalization, and filtering, they use segmentation strategies which follow one of three common strategies. Contiguous copy number regions are identified either by using a Hidden Markov Model (HMM) for the segmentation and imputation of copy number (Shah et al., 2006), an objective function to approximate the underlying constant function of the data and the introduced variation by copy

number (Nilsen et al., 2012), or a sliding window approach, which approximates changes via statistical testing (Ha et al., 2012; Olshen et al., 2004a).

From the plethora of available copy number calling methods, the four most widely used are Ginkgo (Garvin et al., 2015), HMMcopy (Laks et al., 2018; Shah et al., 2006), AneuFinder (Bakker et al., 2016), and CopyNumber (Nilsen et al., 2012). CopyNumber pools cells and segments them together via an objective function and Ginkgo uses a modified version of Circular binary segmentation (CBS) with a sliding window for segmentation (Olshen et al., 2004a). Both of these require a following step to estimate the absolute copy number (Fan et al., 2019; X. Wang et al., 2018). Hidden Markov Model based approaches (AneuFinder, HMMcopy) do not require this extra step as segmentation and imputation of copy number is done at the same time. When Ginkgo, HMMcopy and CopyNumber were benchmarked on simulated and real datasets, HMMcopy outperformed other methods in terms of speed, and Ginkgo outperformed other method in terms of accuracy, but none of the methods exceeded 80% accuracy (Fan et al., 2019). A similar comparison between CBS and HMM methods showed high accuracy for >5 Mbp copy number changes (CBS was more sensitive to deletions an HMM to amplifications) in simulated data, but decreased substantially when CNVs were rare and present in small cell populations (Knouse et al., 2016). These results indicate that the inherent noise of sparse single-cell DNA-seq data is still challenging for CNV detection and that progress has to be made in improving the quality and quantity of cells profiled.

Throughput presents an additional technical challenge for single-cell genome sequencing. Depending on the queried biological problem, two types of approaches have been used in single-cell studies. Previous work focusing on a specific cell population or a rare cell types, such as neurons (Cai et al., 2014; McConnell et al., 2013) or circulating tumor cells (Ni et al., 2013), have selected or enriched for cells expressing distinguishing marker genes. These cells then were deeply profiled for SNPs and smaller CNVs in downstream analyses. Classically, these methods have used 96 well plates with individual cells genome amplified in wells (Gawad et al., 2016; N. E. Navin,

2015). An alternative approach to this is to increase the throughput of low coverage cells with the hope that a representative portion of the studied tissue is sampled. The number of cells needed to be a representative sample is dependent on many factors relating to the heterogeneity of the studied tissue and the false negative and false positive rates resulting from genome amplification and sequencing. For these studies, the power lies in finding the inherent relational structure of low coverage cells based on shared large scale CNVs. Aggregate groups of similar cells then can be used to get a higher resolution view of substructures, and potentially order events based on development or evolution. This approach has been particularly powerful in cancer research, as the selective pressures and evolutionary history of clonal population were revealed in breast cancer (Casasent et al., 2018; Gao et al., 2016; C. Kim et al., 2018b; N. Navin et al., 2011; Y. Wang et al., 2014), bladder cancer (Y. Li et al., 2012), colon cancer (C. Yu et al., 2014), and in single circulating tumor cells of lung cancer patients (Ni et al., 2013).

Throughput of single-cell methods for profiling transcriptomes has increased significantly in the recent years via the use of physical compartmentalization of cells using microfluidics and aqueous droplet technologies. However, this has translated only recently into single-cell genomics due to the difficulty of preforming genome amplification within physical compartments (*e.g.* microfluidics or microwell equipment, Fig 1.1D; K. Zhang, 2017). Virtual compartmentalization of single cells via a series of tagging, mixing and sampling of mixed cells has emerged as an alternative approach. Combinatorial indexing relies on first tagging and indexing cells with uniquely indexed Tn5 transposase, followed by the mixing all cells, down-sampling cells, and tagging them a second time via PCR. The combination of barcodes introduced in the two rounds of indexing allow for individual cells to be computationally distinguished. This removes the need for need for microfluidics or microwell equipment as both rounds of indexing happen in 96 well plates. In the first round, each well has a separate species of Tn5 inserting well-specific indexes into open regions of chromatin. Similarly, PCR reactions are well-specific by adding indexed adapters during the second round of indexing. It is necessary that tagmented nuclei remain intact after the first

round of indexing and redistribution of cells into the PCR wells. Scalability of this strategy only relies on the number of available indexes introduced in the two rounds, making this approach very high throughput (D. a Cusanovich et al., 2015; K. Zhang, 2017). This method was first implemented for chromatin accessibility profiling by modifying the assay for transposase-accessible chromatin (ATAC-seq) so that the introduced transposases introduces barcodes into open regions of chromatin in the first round of indexing (Buenrostro et al., 2013a; D. A. Cusanovich et al., 2015). In section 1.2.2, I will focus on studies conducted on understanding chromatin accessibility across tissues and diseased states by using this combinatorial indexing method.

Combinatorial-indexing strategies have also been extended to profile other properties including transcription, chromatin folding, and DNA methylation (Cao et al., 2017; Mulqueen et al., 2018; Ramani et al., 2017; Yin et al., 2018). We set out to develop a combinatorial indexing method to profile the whole genome, by making it accessible to Tn5 transposition, while retaining intact nuclear scaffolds for the redistribution into the second round of indexing. We aimed to profile copy number variation across a significant number of cells required to detect low frequency aneuploidy in the brain or to accurately track clonal evolution in cancer (Vitak et al., 2017a).

1.2.1 Cell types and cell states in healthy complex tissues

The robustness of a tissue's ability to respond to environmental changes relies on its cellular heterogeneity. Cells of a tissue formulate characteristic responses to external and internal impulses in an organized manner (Altschuler & Wu, 2010). This is aided by the differentiation of cells into cell types and lineages. Separation between cell types is attenuated into distributions as cells fall into cell states, which are much more plastic, as cells can transition between states and are therefore defined by the conditions giving rise to them and the time scale they are happening on (Janes, 2016). This combination of cell type heterogeneity and cell state regulatory heterogeneity has classically been portrayed as a landscape of cellular "potential energy" by Waddington (Fig 1.2), where cell types lie in valleys separated by energetic barriers that represent lineage commitment, and cell states are represented by the width of these valleys which can be explored via state

switching (Janes, 2016; Waddington, 1957). The process of commitment to cell type and state is much more dynamic than what is possible on the level of genetics as somatic mutations accumulate randomly, and selection on mutations happens over long timescales. In addition, the ability to reverse cell lineage commitment artificially into stem cells proves the dynamic nature of cell lineage and state maintenance (Ladewig et al., 2013; Takahashi & Yamanaka, 2016). Indeed, recent advancements are revealing cell fates to be much more plastic than previously thought and bring into question the hierarchical structure of Waddington's landscape, indicating that the landscape can be turned on its developmental axis. This allows for single cells to de- and re-differentiate into alternate lineages, as necessary genes can be silenced or reactivated via proper stimuli (Ladewig et al., 2013).

Plastic changes in cell fates and states can be conveyed via the dynamic control of chromatin architecture in a cell. Chromatin is organized around nucleosome core particles, which consist of histone octamers (two H2A, H2B, H3 and H4 molecules) and ~146 base pairs of DNA wrapped around them (Venkatesh & Workman, 2015). These organizational blocks of nucleosome core particles are connected together by linker DNA forming a "beads on a string" like primary chromatin structure (Fyodorov et al., 2018). The higher-order organization of this topology is controlled by H1 linker histone proteins. These modulate nucleosome stability by binding to DNA entering and exiting the nucleosome core complex (Fyodorov et al., 2018; Hergeth & Schneider, 2015). This imparts regulatory function by "condensing" (hetero-) and "opening" (eu-) chromatin in a region therebyz blocking or allowing access of transcriptional machinery to the genes found there. This process is controlled via the localized epigenetic modification of DNA and histones by modifying chemical and structural properties of a region. In turn, these covalent epigenetic marks are recognized by chromatin reorganizers, which can then mobilize histones and change chromatin compaction (Valencia & Kadoch, 2019).

The most prevalent covalent modification of DNA in vertebrates is the methylation of cytosines (5-methylcytosine, 5mC), which are deposited by methyltransferases (DNMTs) at CpG

dinucleotides (Lyko, 2018; Valencia & Kadoch, 2019). DNMTs vary in their regulatory roles in a cell. Novel methylation is deposited by DNMT3A/B, and DNMT1 is thought to serve the role of maintenance by methylating hemimethylated sites (M. Okano et al., 1998; Masaki Okano et al., 1999; Sen et al., 2010). Most CpG sites are methylated and overlap with regions of high nucleosome occupancy and transcriptionally silenced genes (Collings & Anderson, 2017; Jones, 2012). The remaining non-methylated CpG sites can frequently be found clustered closely in CpG islands near active mammalian promoters (Kulis & Esteller, 2010). While it is not fully understood, CpG methylation is thought to suppress transcription via several mechanisms. Methyl groups can block DNA recognition at important sites (*e.g.* GC boxes) required by transcription factors (TFs, *e.g.* SP1, SP3) for transcriptional activation. Alternatively, methylated CpGs can also be preferentially bound by factors that further recruit repressive histone modifiers (such as histone deacetylases) (Handy et al., 2011). The methylation process of CpG sites is directly opposed by the iterative oxidization of 5mC into relatively stable (5-hydroxymethylcytosine) and transient (5-formyl-methylcytosine, 5-carboxyl-methylcytosine) forms of epigenetic marks by ten-eleven translocation (TET) enzymes (Rasmussen & Helin, 2016). Overall, TET enzymes preferentially bind CpG dinucleotides through their catalytic domains (Hu et al., 2013). TET1 and TET3 have an extra CXXC domain, which increases their affinity for 5mC-, 5hmC- and 5caC-modified CpGs (Jin et al., 2016; Y. Xu et al., 2011). Studies have established that distinct methylation patterns of CpG sites are required for cell differentiation during development and in somatic cells (Schübeler, 2015; Valencia & Kadoch, 2019).

Post translational histone modifications also play an important role in cell fate and state specification within a tissue. There are more than 150 histone-modifying proteins, which deposit >200 distinct types of histone tail and globular domain modifications to regulate the positioning and function of other chromatin regulatory proteins and protein complexes (Audia & Campbell, 2016; Khare et al., 2012; Tessarz & Kouzarides, 2014). These proteins are generally classified as writers and are directly opposed by proteins acting as erasers of histone modifications. The writing

and erasure of histone marks regulate transcription via the combinatorial positioning of activating or repressive marks at promoters and regulatory elements (Choukrallah & Matthias, 2014). Acetylation of core histone N-terminal extensions has been extensively studied in the context transcriptional activation. Acetyl groups are added to lysines by histone acetyltransferases (HATs) and removed by histone deacetylases (HDACs) (Gallinari et al., 2007). Acetylation negates the positive charge of lysines of the histone tail N-terminal extensions and as a result decreases the affinity of the histone tail for the negatively charged DNA backbone. This leads to a more relaxed chromatin where activating acetylation marks, such as H3K27ac, can further be recognized by chromatin remodelers recruited by the transcriptional machinery to reposition, eject, slide or alter the composition of these "loosened" nucleosomes (Gallinari et al., 2007; Martin & Zhang, 2005; X. J. Yang, 2004). This process is directly opposed by HDACs that strengthen histone tail-DNA interactions via the removal of an acetyl group. Compaction of chromatin can be further increased via the methylation of certain lysine residues (such as the methylation of lysine 9 and the methylation of lysine 27 of histone H3) by methyl transferases (Gallinari et al., 2007; Martin & Zhang, 2005). Proteins with chromodomains, (*e.g.* HP1) can recognize marks such as H3K9trime3 and promote DNA methylation and the assembly of heterochromatin (Lachner et al., 2001; Martin & Zhang, 2005). Methylation can also serve as activating marks (*e.g.* H3K4me3) and are removed demethylases (Choukrallah & Matthias, 2014). The "histone code" is still not fully understood in its complexity as it varies based on the epigenetic mark, the histone and the modified residue. Additional epigenetic marks include phosphate groups (kinase and phosphatase proteins), arginine methylation, ubiquitination, citrullination, SUMOylation, ADP ribosylation, deamination and crotonylation (Valencia & Kadoch, 2019).

Finally, chromatin topology is modified by large chromatin remodeling complexes consisting (CRCs) of >100 protein subunits. CRCs recognize epigenetic marks and employ ATP hydrolysis to slide or eject nucleosomes thereby increasing DNA accessibility to DNA binding proteins and the transcriptional machinery (Clapier & Cairns, 2009). CRCs are grouped into four

classes: SWI/SNF (mSWI/SNF (BAF)), imitation SWI, INO80, and nucleosome remodeling and deacetylation chromodomain helicase DNA-binding complexes (Clapier & Cairns, 2009; Valencia & Kadoch, 2019). Further subclassifications of these groups is dependent on the combination of accessory subunits within a CRC (often with DNA and histone binding roles) to the SWI/SNF2 like core ATPase/helicase unit shared across all classes (Poynter & Kadoch, 2016).

Lineage and cell state specification requires the establishment of characteristic transcriptional programs. These are regulated by transcription factors that bind DNA regulatory elements (REs) of specific genes. Cell types and states can have unique sets of expressed TFs associated with them, however, the same TF can have a different binding profile depending on cell type as binding motifs are affected by chromatin structure and epigenetic modifications. In addition, TFs interact with DNA methylation, nucleosome remodeling, and histone post transcriptional modifications thereby having the ability to reorganize focal chromatin structure. (Ahsendorf et al., 2017; Choukrallah & Matthias, 2014).This interplay between TFs and chromatin stucture is often portrayed as the scaffolding mesh of cellular epigenetic and transcriptional regulation under Waddington's landscape (Fig 1.2, Ladewig et al., 2013; Takahashi & Yamanaka, 2016)

Cells from the same genetic clonal lineage can occupy different cell types and states (Mathis et al., 2017), which can be defined by patterns in the expression of a variety of markers, identified via immunohistochemistry (IHC) and flow and mass cytometry (Giesen et al., 2014; Potts et al., 2012; Reuben et al., 2017). This, however, requires prior knowledge for accurate population stratification. Genome-wide analyses such as RNA sequencing, bisulfite sequencing, or the assay-for-transposase-accessible chromatin sequencing, can reveal cell states via shared patterns expression, methylation or chromatin accessibility between sets of cells. However, all assays have to be performed on the single-cell level, due to the heterogeneity of complex tissues (Goldman et al., 2019).

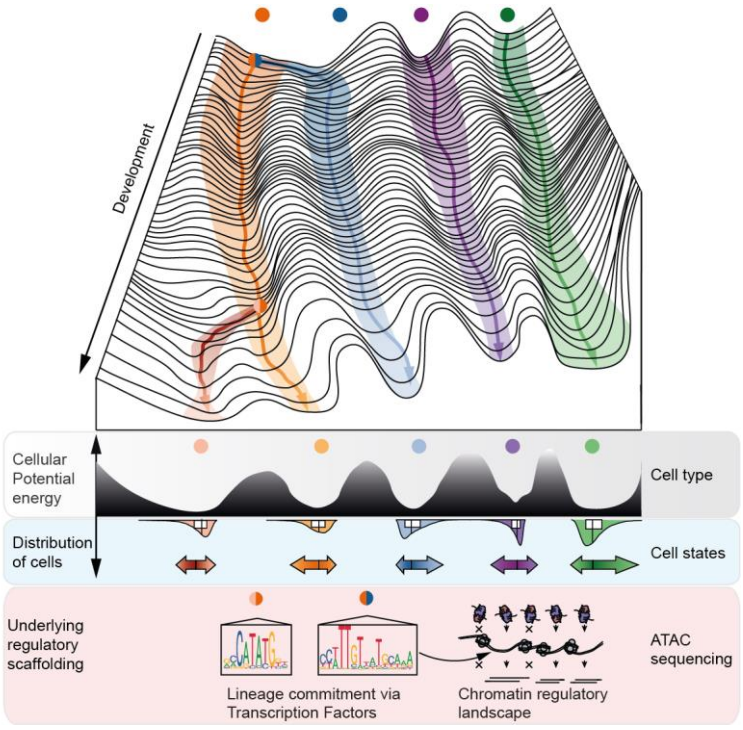

Figure 1.2 Waddington landscape of cell fate. Cells are portrayed as balls rolling down the epigenetic landscape of cellular potential energy. Cell types are separated by barriers which a cell cannot normally cross and cell states are the possible permutations, which cells can occupy upon external stimuli. Commitment into cell lineages can be decided via changes in chromatin architecture as a response to epigenetic modifications of DNA and nearby histones. Thereby, important lineage transcription factors can be recruited to promoters and enhancers of lineage commitment genes. Chromatin architecture can be profiled via ATAC-seq to understand the underlying regulatory architecture.

1.2.2 Single-cell methods for cell type and cell state detection in healthy complex tissues

Recent improvements in single-cell technologies have led to large-scale single-cell profiling of cell type and cell state heterogeneity across a wide range of tissues. Much of the focus has been on the mammalian brain (primarily using mouse as a model) due to the large functional variability of this organ defined by a large number of transcriptional and epigenetic cell types present across all regions. The development of high throughput microfluidics, microwell and droplet based approaches (Fig 1.3A) for single-cell RNA sequencing (scRNA-seq, Han et al., 2018; Jaitin et al., 2014; Klein et al., 2015; Macosko et al., 2015; Rosenberg et al., 2018; Streets et al., 2014) have led the profiling of expression programs of cell types in the mouse cerebral cortex (Tasic et al., 2016;

A. Zeisel et al., 2015; Amit Zeisel et al., 2018), retina (Shekhar et al., 2016), hypothalamic arcuate nucleus (Campbell et al., 2017), entopeduncular nucleus (Wallace et al., 2017), amygdala (Y. E. Wu et al., 2017), and more recently across nine major regions of the murine brain (Saunders et al., 2018) and 19 regions across the mouse nervous system (Amit Zeisel et al., 2018). While single-cell RNA methods show the dominant expression programs occurring in different cell types, they do not inform on the chromatin accessibly changes and transcription factors playing roles in cell-type commitment. The assay for transposase-accessible chromatin sequencing can gauge chromatin architecture through probing the genome via hyperactive Tn5 transposases (Buenrostro et al., 2013a). Though single-cell ATAC sequencing (scATAC-seq) technologies have been developed for microfluidics, plate and droplet-based platforms (Buenrostro et al., 2013b, 2015; X. Chen et al., 2018; Mezger et al., 2018; Satpathy et al., 2019), the inherent use of the hyper active Tn5 transposase made this method easily adaptable for combinatorial indexing (Fig 1.3A), making it one of the most popular methods for single-cell ATAC-seq (D. A. Cusanovich et al., 2015). Since its inception, this method has been widely used across multiple species and tissues ranging from fly embryonic development (D. A. Cusanovich, Reddington, et al., 2018), to mouse (Preissl et al., 2018) to human cortex (Lake et al., 2017), to myogenesis (Pliner et al., 2018a), to hematopoietic differentiation (Buenrostro et al., 2018), to an atlas of 13 mouse tissues (D. A. Cusanovich, Hill, et al., 2018), to most recently mammary gland development (see section (Chung et al., 2019). While cell type specific chromatin architecture changes have been identified across multiple regions of the brain, the hippocampus, harboring a wide range of cells involved with memory formation, has not been profiled. We set out to create a comprehensive cell atlas for the murine hippocampus using sci-ATAC-seq (Sinnamon et al., 2019c).

1.2.3 Single state plasticity in breast cancer

Cell state heterogeneity adds an extra layer of complexity on top of genomic intra-tumor heterogeneity. Cells within tumors occupy cellular states with aberrant epigenetic architecture supporting transcriptomic changes and cellular signaling reminiscent of native states within their

organs and tissues of origin. This is often a result of mutations in genes that regulate focal and global chromatin architecture. Changes in DNA and histone modifying enzyme activity and CRC functionality have been observed across a wide range of cancers (Valencia & Kadoch, 2019). In acute myeloid leukemia (AML) global levels of hypomethylation have been observed with focal hypermethylation patterns of promoters and enhancers of important tumor suppressor genes, with ~7–10% of all patients harboring deletion or truncating mutations in *TET* genes and ~25% of AML cases showing *DNMT3A* mutations (Cancer & Atlas, 2013; Kulis & Esteller, 2010; Rasmussen & Helin, 2016; Liubin Yang et al., 2015). On the level of histone modifications, both hematological malignancies, breast and colorectal cancers show upregulation of several classes (I, II and IV) of HDACs (Audia & Campbell, 2016; Valencia & Kadoch, 2019; West & Johnstone, 2014). Similarly, misregulation of components of the Policomb gene complex tasked with transcriptional silencing via histone methylation (*e.g.* H3K27me3) and ubiquitination has also been observed in leukemia and breast cancer (Bachmann et al., 2006; Chittock et al., 2017; Score et al., 2012). CRCs accrue mutations as well, as more than 20% of all cancers show mutations in mSWI/SNF-encoding genes(Valencia & Kadoch, 2019). This shows the importance of understanding the role of epigenetic changes in tumorigenesis, cancer progression and treatment resistance.

In breast cancer, cells are often described along three axes of differentiation. Cells can range from stem-like to differentiated, basal-like to luminal, and epithelial to mesenchymal (Roy Z. Granit et al., 2014). These also relate to cell states that have classically been described in breast cancer based on their relatedness to undifferentiated and differentiated cell types present in the adult mammary gland. Mammary stem cells (MASC) are a bipotent cell type present throughout life, which during human mammary gland development give rise to two cell lineages: luminal progenitors and basal (myoepithelial) progenitors, which in turn can develop into differentiated basal and luminal cell types (Chung et al., 2019; Visvader & Stingl, 2014). A recent study showed this process to be strongly epigenetically regulated as single-cell ATAC sequencing on embryonic age 18 mouse fetal MASC cells exhibited poised basal like and luminal like populations, while

scRNA-seq could not distinguish between these two populations (Chung et al., 2019). Myoepithelial cells can further transition into mesenchymal cell states during later stages of development, such as puberty and pregnancy, where rapid changes occur to the mammary gland. Similarly, epithelial-to-mesenchymal transition (EMT) can occur in wound healing. Next to differentiation EMT is also a strongly epigenetically regulated process (Y. Wu et al., 2016). Known transcription factors can signal a push of cell populations to differentiate along a luminal lineage (ESR1, FOXA1, GATA3) or basal lineage (TP63, SLUG, EGR1), and can promote EMT (ZEB1, SNAIL, and TWIST) or mark the reverse process of mesenchymal-to-epithelial transition (Banyard & Bielenberg, 2015; Gascard et al., 2015b; Roy Z. Granit et al., 2014; Hardy et al., 2010; Lamouille et al., 2014; Micalizzi et al., 2010; Risom et al., 2018; Y. Wu et al., 2016).

Recently, nanogrid single nucleus RNA-seq of triple negative breast cancers showed tumor cells of the same tumors to occupy heterogenous combinations of primarily basal-like cells, alongside luminal-A, luminal-B, normal-like and HER2+ cells, indicating that tumor cells of different origin can plasticly shift between cell states (Gao et al., 2017). Similarly, patient-derived cell lines also show this diversity of basal, luminal, and stem-like cells, which, even after isolation of an individual phenotype, reconstitute the fractions of states required for normal growth (Gupta et al., 2011b). This indicates a plasticity in cell states that tumor cells can use to restore heterogeneity. This becomes very important for a tumor's survival as on top of Darwinian clonal selection tumors can exhibit transcriptional and epigenetic reprogramming within clones to repurpose existing regulatory developmental pathways to shift cell state equilibrium as a response to extrinsic stimuli (C. Kim et al., 2018b). This strategy is widely employed in breast cancers where, upon treatment, cells can shift into drug resistant persistor (DRP) populations for the duration of the treatment independent of their genomic background. This requires fast epigenetic re-modeling, which when closely studied resembles that of processes observed during differentiation and EMT (Hinohara & Polyak, 2019b). This has elicited combination treatments in breast cancers, where the epigenetic modeling is blocked next to targeted therapies. In a recent study of luminal ER+ breast

cancer, the authors tested whether KDM5B, a regulator of transcriptomic heterogeneity through H3K4me3 demethylation, affects resistance to the endocrine drug fulvestrant. Inhibition of KDM5 showed decreased transcriptional heterogeneity and increased sensitivity to the drug (Hinohara et al., 2018). Similarly, Risom et al., 2018 showed DTPs of basal like TNBC cells to respond favorably to the MEK inhibitor Trametinib and PI3K/mTOR BEZ235 inhibitor treatment in combination with JQ1 and inhibitor of the BET bromodomain chromatin remodeler proteins. Interestingly, this study also showed a shared response across genetically distinct basal cell lines towards a basoluminal cell state upon Trametinib treatment, potentially indicating a targetable chromatin state. While these studies have evaluated intra tumor heterogeneity on the level of transcriptomic and epigenetic heterogeneity separately, no comprehensive integrated analysis of modalities has been done on breast cancer DTP formation to date. This has primarily been due a lack of reliable cross modality single-cell integration methods.

1.2.4 Computational methods for dissecting heterogeneity in cell types and cell states

Single-cell RNA and ATAC sequencing provide valuable snapshots of the regulatory heterogeneity across studied tissues. They can inform us of cell types and cell states present in our data and can help us infer the ordering of regulatory changes in the case of development.

Due to the early development of reliable high-throughput scRNA-seq methods, this technology has been applied to a wide range of studies across a multitude of tissues. Tools development of computational methods have followed along (Fig 1.3B). Conventional pipelines follow the same overall order of analyses. Pre-processing and quality control steps first assign aligned duplicate removed high quality sequencing reads to cells based on a unique molecular identifier (UMIs) and create a high dimensional single-cell expression matrix. Low expression transcripts are generally removed at this point along with cells with a low number of transcripts. Then, the cell expression matrix is normalized so that technical differences in cell read depth and gene expression are corrected for (Luecken & Theis, 2019). Methods for this can be as simple as linear regression on counts per million in bins of cell read depth to non-linear methods correcting

for multiple sources of variation (Cole et al., 2019; Luecken & Theis, 2019). Corrected pseudo-count matrixes are then log-transformed to mitigate the mean-variance relationship of the data and reduce skewness. Additional sources of variation, such as mitochondrial reads bias, read dropout rate, batch and cell cycle effects can be corrected for afterwards. These log-transformed values can then be more easily interpreted for fold changes in downstream analyses (Luecken & Theis, 2019). At this point the reduction of dimensionality of the data begins first by selecting for highly variable genes across cells (~1,000-5,000 genes), followed by dimensionality reduction methods that project the data into as low a number of interpretable dimensions as possible while retaining its inherent structure (Heimberg et al., 2016; Luecken & Theis, 2019). These methods can be linear or non-linear, with Principal Component Analysis (PCA) being the most popular linear method (Pearson, 1901). At this point, data is projected into two or three dimensions for visualization purposes. While for the initial dimensionality reduction linear methods are preferred for easier interpretability, non-linear methods are used for visualization purposes to show greater separation between cell types (Luecken & Theis, 2019). Uniform Manifold Approximation and Projection method (UMAP), and t-distributed stochastic neighbor embedding (t-SNE) are the two most widely used methods (Becht et al., 2018a; Wattenberg et al., 2017). Downstream analyses depend on the biological question of the study, but often continue with single-cell clustering, cluster annotation, anddifferential expression, gene ontology, and gene regulatory network analysis (Luecken & Theis, 2019). For streamlined analysis, comprehensive platforms now exist which contain a collection of independently developed tools for all stages of analysis. Of the many existing platforms, the three most popular are Seurat and Scater developed for R and SCANPY for the python environment (Butler et al., 2018; McCarthy et al., 2017; Wolf et al., 2018).

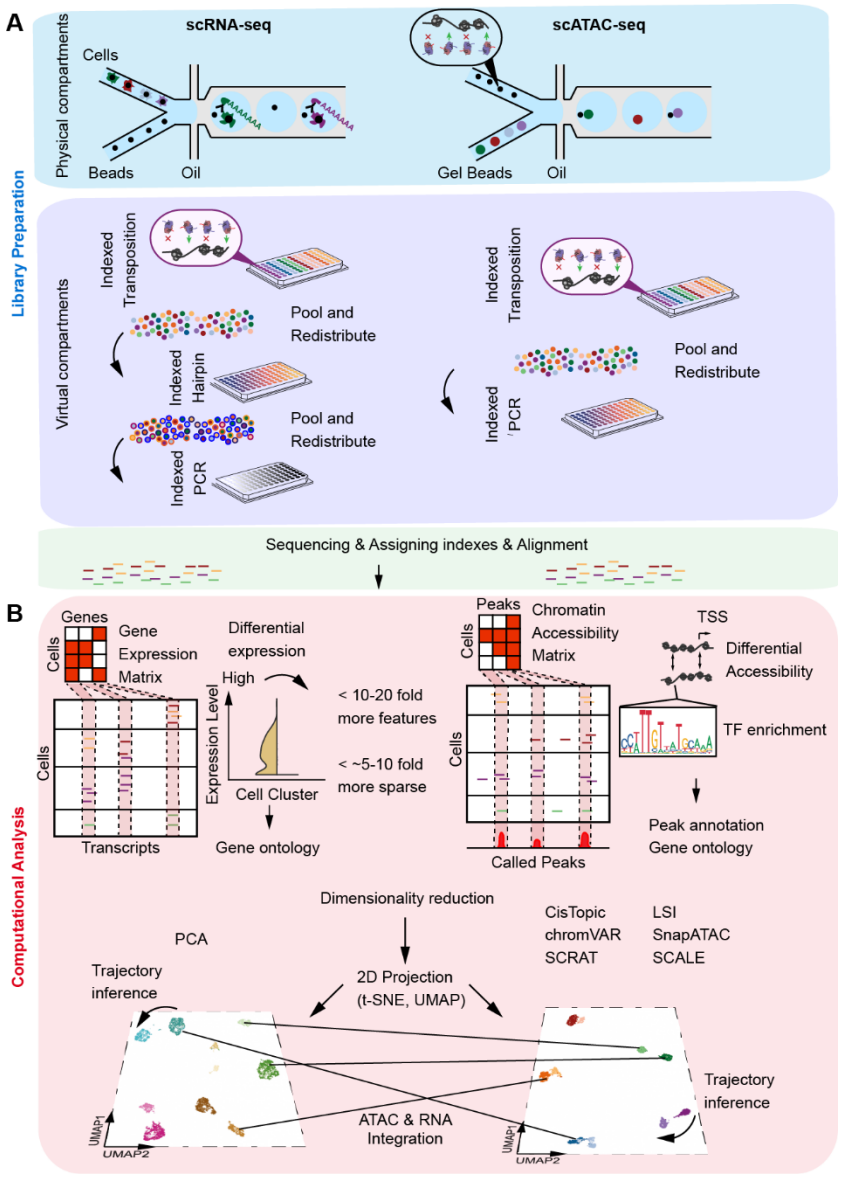

Figure 1.3 Methods for single-cell expression and chromatin accessibility. (A) Examples of droplet based and combinatorial indexed based scRNA-seq and scATAC-seq library preparation (B) Downstream analyses of the feature matrix and the information gained.

There are fewer analysis methods available for single-cell ATAC sequencing compared to scRNA-seq due to the relative novelty of the technique and the inherent challenges of this type of data. Analysis of data begins similarly to scRNA-seq by first demultiplexing, deduplicating and assigning aligned reads to single cells. Quality of reads can be assessed based on several features, one of which is the distribution of fragment sizes, which follows a decreasing trend with increasing fragment size that also shows periodic peaks corresponding to mono-, di- and tri- nucleosome

occupation. Regions that are nucleosome free are expected to be enriched at transcription start sites of genes (H. Chen et al., 2019; Yan et al., 2020). Following quality control, common peaks across cells are identified based on the read distribution relative to a random background (*e.g.* MACS2, Gaspar, 2018; Yong Zhang et al., 2008). Reads from cells in common peaks can be used to create the chromatin accessibility feature matrix, similar to the expression matrix. However, compared to scRNA-seq sequencing, single-cell ATAC sequencing can produce 10-20 fold more features due to the relative number of available enhancers (generally >50% of peaks) and promoters (~25% peaks) compared to transcripts, potentially allowing for better separation of cell states (Corces et al., 2016; Yan et al., 2020). This data is generally very sparse per cell due to the low copy number of DNA relative to RNA, resulting in only 1-10% of expected sites observed in a single cell relative to 10-45% expected genes detected per cell in scRNA-seq studies (H. Chen et al., 2019). Computational methods combat this inherent sparsity by finding or defining features that are most informative at explaining the variability between cells and identifying cell states. Methods such as chromVAR, SCRAT and Cicero aggregate reads in biologically linked regions, either by finding shared transcription factor accessibility (chromVAR) or by linking regulatory regions (SCRAT, Cicero) in the proximity of gene transcription start sites or through co-accessibility (Ji et al., 2017; Pliner et al., 2018b; Schep et al., 2017). Other methods use unsupervised statistical and machine learning methods to find co-regulated sites. Latent semantic indexing (LSI) was one of the first approaches developed, based on natural language processing. These methods first normalize the chromatin accessibility via a term frequency-inverse document frequency transformation (TF-IDF) and then reduce dimensionality using singular value decomposition (SVD). Later iterations of this method use an extra round of TF-IDF and SVD on peaks defined by clusters of the first round of LSI (D. A. Cusanovich, Reddington, et al., 2018; D. a Cusanovich et al., 2015). Similarly based in natural language processing, cisTopic applies latent Dirchlet allocation (LDA), a probabilistic topic-based modelling approach, which simultaneously infers a topic by cells matrix informing on cell states, and a topic by sites matrix that can help identify cis-regions specific to these topics

(Bravo González-Blas et al., 2019). Finally, SnapATAC uses a windowing approach across the genome followed by regression-based correction for library size and PCA. This method is the least computationally expensive and therefore very scalable (Fang et al., 2019). A recent study comparing 10 methods based on their ability to separate cell types in 13 synthetic and real datasets showed SnapATAC, LSI and CisTopic to outperform other compared methods (H. Chen et al., 2019). Scalability has become a central issue with larger single-cell ATAC-seq datasets, and new machine learning based approaches, such as the single-cell ATAC-seq analysis via latent feature extraction (SCALE) show promise (Xiong et al., 2019). An additional problem lies with linking defined important features to cell states and interpretability as peaks often lack annotation. Unsupervised methods (*e.g.* Cicero) can use co-accessibility between peaks across cells to define co-regulated chromatin hubs (Pliner et al., 2018b). Other methods, (*e.g.* Cistopic) inherently link sites within topics (Bravo González-Blas et al., 2019). Differential accessibility between cell states can also define cell state specific sites. These sites can be annotated by methods (*e.g.* HOMER, ChIPseeker) that use proximity of sites to genes or regulatory regions to establish linkage (Benner et al., 2017; G. Yu et al., 2015). Similarly, gene ontology of grouped sites can be approximated (Gu Z., 2019). Finally, single-cell ATAC-seq has the inherent property to inform on putative binding sites of transcription factors. The enrichment of TF binding sites in cell state specific sites relative to a background can inform on TFs specific to those states and chromVAR can provide information on the changes in global chromatin accessibility changes at defined TFs binding sites across our samples (Benner et al., 2017; Schep et al., 2017).

Single-cell studies capture a genetic snapshot across all cells, where each cell represents a sample from the distribution of biological processes occurring. Therefore, a continuous ordering of events can be reconstructed from the observed heterogeneity of cells. This is done via trajectory inference, where dynamical models of gene expression and chromatin accessibility are used to infer a path across cellular space based on cell to cell transitions (H. Chen et al., 2019; Luecken & Theis, 2019; Schier, 2020). After cells are ordered along a trajectory, pseudotime can be calculated relative

to a root cell. Pseudotime is often interpreted as a proxy for developmental or treatment response. Depending on software, trajectories can range in intricacy from linear to bifurcating paths and loops. A wide range of software exists for trajectory inference and the best choice for analysis often depends on the biological problem and the method used for assaying. Including time points can often help with finding biologically relevant trajectories (Luecken & Theis, 2019). One of the first software designed for inferring trajectories was Monocle (Qiu et al., 2017), which has since been expanded from only doing ordering on single-cell RNA data to incorporating single-cell ATAC data as well based on Cicero (Pliner et al., 2018b; Yan et al., 2020).

Single-cell RNA and ATAC sequencing can show diverse, often complementary aspects of a biological question. Recent years have seen the development of co-assays that can shed light on chromatin organizational and transcriptional changes in parallel. These assays are limited, however, in the amount of information that can be garnered from a single-cell. This results in each modality of data often having lower quality compared to individual assays performed in two separate experiments (Zhu, 2020). This exchange of co-temporal data acquisition with data quality has created a need for computational methods to establish cross assay integration between individual experiments. These methods can bridge experimental and assay differences while reconciling heterogeneity across modalities and therefore helping with interpretability. This is achieved by projecting data from different assays into a common latent space from which missing data in each of the modalities can be predicted for via inference learning. The most common inference strategies used currently rely on some form of canonical correlation analysis, nonnegative matrix factorization, or variational autoencoder (Efremova & Teichmann, 2020). Two recently published methods showed scATAC-seq and scRNA-seq integration, with the latter used as a reference (Stuart et al., 2018; Welch et al., 2019). These methods offer a great opportunity to infer gene regulatory networks and shared trajectories to understand tumorigenesis and development (Welch et al., 2017).

1.3 Summary

Genetic, epigenetic, and transcriptomic heterogeneity play a significant role in tissue functional maintenance and robustness to external stimuli. Developmental processes give rise to cell types and states, in addition to tissues accumulating somatic mutations. The latter give rise to evolutionary advantages in dividing cell populations potentially resulting in selection for somatic mutations at cancer associated genes. Similarly, the natural processes of cell type maintenance can mutate, resulting in aberrant forms of their original functions. This results in the eventual shift of the genetic and epigenetic regulatory equilibrium within an individual cell and the formation of cancer. Both epigenomic and genomic processes result in intra tumor heterogeneity, which can aid tumors in evading therapeutic pressure via Darwinian selection and cell state plasticity. Single-cell technologies (discussed above) serve as a platform to accurately profile the genomic and epigenomic heterogeneity in healthy tissues and their diseased counterparts. In this dissertation I show examples of these processes by (i) profiling the genomes of thousands of cells in healthy and diseased tissues (Chapter 2), (ii) mapping the chromatin landscape of the murine hippocampus (Chapter 3), and (iii) looking at the development of Trametinib resistance through cell state plasticity across basal-like triple negative breast cancer cell lines (Chapter 4).

Chapter 2: Sequencing thousands of single-cell genomes with combinatorial indexing

Sarah A. Vitak*, **Kristof A. Torkenczy***, Jimi L. Rosenkrantz, Andrew J. Fields, Lena Christiansen, Melissa H. Wong, Lucia Carbone, Frank J. Steemers, and Andrew Adey

* These authors contributed equally to this work.

This chapter has been reformatted for inclusion for this dissertation from the manuscript titled: "Sequencing thousands of single-cell genomes with combinatorial indexing" published in Nature Methods, March 2017. A. A. designed and supervised the study. I processed all sequence data and designed and wrote the copy number analysis pipeline and connected analyses. S.A.V. carried out all SCI-seq and GM12878 DOP library preparations, designed experiments, and performed all sequencing. A.A., S.A.V., and I wrote the manuscript. All authors contributed and edited the manuscript. J.L.R. constructed QRP and DOP libraries on Rhesus samples. A.J.F. prepared all GM12878 QRP library construction and co-prepared all SCI-seq libraries using xSDS for nucleosome depletion. M.H.W. provided tumor samples and aided in the analyses of those samples. L. Carbone supervised and provided all samples for Rhesus work. F.J.S. contributed to experimental design and contributed to the manuscript, L. Christiansen produced all transposase complexes used in this study.

All supplemental figures and tables are not included in this dissertation and should be referred to at https://www.ncbi.nlm.nih.gov/pubmed/28135258.

2.1 Abstract

Single-cell genome sequencing has proven to be a valuable tool for the detection of somatic variation, particularly in the context of tumor evolution. Current technologies suffer from high per-cell library construction costs which restrict the number of cells that can be assessed, thus imposing limitations on the ability to quantitatively measure genomic heterogeneity within a tissue. Here, we present Single-cell Combinatorial Indexed Sequencing (SCI-seq) as a means of simultaneously generating thousands of low-pass single-cell libraries for somatic copy number variant detection. In total, we constructed libraries for 16,698 single cells from a combination of cultured cell lines, primate frontal cortex tissue, and two human adenocarcinomas, including a detailed assessment of subclonal variation within a pancreatic tumor. This novel technology facilitates low-cost, deep characterization of somatic copy number variation in single cells, providing a foundational knowledge across both healthy and diseased tissues.

2.2 Introduction

The booming field of single-cell sequencing continues to shine light on the abundance and breadth of genomic heterogeneity between cells in a variety of contexts, including somatic aneuploidy in the mammalian brain (Cai et al., 2014; Knouse et al., 2014; McConnell et al., 2013; Rehen et al., 2001), and intra-tumor heterogeneity (Eirew et al., 2015; Gao et al., 2016; Gawad et al., 2014; N. Navin et al., 2011). These studies have taken one of two approaches: high depth of sequencing per cell for single nucleotide variant detection (Cai et al., 2014; Zong et al., 2012) , or low-pass sequencing to identify copy number variants (CNVs) and aneuploidy (Baslan et al., 2012; Knouse et al., 2016; McConnell et al., 2013). In the latter approach, the lack of an efficient, cost-effective method to produce high numbers of single-cell libraries has prevented the ability to quantitatively measure the frequency of CNV-harboring cells at population-level scale, or provide a robust analysis of heterogeneity in the context of cancer (Gawad et al., 2016).

Recently, we established a method to produce thousands of individually barcoded libraries of linked sequence reads using a transposase-based combinatorial indexing strategy (CPT-seq) (Adey et al., 2010, 2014; Amini et al., 2014a) which we applied to haplotype resolution (Amini et al., 2014a) and *de novo* genome assembly (Adey et al., 2014). This concept was then integrated with the chromatin accessibility assay, ATAC-seq (Buenrostro et al., 2013b), to produce profiles of active regulatory elements in thousands of single cells (Cusanovich et al., 2015) (sci-ATAC-seq, Figure 2.1A). In this method, nuclei are first barcoded by the incorporation of one of 96 indexed sequencing adaptors via transposase. The 96 reactions are then combined and 15-25 of these randomly indexed nuclei are deposited into each well of a PCR plate by Fluorescence Activated Nuclei Sorting (FANS, Supplementary Figure 1). The probability of any two nuclei having the same transposase barcode is therefore low (6-11%) (Cusanovich et al., 2015). Each PCR well is then uniquely barcoded using indexed primers. At the end of this process, each sequence read

contains two indexes: Index 1 from the transposase plate, and Index 2 from the PCR plate, which facilitate single-cell discrimination. As proof of principle, Cusanovich and colleagues produced over 15,000 sci-ATAC-seq profiles and used them to separate a mix of two cell types by their accessible chromatin landscapes (Cusanovich et al., 2015). We reasoned that a similar combinatorial indexing strategy could be extended to single-cell whole genome sequencing.

2.3 Results

2.3.1 Nucleosome depletion for uniform genome coverage

The key hurdle to adapt combinatorial indexing to produce uniformly distributed sequence reads is the removal of nucleosomes bound to genomic DNA without compromising nuclear integrity. The sci-ATAC-seq method is carried out on native chromatin, which permits the conversion of DNA into library molecules only within regions of open chromatin (1-4% of the genome). This restriction is desirable for epigenetic characterization; however, for CNV detection, it results in biological bias and severely limited read counts (~3,000 per cell) (Cusanovich et al., 2015). We therefore developed two strategies to unbind nucleosomes from genomic DNA while retaining nuclear integrity for SCI-seq library construction. The first, Lithium Assisted Nucleosome Depletion (LAND), utilizes the chaotropic agent, Lithium diiodosalycylate, to disrupt DNA-protein interactions in the cell, therefore releasing DNA from histones. The second, crosslinking with SDS treatment (xSDS), uses the detergent, SDS, to denature histone proteins and render them unable to bind DNA. However, SDS has a disruptive effect on nuclear integrity, thus necessitating a crosslinking step prior to denaturation in order to maintain intact nuclei.

To test the viability of these strategies, we performed bulk (30,000 nuclei) preparations on the HeLa S3 cell line, for which chromatin accessibility and genome structure has been extensively profiled (Adey et al., 2013; Dunham et al., 2012), and carried out LAND or xSDS treatments along with a standard control. In all three cases nuclei remained intact – a key requirement for the SCI-seq workflow (Figure 2.1B). Prepared nuclei were then carried through standard ATAC-seq library

construction (Buenrostro et al., 2013b). The library prepared from untreated nuclei produced the expected ATAC-seq signal with a 10.8 fold enrichment of sequence reads aligning to annotated HeLa S3 accessibility sites. Both the LAND and xSDS preparations had substantially lower enrichments of 2.8 and 2.2 fold respectively, close to the 1.4 fold observed for shotgun sequencing (Figure 2.1C, Supplementary Table 1). Furthermore, the projected number of unique sequence reads present in the LAND and xSDS preparations were 1.7 billion and 798 million respectively, much greater than for the standard library at 170 million, suggesting a larger proportion of the genome was converted into viable sequencing molecules.

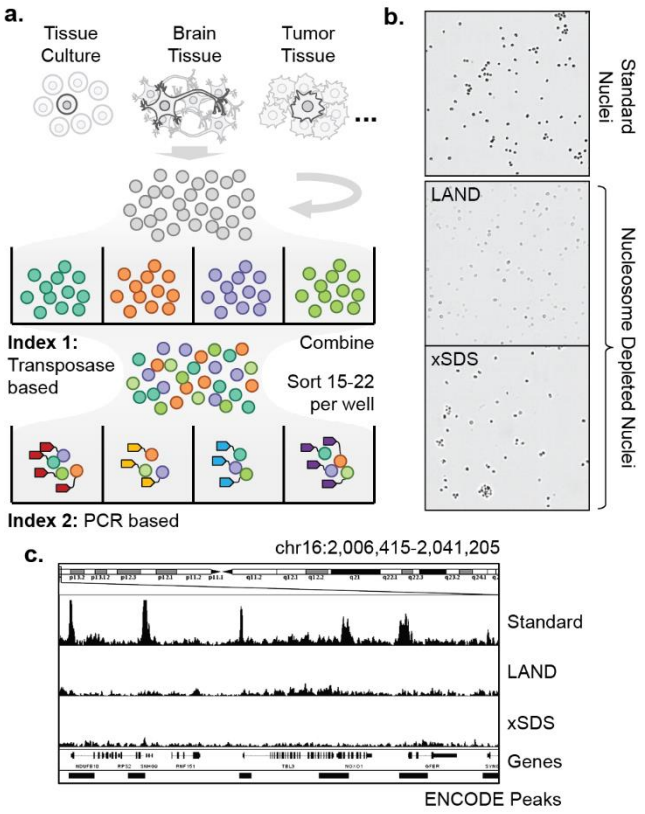

Figure 2.1 Single-cell combinatorial indexing with nucleosome depletion. (**A**) Single-cell combinatorial indexing workflow. (**B**) Standard isolated nuclei and nucleosome depleted nuclei using Lithium Assisted Nucleosome Depletion (LAND) or by crosslinking and SDS treatment (xSDS) produce intact nuclei. (**C**) Nucleosome depletion produces genome-wide uniform coverage that is not restricted to sites of chromatin accessibility.

2.3.2 SCI-seq with nucleosome depletion

To assess the performance of nucleosome depletion with our single-cell combinatorial

indexing workflow, we first focused on the deeply profiled, euploid lymphoblastoid cell line GM12878 (Adey et al., 2014; Amini et al., 2014a; Dunham et al., 2012). We produced a total of six SCI-seq libraries with a variety of LAND conditions, each using a single 96-well plate at the PCR indexing stage, and a single xSDS library with 3×96-well PCR plates. To serve as a comparison to existing methods, we prepared 42 single-cell libraries using quasi-random priming (QRP, 40 passing QC) and 51 using degenerate oligonucleotide primed PCR (DOP, 45 passing QC). Finally, we karyotyped 50 cells to serve as a non-sequencing means of aneuploidy measurement (Supplementary Table 2).

For each SCI-seq preparation, the number of potential index combinations is 96 (transposase indexing) × N (PCR indexing, 96 per plate); however, not all index combinations represent a single-cell library, as each PCR well contains only 15-25 transposase-indexed nuclei. To identify non-empty index combinations, we generated a $\log_{10}$ transformed histogram of unique (*i.e.* non-PCR duplicate), high-quality (MQ ≥ 10) aligned reads for each potential index combination. This resulted in a bimodal distribution comprised of a low-read-count, noise component centered between 50 and 200 reads, and a high-read-count, single-cell component centered between 10,000 and 100,000 reads (Fig 2.2A,B, right; Supplementary Figure 2, Supplementary Software). We then used a mixed model to identify indexes that fall in this high-read-count component (Supplementary Figure 3), which resulted in 4,643 single-cell libraries across the six SCI-seq preparations that used LAND for nucleosome depletion and 3,123 for the xSDS preparation.

To confirm that the majority of putative single-cell libraries contain true single cells, we carried out four SCI-seq library preparations on a mix of human and mouse cells using LAND (2,369 total cells) with either 22 or 25 nuclei per PCR well, and one preparation using xSDS split between two FANS conditions (1,367 total cells). For each experiment we analyzed the proportion of putative single cells with ≥ 90% of their reads that aligned exclusively to the human or mouse genome. The remaining cells represent human-mouse collisions (*i.e.* doublets) and make up approximately half of the total collision rate (the remaining half being human-human or mouse-

mouse). The total collision rates varied between 0-23.6%, and were used to decide upon 22 nuclei per well with restrictive sorting conditions for a target doublet frequency of <10%, comparable to sci-ATAC-seq (Cusanovich et al., 2015) or high throughout single-cell RNA-seq technologies (Macosko et al., 2015).

The unique read count produced for each single-cell library in a SCI-seq preparation is a function of library complexity (unique molecules in library) and sequencing depth. Due to the inhibitive cost of deeply sequencing every preparation during development, we implemented a model to project the anticipated read count and PCR duplicate percentage that would be achieved with increased sequencing depth (Figure 2.2C, Methods). As a means of quality assessment, we identified the depth in which a median of 50% of reads across cells are PCR duplicates (M50), which represents the point in which diminishing returns of additional sequencing become excessive (*i.e.* greater than 50% of additional reads provide no new information), along with several other metrics (Supplementary Table 3). To evaluate our projections, we built a model on a subset of the sequenced reads and compared the projected metrics with those from the actual depth. This analysis showed our model accurately predicted the median unique read count within a median of 0.02% (maximum 2.25%, mean 0.41%) across all libraries. We further tested our projections by selecting a subset of PCR wells from several preparations and performed additional sequencing which produced unique reads counts for each cell that were within a median of 0.13% (maximum 3.56%, mean 0.72%) of what was predicted by our model (Supplementary Figure 5).

Coverage uniformity was assessed using two previously described metrics – mean absolute deviation (MAD) (Garvin et al., 2015), and mean absolute pairwise deviation (MAPD) (Cai et al., 2014), which indicated substantially increased uniformity using the xSDS strategy over LAND (MAD: mean 1.57-fold improvement, $p = <1x10^{-15}$; MAPD: 1.70-fold improvement, $p = <1x10^{-15}$, Welch's t-test); however, the deviation of the xSDS preparation is still greater than for QRP and DOP methods, though similar to multiple displacement amplification methods (Figure 2.2D). While LAND preparations had an increased coverage bias, the method produced higher unique

read counts per cell (*e.g.* M50 of 763,813 for one of three HeLa LAND preparations) when compared to xSDS (*e.g.* M50 of 63,223 for the GM12878 preparation). For all libraries, we observed the characteristic 9 basepair overlap of adjacent read pairs due to the mechanism of transposition(Adey et al., 2010; Goryshin et al., 1998), indicating we are able to sequence molecules on either side of a transposase insertion event (Supplementary Figure 6).

2.3.3 Copy number variant calling using SCI-seq

For any single-cell genome sequencing study, determining how to filter out cells that may have failed during library construction at the risk of removing true aneuploid cells is a significant challenge. Therefore, we initially proceeded with CNV calling assessment on our SCI-seq preparation without any filtering and compared them to the QRP and DOP methods. For all preparations, we proceeded with only cells for which a minimum of 50,000 unique, high quality aligned reads were sequenced (868 across all LAND libraries, 1,056 for the xSDS library). We then used Ginkgo (Garvin et al., 2015), Circular Binary Segmentation (CBS) (Olshen et al., 2004b), and a Hidden Markov Model (HMM) (Ha et al., 2012), with variable-sized genomic windows (target median of 2.5 million bp) for CNV calling (Supplementary Figure 7) and conservatively retained the intersection of all three methods. To compare our sequencing-based calls with karyotyped cells, we focused on chromosome-arm level events (Figure 2.2E,F). Consistent with the coverage uniformity differences, our LAND SCI-seq preparations produced a high aneuploidy rate (61.9%), suggesting an abundance of false positives due to lack of coverage uniformity (Figure 2.2E,G). However, the xSDS nucleosome depletion strategy with SCI-seq resulted in an aneuploidy frequency of 22.6%, much closer to the karyotyping results (Figure 2.2E,H), and the two standard methods of single-cell sequencing at 15.0% and 13.5% for DOP and QRP respectively (Supplementary Figure 8).

We next explored the range of resolution that can be achieved using SCI-seq, and determined filtering criteria based on MAD and MAPD scores across a variety of resolutions and

read count thresholds (Supplementary Figure 9). This analysis revealed a greater range of variability in our SCI-seq preparations that is largely driven by the wider range of unique reads per cell when compared to standard methods. We then applied a MAD filter across all methods of 0.2 and recalculated the aneuploidy rate. Post variance filtering, the aneuploidy rates for xSDS, DOP, and QRP were 12.2%, 9.7%, and 10.5% respectively, all below the rate determined by karyotyping, yet closer to one another prior to filtering (**Supplementary Figure 10**).

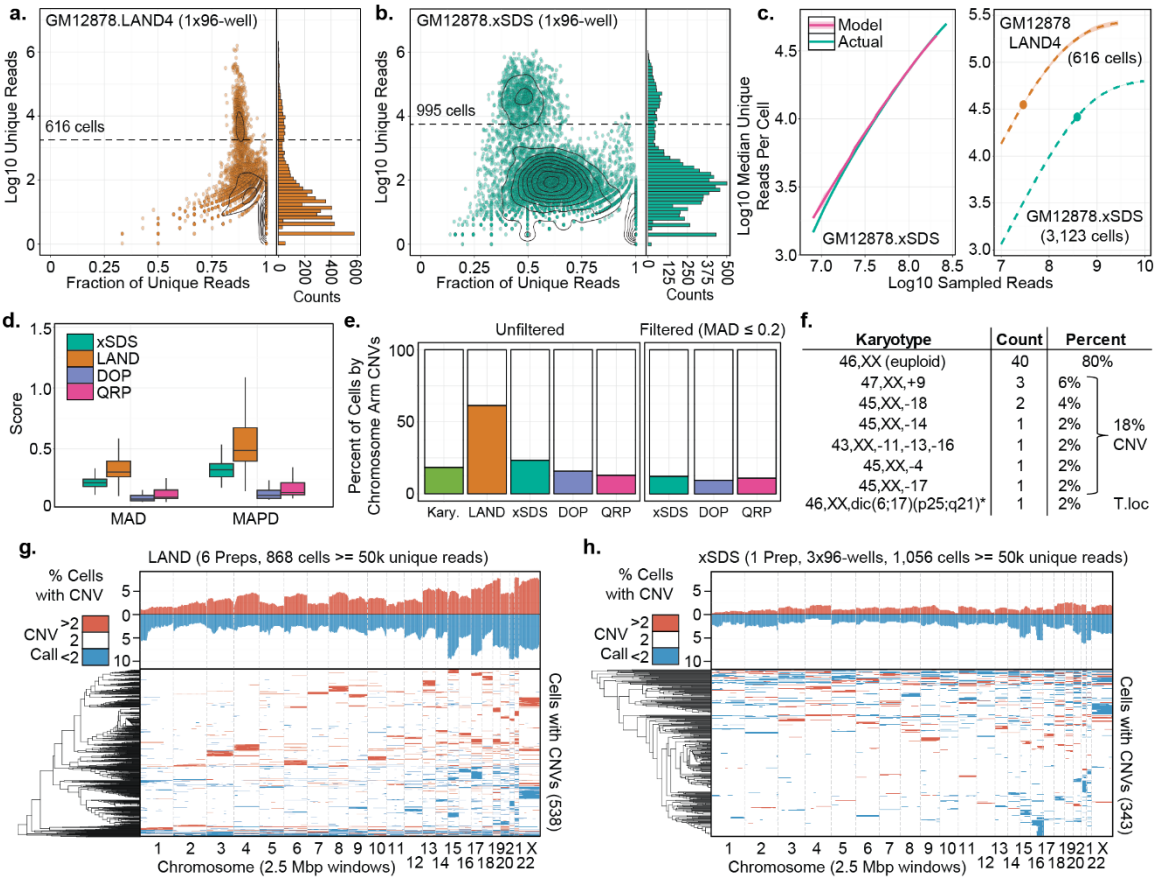

Figure 2.2 Comparison of LAND and xSDS nucleosome depletion methods with SCI-seq. (**a**) Log10 unique read count (y-axis) and histogram (right panel), by fraction of unique reads (x-axis) to indicate complexity for one of six LAND SCI-seq preparations on GM12878. Dashed line represents single-cell read cutoff. (**b**) As in (**a**) but for xSDS nucleosome depletion for one of three PCR plates. (**c**) Left, model built on downsampled reads for the GM12878 xSDS preparation and used to predict the full depth of coverage. Right, Projections for one of the LAND preparations and the full xSDS preparation. Points represent actual depth of sequencing. (**d**) Coverage uniformity scores for SCI-seq using LAND or xSDS and for quasi-random priming (QRP) and degenerate oligonucleotide PCR (DOP). (**e**) Summary of the percentage of cells showing aneuploidy at the chromosome arm level across all preparations with and without imposing a variance filter. (**f**) Karyotyping results of 50 GM12878 cells. (**g-h**) Summary of windowed copy number calls and

clustering of GM12878 single cells produced using LAND (**g**) or xSDS (**h**). Top represents a chromosome-arm scale summary of gain or loss frequency for all cells, bottom is the clustered profile for cells that contain at least one CNV call.

2.3.4 Copy number variation in the Rhesus brain

Estimates of the frequency of aneuploidy and large-scale CNVs in the mammalian brain have varied widely from <5% to 33%(Cai et al., 2014; Knouse et al., 2014; McConnell et al., 2013; Rehen et al., 2001). This uncertainty largely stems from the inability to profile sufficient numbers of single cells to produce quantitative measurements. The Rhesus macaque is an ideal model for quantifying the abundance of aneuploidy in the brain, as human samples are challenging to acquire and are confounded by a high variability of lifetime environmental exposures. Furthermore, the Rhesus brain is phylogenetically, structurally and physiologically more similar to humans than rodents (Rosenkrantz & Carbone, 2017).

We applied SCI-seq to archived frontal cortex tissue, to demonstrate the versatility of our platform, by performing both LAND and xSDS SCI-seq methods along with 38 cells using QRP (35 passing QC), and 35 cells using DOP (30 passing QC). All samples were from adjacent ~150 mm^3 sections of frontal cortex (Individual 1). Our low-capacity LAND preparation (16 PCR indexes) produced 340 single-cell libraries with a median unique read count of 141,449 (248 cells ≥ 50,000 unique reads), and our xSDS preparation generated 171 single-cell libraries with a median unique read count of 55,142 (92 cells ≥ 50,000 unique reads). The number of cells produced in our xSDS preparation was lower than expected, largely due to nuclei aggregates during sorting that may be remedied by additional cell dis-aggregations steps.

Across all methods of library construction we observed greater discrepancies between the three CNV calling approaches than in the human analyses (**Supplementary Figure 11-14**). We believe that this variability is due to the lower quality of the Rhesus reference genome (284,705 contigs < 1 Mbp) when compared to human, emphasizing the need for "platinum" quality reference genomes (Callaway, 2014). We therefore focused on the HMM results for sub-chromosomal calls

(**Figure 2.3A**) and performed aneuploidy analysis using the intersection of CBS and HMM calls. Consistent with our cell line results, the LAND preparation produced a much higher aneuploidy rate (95.1%), suggestive of false positives stemming from coverage nonuniformity (**Supplementary Figure 15,16**). The xSDS SCI-seq unfiltered aneuploidy rate (25.0%) was close to the DOP preparation (18.5%), with QRP producing a much lower rate (3.1%; **Figure 2.3B**). After imposing a variance filter for cells with a MAD score of 0.2 or lower, the aneuploidy rates dropped to 12.0% for the xSDS preparation, 8.7% for the DOP, and stayed the same for the QRP preparation at 3.1%. These rates were similar to those produced by xSDS SCI-seq on a 200 mm^3 section of frontal cortex from a second individual (381 single-cells, median read count of 62,731, 213 cells ≥ 50,000 unique reads) which produced unfiltered and filtered aneuploidy rates of 12.1% and 10.3% respectively (**Supplementary Figure 17**).

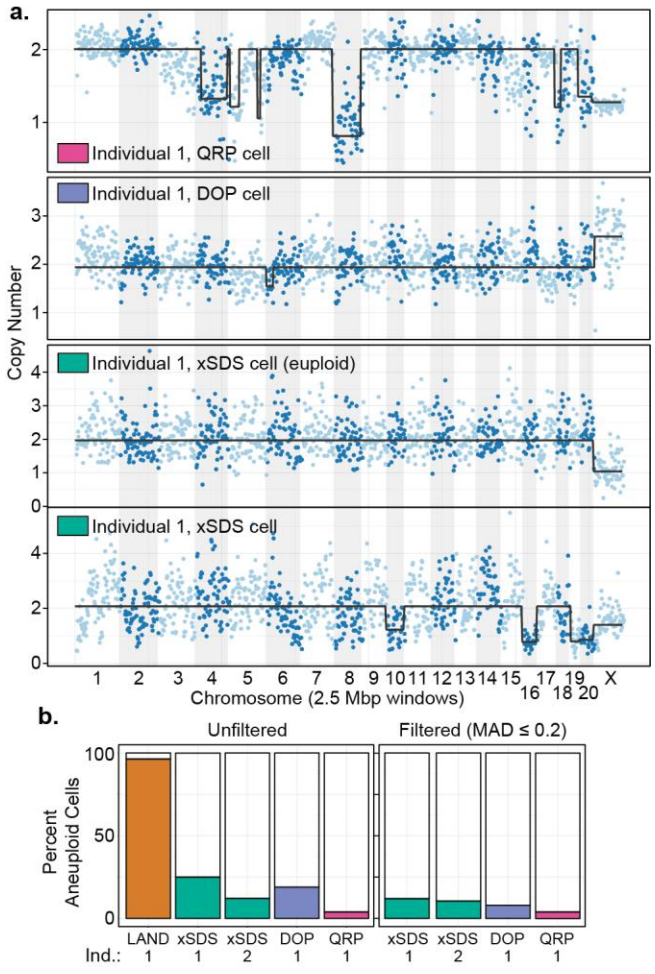

Figure 2.3 Somatic CNVs in the Rhesus brain. (**a**) Example single cells with copy number variants, and one representative euploid cell for the SCI-seq preparation (HMM). (**b**) Frequency of aneuploidy as determined by each of the methods with and without filtering.

2.3.5 SCI-seq on primary tumor samples reveals clonal populations

One of the primary applications of single-cell genome sequencing is in the profiling of tumor heterogeneity and understanding clonal evolution in cancer as it relates to treatment resistance (Eirew et al., 2015; Gao et al., 2016; Gawad et al., 2014; N. Navin et al., 2011). We carried out a single xSDS SCI-seq preparation on a freshly acquired stage III pancreatic ductal adenocarcinoma (PDAC) sample measuring approximately 250 mm^3 which resulted in 1,715 single-cell libraries sequenced to a median unique read count of 49,272 per cell (M50 of 71,378; 846 cells ≥ 50,000 unique reads at the depth the library was sequenced; **Figure 2.4A**). We first carried out CNV calling using our GM12878 library as a euploid baseline for comparison to identify a set of high-confidence euploid cells (298, 35.2%) which were then used as a new baseline specific to the individual and preparation (**Supplementary Figure 17-19**). We next assumed that subchromosomal copy number alterations caused by genome instability are more informative for the identification of subclonal populations than whole chromosome aneuploidy due to errors during cell division. We therefore developed a strategy to identify putative copy number breakpoints at low resolution to be used as new window boundaries (Methods, **Supplementary Figure 20**) followed by stratification via principle components analysis (PCA) and k-means clustering. We initially applied this method to our HeLa libraries (2,361 single cells in total), revealing no distinct heterogeneity and further supporting the stability of the HeLa cell line (Adey et al., 2013) (**Supplementary Figure 21-24**), and then on our primary PDAC sample, which revealed an optimum cluster count of 4 by silhouette analysis (**Figure 4B,C**).

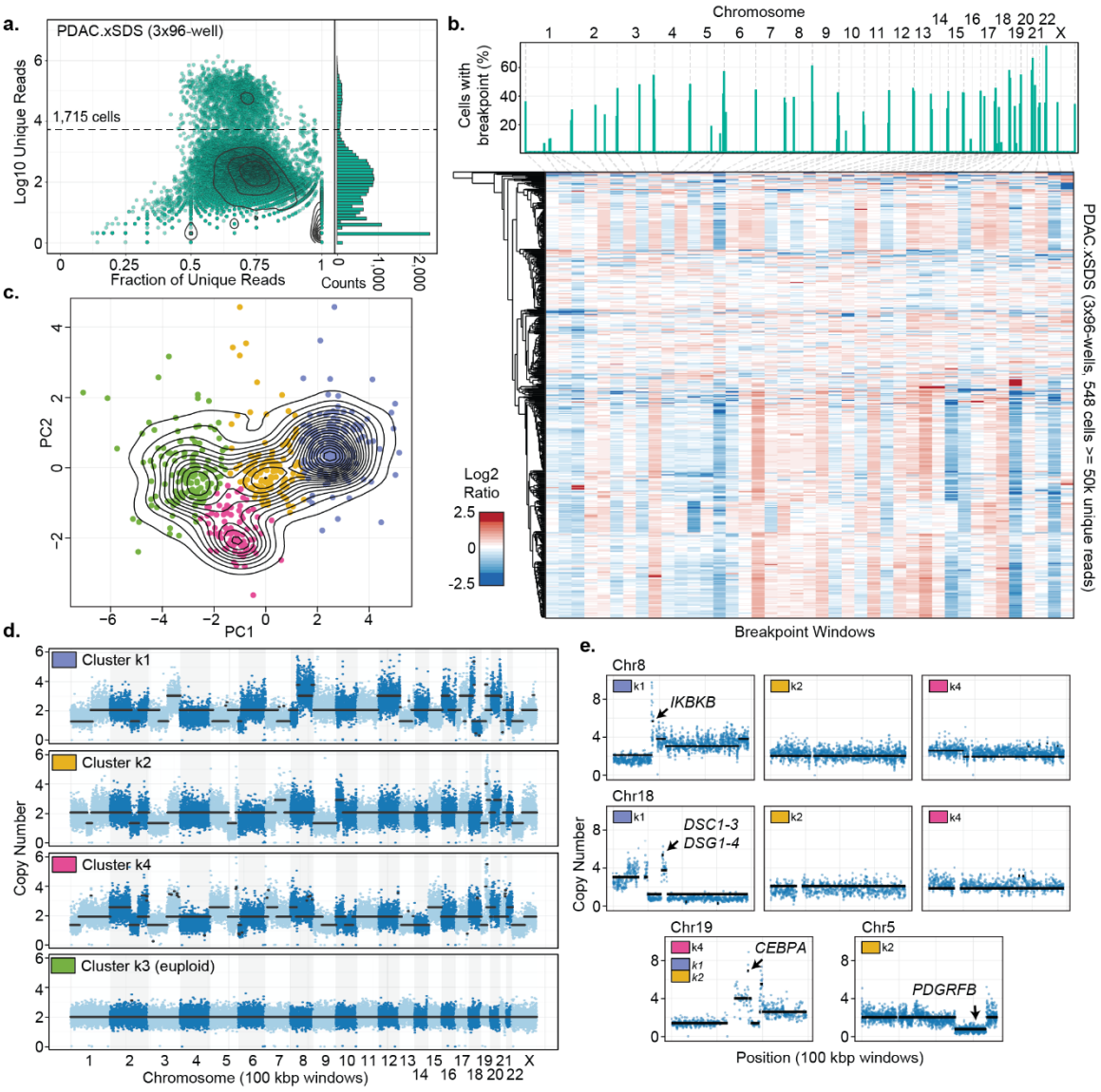

Figure 2.4 SCI-seq analysis of a stage III human Pancreatic Ductal Adenocarcinoma (PDAC). (**a**) SCI-seq library complexity as in Figure 1a. (**b**) Breakpoint calls (top) and breakpoint window matrix of log2 sequence depth ratio. (**c**) Principle component analysis and k-means clustering on breakpoint matrix. (**d**) 100 kbp resolution CNV calling on aggregated cells from each cluster. (**e**) Cluster specific CNVs and *CEBPA* amplification present in all clusters (k4 shown).

The first of these clusters (k3, green) is a population of euploid cells that were not considered high confidence euploid in the initial analysis, and thus not removed. When including these, the euploid population rises to 389 for a final tumor cell purity of 46.0%, within the expected range for PDAC (Waddell et al., 2015). For the remaining three clusters, k1 (purple, 199 cells), k2 (gold, 115 cells), and k4 (pink, 91 cells), we aggregated all reads from cells proximal to each centroid

(Methods) and carried out CNV calling using 100 kbp windows, a 25-fold greater resolution than the initial analysis, and then determined absolute copy number states (Adey et al., 2013) (Figure 2.4D). Across the three tumor clusters, a substantial portion of copy number segments were shared (44.8%), suggesting that they arose from a common progenitor population. This includes a highly rearranged chromosome 19 which harbors a focal amplification of *CEBPA*, which encodes an enhancer binding protein, at copy number 7 which is frequently mutated in AML (De Kouchkovsky & Abdul-Hay, 2016), and has recently been shown to have altered epigenetic regulation in pancreatic tumors (Kumagai et al., 2009) (Figure 2.4E). An all-by-all pairwise comparison revealed clusters k2 and k4 as the most similar, sharing 65.9% of copy number segments, followed by k1 and k4 at 58.3%, and k1 and k2 at 55.0%. We then assessed cluster-specific CNVs and discovered several that contain genes of potential functional relevance (Figure 2.4E). These include a focal amplification to copy number 6 of *IKBKB* in cluster k1, which encodes a serine kinase important in the NF-κB signaling pathway (Perkins, 2007); another focal amplification to copy number 5 in cluster k1 containing genes *DSC1,2,3* and *DSG1,2,3,4* all of which encode proteins involved in cell-cell adhesion and cell positioning and are often mis-regulated in cancer (Stahley & Kowalczyk, 2015); and the deletion of a region containing *PDGRFB* specific to cluster k2, which encodes a tyrosine kinase cell surface receptor involved in cell proliferation signaling, and is frequently mutated in cancer (Forbes et al., 2015).

Lastly, we applied xSDS SCI-seq to a frozen stage II rectal adenocarcinoma measuring 500 mm^3. During preparation we noticed a high abundance of nuclear debris and ruptured nuclei which likely attributed to the decreased yield of the preparation (16 PCR indexes) of 146 single-cell libraries (median unique read count of 71,378; M50 of 352,168; 111 cells ≥ 50,000 unique reads). We then carried out the same CNV calling approach as with the PDAC sample; however clear subpopulations or high frequency breakpoints were not observed, and therefore subclonal populations could not be identified (Supplementary Figure 25). This may be a result of nuclear deterioration due to irradiation, a common treatment for rectal cancers, underscoring the challenge

of producing high-quality single-cell or nuclei suspensions shared by all single-cell methods (Gawad et al., 2016).

2.4 Discussion

We have developed a novel approach, SCI-seq, which utilizes nucleosome depletion in a combinatorial indexing workflow to produce thousands of single-cell genome sequencing libraries. In total we produced 16,698 single-cell libraries (of which 5,395 were sequenced to a depth sufficient for CNV calling) from myriad samples using SCI-seq, including primary tissue isolates representative of the two major areas of single-cell genome research: somatic aneuploidy, and cancer. In addition to the advantages of throughput, the platform does not require specialized microfluidics equipment or droplet emulsification techniques. Using our more uniform nucleosome depletion strategy, xSDS, we were able to achieve resolution on the order of 250 kbp, though we suspect further optimization, such as alternative crosslinking agents, may provide sufficient depth for improved resolution. We also demonstrate the ability to identify clonal populations that can be aggregated to facilitate high resolution CNV calling by applying this strategy to a pancreatic ductal adenocarcinoma which revealed subclone-specific CNVs that may impact proliferation, migration, or possibly drive other molecular subtypes (P. Bailey et al., 2016).

While the technology is currently limited to copy number variant detection, it may be possible to include *in situ* pre-amplification within the nuclear scaffold prior to SCI-seq or the incorporation of T4 *in vitro* transcription, such as in THS-seq (Sos et al., 2016), an ATAC-seq variant, to boost the resulting coverage and facilitate single nucleotide variant detection. While optimization is possible, as with any new method, we believe that the throughput provided by SCI-seq will open the door to deep quantification of mammalian somatic genome stability as well as serve as a platform to assess other properties of single cells including DNA methylation and chromatin architecture.

2.5 Methods

2.5.1 Sample preparation and nuclei isolation.

Tissue culture cell lines were trypsinized then pelleted if adherent (HeLa S3, ATCC CCL-2.2; NIH/3T3, ATCC CRL-1658) or pelleted if grown in suspension (GM12878, Coriell; karyotyped at the OHSU Research Cytogenetics Laboratory), followed by one wash with ice cold PBS. They were then carried through crosslinking (for the xSDS method) or directly into nuclei preparation using Nuclei Isolation Buffer (NIB, 10 mM TrisHCl pH7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% igepal, 1x protease inhibitors (Roche, Cat. 11873580001)) with or without nucleosome depletion. Tissue samples (RhesusFcx1, RhesusFcx2, PDAC, CRC) were dounce homogenized in NIB then passed through a 35µm cell strainer prior to nucleosome depletion. The frozen Rhesus frontal cortex samples, RhesusFcx1 (4 yr. female) and RhesusFcx2 (9 yr. female), were obtained from the Oregon National Primate Research Center as a part of their aging nonhuman primate resource.

2.5.2 Standard Single-cell Library Construction

Single-cell libraries constructed using quasi-random priming (QRP) and degenerate oligonucleotide primed PCR (DOP) were prepared from isolated nuclei without nucleosome depletion and brought up to 1 mL of NIB, stained with 5 µL of 5 mg/ml DAPI (Thermo Fisher, Cat. D1306) then FANS sorted on a Sony SH800 in single-cell mode. One nucleus was deposited into each single well containing the respective sample buffers. QRP libraries were prepared using the PicoPlex DNA-seq Kit (Rubicon Genomics, Cat. R300381) according to the manufacturer's protocol and using the indexed PCR primers provided in the kit. DOP libraries were prepared using the SeqPlex DNA Amplification Kit (Sigma, Cat. SEQXE-50RXN) according to the manufacturer's protocol, but with the use of our own custom PCR indexing primers that contain 10 bp index sequences. To avoid over-amplification, all QRP and DOP libraries were amplified with the addition of 0.5 µL of 100X SYBR Green (FMC BioProducts, Cat. 50513) on a BioRad CFX thermocycler in order to monitor the amplification and pull reactions that have reached mid-

exponential amplification.

2.5.3 Nucleosome Depletion

Lithium assisted nucleosome depletion (LAND): Prepared Nuclei were pelleted and resuspended in NIB supplemented with 200 µL of 12.5 mM lithium 3,5-diiodosalicylic acid (referred to as Lithium diiodosalycylate in main text, Sigma, Cat. D3635) for 5 minutes on ice prior to the addition of 800 µL NIB and then taken directly into flow sorting.

Crosslinking and SDS nucleosome depletion (xSDS): Crosslinking was achieved by incubating cells in 10 mL of media (cell culture) or nuclei in 10 mL of HEPES NIB (20 mM HEPES, 10 mM NaCl, 3mM MgCl2, 0.1% igepal, 1x protease inhibitors (Roche, Cat. 11873580001)) (tissue samples) containing 1.5% formaldehyde at room for 10 minutes. The crosslinking reaction was neutralized by bringing the reaction to 200 mM Glycine (Sigma, Cat. G8898-500G) and incubating on ice for 5 minutes. Cell culture samples were crosslinked and then washed once with 10 ml ice cold 1x PBS and had nuclei isolated by incubating in NIB buffer on ice for 20 minutes and pelleted once again. Nuclei were then resuspended in 800 uL 1x NEBuffer 2.1 (NEB, Cat. B7202S) with 0.3% SDS (Sigma, Cat. L3771) and incubated at 42°C with vigorous shaking for 30 minutes in a thermomixer (Eppendorf). SDS was then quenched by the addition of 200 µL of 10% Triton-X100 (Sigma, Cat. 9002-93-1) and incubated at 42°C with vigorous shaking for 30 minutes.

2.5.4 Combinatorial indexing via tagmentation and PCR

Nuclei were stained with 5 µL of 5mg/ml DAPI (Thermo Fisher, Cat. D1306) and passed through a 35 µm cell strainer. A 96 well plate was prepared with 10 µL of 1x Nextera® Tagment DNA (TD) buffer from the Nextera® DNA Sample Preparation Kit (Illumina, Cat. FC-121-1031) diluted with NIB in each well. A Sony SH800 flow sorter was used to sort 2,000 single nuclei into each well of the 96 well tagmentation plate in fast sort mode. Next, 1 µL of a uniquely indexed 2.5 µM transposase-adaptor complex (transposome) was added to each well. These complexes and

associated sequences are described in Amini *et. al*. 2015, Ref. 14. Reactions were incubated at 55°C for 15 minutes. After cooling to room temperature, all wells were pooled and stained with DAPI as previously described. A second 96 well plate, or set of 96 well plates, were prepared with each well containing 8.5 µL of a 0.058% SDS, 8.9 nM BSA solution and 2.5 µL of 2 uniquely barcoded primers at 10 µM. 22 post-tagmentation nuclei from the pool of 96 reactions were then flow sorted on the same instrument but in single-cell sort mode into each well of the second plate and then incubated in the SDS solution at 55°C for 5 minutes to disrupt the nuclear scaffold and disassociate the transposase enzyme. Crosslinks were reversed by incubating at 68°C for an hour (xSDS). SDS was then diluted by the addition of 7.5 µL of Nextera® PCR Master mix (Illumina, Cat. FC-121-1031) as well as 0.5 µL of 100X SYBR Green (FMC BioProducts, Cat. 50513) and 4 µL of water. Real time PCR was then performed on a BioRad CFX thermocycler by first incubating reactions at 72°C for 5 minutes, prior to 3 minutes at 98°C and 15-20 cycles of [20 sec. at 98°C, 15 sec. at 63°C, and 25 sec. at 72°C]. Reactions were monitored and stopped once exponential amplification was observed in a majority of wells. 5 µL of each well was then pooled and purified using a Qiaquick PCR Purification column (Qiagen, Cat. 28104) and eluted in 30 µL of EB.

2.5.5 Library quantification and sequencing

Libraries were quantified between the range of 200bp and 1 kbp on a High Sensitivity Bioanalyzer kit (Agilent, Cat. 5067-4626). Libraries were sequenced on an Illumina NextSeq® 500 loaded at 0.8 pM with a custom sequencing chemistry protocol (Read 1: 50 imaged cycles; Index Read 1: 8 imaged cycles, 27 dark cycles, 10 imaged cycles; Index Read 2: 8 imaged cycles, 21 dark cycles, 10 imaged cycles; Read 2: 50 imaged cycles) using custom sequencing primers described in Amini *et. al*. 2015, Ref.14. QRP and DOP libraries were sequenced using standard primers on the NextSeq® 500 using high-capacity 75 cycle kits with dual-indexing. For QRP there is an additional challenge that the first 15 bp of the read are highly enriched for "G" bases, which are non-fluorescent with the NextSeq® 2-color chemistry and therefore cluster identification on the

instrument fails. We therefore sequenced the libraries using a custom sequencing protocol that skips this region (Read 1: 15 dark cycles, 50 imaged cycles; Index Read 1: 10 imaged cycles; Index Read 2: 10 imaged cycles).

2.5.6 Sequence Read Processing

Software for processing SCI-seq raw reads can be found in the accompanying **Supplementary Software** or downloaded from http://sci-seq.sourceforge.net. Sequence runs were processed using bcl2fastq (Illumina Inc., version 2.15.0) with the --create-fastq-for-index-reads and --with-failed-reads options to produce fastq files. Index reads were concatenated (36 bp total) and used as the read name with a unique read number appended to the end. These indexes were then matched to the corresponding index reference sets allowing for a hamming distance of two for each of the four index components (i7-Transposase (8 bp), i7-PCR (10 bp), i5-Transposase (8 bp), and i5-PCR (10 bp)), reads matching a quad-index combination were then renamed to the exact index (and retained the unique read number) which was subsequently used as the cell identifier. Reads were then adaptor trimmed, then paired and unpaired reads were aligned to reference genomes by Bowtie2 and merged. Human preparations were aligned to GRCh37, Rhesus preparations were aligned to RheMac8, and Human/Mouse mix preparations were aligned to a combined human (GRCh37) and mouse (mm10) reference. Aligned bam files were subjected to PCR duplicate removal using a custom script that removes reads with identical alignment coordinates on a per-barcode basis along with reads with an alignment score less than 10 as reported by Bowtie2.

2.5.7 Single-cell Discrimination

For each PCR plate, a total of 9,216 unique index combinations are possible (12 i7-Transposase indexes × 8 i5-Transposase indexes × 12 i7-PCR indexes × 8 i5-PCR indexes), for which only a minority should have a substantial read count, as the majority of index combinations should be absent – *i.e.* transposase index combinations of nuclei that were not sorted into a given PCR well. These "empty" indexes typically contain very few reads (1-3% of a run) with the

majority of reads falling into *bona fide* single-cell index combinations (97-99% of a run). The resulting histogram of $\log_{10}$ unique read counts for index combinations (**Supplementary Figure 3**) produces a mix of two normal distributions: a noise component and a single-cell component. We then used the R package "mixtools" to fit a mixed model (normalmixEM) to identify the proportion (λ) mean (μ) and standard deviation (σ) of each component. The read count threshold to qualify as a single-cell library was taken to be the greater of either one standard deviation below the mean of the single-cell component in $\log_{10}$ space, or 100 fold greater than the mean of the noise component (+2 in $\log_{10}$ space), and had to be a minimum of 1,000 unique reads.

2.5.8 Human-Mouse Mix Experiments

We took one of two approaches to mix human (GM12878 or HeLa S3) and mouse (3T3) cells: i) mixing at the cell stage (HumMus.LAND1 and HumMus.LAND2) or ii) mixing at the nuclei stage (HumMus.LAND3, HumMus.LAND4, and HumMus.xSDS). The reason we employed the latter was to control for nuclei crosslinking or agglomerating together that could result in doublets. Libraries were constructed as described above, for instances where two distinct DAPI-positive populations were observed during flow sorting, included both populations in the same gate so as not to skew proportions. Reads were processed as in other experiments, except reads were instead aligned to a reference comprised of GRCh37 (hg19) and mm10. The mapping quality 10 filter effectively removed reads that aligned to conserved regions in both genomes and then for each identified single cell, reads to each species were tallied and used to estimate collision frequency. For our early LAND preparations we sorted 25 indexed nuclei per PCR well and produced total collision rates (*i.e.* twice the human-mouse collision rate) of 28.1% and 10.4%. For the second two LAND preparations we sorted 22 nuclei per PCR well, which produced a total collision rate of 4.3% for one preparation and no detectable collisions in another. We also tested two FANS sorting conditions for our xSDS preparation, one was permissive and allowed a broader range of DAPI fluorescence, and the other more restrictive, and carried out both preparations on

separate sides of the same PCR plate. For the permissive gating we observed a total collision rate of 23.6% with a substantial reduction for the more restrictive gating at 8.1%. Based on these results we decided to continue sorting 22 nuclei per PCR well using the more restrictive FANS

2.5.9 Library Depth Projections

To estimate the performance of a library pool if, or when, it was sequenced to a greater depth, we incrementally sampled random reads from each SCI-seq preparation across all index combinations including unaligned and low-quality reads without replacement at every one percent of the total raw reads. For each point we identified the total number reads that are aligned with high quality (MQ $\geq$ 10) assigned to each single-cell index and the fraction of those reads that are unique, non-PCR duplicates, as well as the corresponding fraction of total reads sampled that were assigned to that index. Using these points we fit both a nonlinear model and a Hanes-Woolfe transformed model to predict additional sequencing for each individual single-cell library within the pool and projected out to a median unique read percentage across cells of 5%. To determine the accuracy of the models, we determined the number of downsampled raw reads of each library that would reach the point in which the median unique read percentage per cell was 90%, which is somewhat less than what was achieved for libraries that were sequenced at low coverage. We then subsampled the pre-determined number of reads for 30 iterations and built a new model for each cell at each iteration and then predicted the unique read counts for each cell out to the true sequencing depth that was achieved. The standard deviation of the true read count across all iterations for all cells was then calculated.

2.5.10 Genome Windowing

Genomic windows were determined on a per-library basis using custom tools. For each chromosome the size of the entire chromosome was divided by the target window size to produce the number of windows per chromosome. The total read count for the chromosome summarized over the pool of all single cells (GM12878 for all human samples where absolute copy number was

determined, as well as for each pooled sample where amplifications or deletions relative to the mean copy number were determined) was then divided by the window count to determine the mean read count per window. The chromosome was then walked and aligned reads from the pool tallied and a window break was made once the target read count per window was reached. Windows at chromosome boundaries were only included if they contained more than 75% of the average reads per window limit for that chromosome. By using dynamic windows we accounted for biases, such as highly repetitive regions, centromeres and other complex regions that can lead to read dropout in the case of fixed size bins[22].

2.5.11 GC Bias Correction

Reads were placed into the variable sized bins and GC corrected based on individual read GC content instead of the GC content of the dynamic windows. We posit that the large bin sizes needed for single-cell analysis average out smaller scale GC content changes. Furthermore, SCI-seq does not involve pre-amplification where large regions of the genome are amplified, therefore GC bias originates solely from the PCR and is amplicon-specific. To calculate correction weights for the reads we compared the fraction of all reads with a given GC to the fraction of total simulated reads with the average insert size at the same GC fraction. This weight was then used in lieu of read counts and summed across all reads in a given window. All regions present in DAC blacklisted regions were excluded from analysis for the human sample analyses (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability)[19]. Following GC correction, all reads were normalized by the average number of reads per bin across the genome. Finally for each window we took the normalized read count of each cell and divided it by the pooled sample baseline to produce a ratio score.

2.5.12 Measures of data variation

To measure data quality, we calculated two different measures of coverage dispersion: the median absolute deviation (MAD), the median absolute pairwise difference (MAPD). For each

score we calculated the median of the absolute values of all pairwise differences between neighboring bins that have been normalized by the mean bin count within the cell (log2 normalized ratios for the MAPD scores). These scores measure the dispersion of normalized binned reads due to technical noise, rather than due copy number state changes, which are less frequent[2,22].

2.5.13 Copy Number Variant Calling

CNV calling was performed on the windowed, GC corrected and bulk sample normalized reads with two available R packages that employ two different segmentation strategies: a Hidden Markov Model approach (HMMcopy, version 3.3.0, Ref. 25) and Circular Binary Segmentation (DNAcopy, version 1.44.0, Ref. 24). Values were Log2 transformed for input (2*log2 for CBS) and copy number calls were made based on the optimized parameters from Knouse et al. 2016, Ref. 11. For optimal sensitivity and specificity to detect copy number calls with sizes ≥5Mb we set the probability of segment extension (E) to 0.995 for HMM and for CBS we chose the significance level to accept a copy number change (α) to be 0.0001. The Log2 cutoffs for calling losses or gains were 0.4 and −0.35 for HMM and 1.32 and 0.6 for CBS. As an additional tool for CNV calling we used Ginkgo[22], which uses an alternative method for data normalization. We uploaded bed files for each cell and a bulk down sampled bed file, which we created with Picard Tools (we used a down sample probability of 0.1). For the analysis we chose to segment single cells with the down sampled bulk bed file and when ploidy was known for the samples we created FACS files to force Ginkgo to normalize to that ploidy. Calls for the three methods were intersected either on a per-window basis or were filtered to only include calls that span ≥ 80% of a chromosome arm and then intersected for aneuploidy analysis.

2.5.14 Tumor breakpoint analysis

Unlike the assessment of sporadic aneuploidy, tumor structural variation is much more complex with a large portion of breakpoints within chromosomes. Further, sporadic aneuploidy within any given subclone of a tumor is less pertinent than an accurate profile of the subpopulations

that are present. We therefore used the HMM and CBS segmented ratio score matrixes to identify breakpoints by tallying up the boundaries of segmented regions across cells. We then used the resulting distribution of shared chromosomal breakpoints across the genome to identify local maxima to account for variability in which specific window the call was made, and then retained those that are present in at least 5% of cells. We then merged all windows within each breakpoint span and calculated the new log2 ratio of each aneuploid cell over the mean values of the euploid population. We then carried out principle components analysis prior to k-means clustering with a k value determined by Silhouette analysis. To minimize the effect of doublets which can account for ~10% of putative single cells and also to exclude low-performance cells, we retained only those in the close proximity to their respective centroids. We then merged sequence reads for all cells within each cluster and then carried out a higher resolution CNV analysis (target window size of 100 kbp) using an HMM strategy followed by absolute copy number state identification and the identification of focal amplifications and deletions using a sliding window outlier strategy[20]. Intra-tumoral clonal relationships are most accurately captured by shared breakpoints as opposed to the drift in copy number of a segment based on the assumption that structural changes involving breaks in the DNA as being more impactful on the cell. We therefore compared cells by assessing the proportion of segments between breakpoints that were identified using the high resolution (100 kbp) CNV analysis that overlapped by at least 90% (to account for noise in the exact window that was called as the copy number change) out of the total number of segments.

Chapter 3: The accessible chromatin landscape of the murine hippocampus at single-cell resolution

John R. Sinnamon*, **Kristof A. Torkenczy***, Michael W. Linhoff, Sarah Vitak, Ryan M. Mulqueen, Hannah A. Pliner, Cole Trapnell, Frank J. Steemers, Gail Mandel, Andrew C. Adey

* Denotes equal contribution

This chapter has been reformatted for inclusion for this dissertation from the manuscript titled: "The accessible chromatin landscape of the murine hippocampus at single-cell resolution" published in Genome Biology, March 2019. A.C.A., J.R.S., and G.M. designed all experiments. A.C.A., J.R.S., K.A.T., and M.W.L. wrote the manuscript. S.A.V. performed all sci-ATAC-seq preparations. A.C.A. and K.A.T. performed computational analysis and wrote software associated with this work. J.R.S. prepared all tissue samples and cultures, and performed analyses and interpretation of the data. M.W.L. aided in analysis of the data and interpretation of findings. R.M.M. contributed to sci-ATAC-seq and *scitools* protocol development and data processing pipelines. H.A.P. and C.T. provided early access to computational tools and aided in the co-accessibility analysis and interpretation. F.J.S. provided reagents and contributed to sci-ATAC-seq method development and implementation. All authors reviewed and approved the manuscript.

All supplemental figures and tables are not included in this dissertation and should be referred to at https://genome.cshlp.org/content/early/2019/04/01/gr.243725.118

3.1 Abstract

Here we present a comprehensive map of the accessible chromatin landscape of the mouse hippocampus at single-cell resolution. Substantial advances of this work include the optimization of single-cell combinatorial indexing assay for transposase accessible chromatin (sci-ATAC-seq), a software suite, *scitools*, for the rapid processing and visualization of single-cell combinatorial indexing datasets, and a valuable resource of hippocampal regulatory networks at single-cell resolution. We utilized sci-ATAC-seq to produce 2,346 high-quality single-cell chromatin accessibility maps with a mean unique read count per cell of 29,201 from both fresh and frozen hippocampi, observing little difference in accessibility patterns between the preparations. Using this dataset, we identified eight distinct major clusters of cells representing both neuronal and non-neuronal cell types and characterized the driving regulatory factors and differentially accessible loci that define each cluster. Within pyramidal neurons, we identified four major clusters, including CA1 and CA3 neurons, and three additional subclusters. We then applied a recently described co-accessibility framework, *Cicero*, which identified 146,818 links between promoters and putative distal regulatory DNA. Identified co-accessibility networks showed cell-type specificity, shedding light on key dynamic loci that reconfigure to specify hippocampal cell lineages. Lastly, we carried out an additional sci-ATAC-seq preparation from cultured hippocampal neurons (899 high-quality cells, 43,532 mean unique reads) that revealed substantial alterations in their epigenetic landscape compared to nuclei from hippocampal tissue. This dataset and accompanying analysis tools provide a new resource that can guide subsequent studies of the hippocampus.

3.2 Introduction

A major goal in the life sciences is to map cell types and identify the respective genomic properties of each of the cell types in complex tissues. Traditional strategies that utilize intact tissue are limited to averaging of the constituent cell profiles. To overcome this limitation, there has been a burst in development of unbiased single-cell genomics assays, leveraging the concept that each single cell can only occupy a single position in the landscape of cell types (Trapnell, 2015). This push into the single-cell space has largely centered on the use of single-cell transcriptional profiling. While profiling the RNA complement has produced valuable information (Saunders et al., 2018; Amit Zeisel et al., 2018), the ability to profile chromatin status, *i.e.* active versus inactive, has lagged behind, leaving open the question as to what extent accessible chromatin profiles are linked to cell specificity, particularly with respect to distal enhancer elements (Corces et al., 2016).

Recently, progress has been made to ascertain chromatin accessibility profiles in single cells using ATAC-seq (assay for transposase-accessible chromatin) technologies. These strategies have been applied to myogenesis (Pliner et al., 2018a), hematopoietic differentiation (Buenrostro et al., 2018), fly embryonic development (D. A. Cusanovich, Reddington, et al., 2018), the mouse (Preissl et al., 2018) and human cortex (Lake et al., 2017), and most recently an atlas of multiple tissues in the mouse, though notably lacking the hippocampus (D. A. Cusanovich, Hill, et al., 2018). The core concept behind the methods utilized in several of these studies is a combinatorial indexing schema whereby library molecules are barcoded twice, once at the transposase stage and then again at the PCR stage. This platform has also been extended to profile other properties including transcription, genome sequencing, chromatin folding, and DNA methylation (Cao et al., 2017; Mulqueen et al., 2018; Ramani et al., 2017; Vitak et al., 2017a; Yin et al., 2018). In this work, we optimized the sci-ATAC-seq assay for analysis of fresh and frozen hippocampal tissue samples to produce single-cell chromatin accessibility profiles in high throughput, with greater information

content – as measured by unique reads per cell. These improvements will also facilitate the use of this technology platform on frozen samples, enabling the assessment of banked tissue isolates.

The hippocampus is critical to the formation and retrieval of episodic and spatial memory (O'Keefe & Dostrovsky, 1971; Scoville & Milner, 1957; Smith & Milner, 1981; Zola-Morgan et al., 1986). Historically, cell types within the hippocampus have been broadly classified by their morphology (Ramon y Cajal 1911; Lorente de No 1934) and electrophysiological properties (Kandel et al., 1961; Kandel & Spencer, 1961; Spencer & Kandel, 1961b, 1961a). More recently types have been identified by their transcriptional profiles (Cembrowski et al., 2016; Lein et al., 2004), and single-cell transcriptomics has also revealed potential subclasses within previously defined cell types (Habib et al., 2017; A. Zeisel et al., 2015). The defined classes of cells within the hippocampus and the existing single-cell transcriptome data allowed us to refine our sci-ATAC-seq method and provide the first single-cell epigenomics profile of the murine hippocampus.

3.3 Results

3.3.1 Single-cell chromatin accessibility profiles from mouse hippocampus

We utilized sci-ATAC-seq to profile two fresh and two frozen mouse total hippocampi to map the accessible chromatin landscape (Methods). Each sample was freshly isolated from an adult (P60) wild type mouse (C57-Bl6) and either processed immediately or flash frozen using liquid nitrogen. Nuclei were isolated and carried through the sci-ATAC-seq protocol with several optimizations from previously described implementations (Methods, Figure 3.1A, and Supplemental Protocol). Briefly, nuclei were isolated by dounce homogenization of tissue in nuclei isolation buffer followed by Fluorescence Assisted Nuclei Sorting (FANS) using DAPI as a stain to select for intact, single nuclei. One of the key improvements to our workflow was the addition of Tween-20 (Sigma) to the nuclei isolation buffer which we believe increased the permeability of the nucleus and removed more of the cell membrane. We then performed sci-ATAC-seq as previously described using a 55°C tagmentation temperature (Methods, Supplemental Protocol).

Sequence reads were processed and subsequent analysis was performed using *scitools* (Supplemental Code).

In total, we produced 2,346 single cells passing quality control (≥1,000 unique reads present in peaks and ≥25% of all unique reads present in peaks, alignment q≥10, not aligned to chrM, unscaffolded, alternative, or random contigs) evenly represented across replicates (2 frozen, 2 fresh). Based on existing single-cell RNA-seq studies we assumed that our cell number should be sufficient for preliminary cell type deconvolution (Habib et al., 2017; A. Zeisel et al., 2015); however, we believe future innovation may enable greater numbers. Cells had a mean unique aligned read count of 29,201, which is notably higher than other high throughput single-cell ATAC-seq workflows to date (Supplemental Table 1). We observed a strong correlation in ATAC signal between the aggregate profiles of the four replicates (Pearson R > 0.99), indicating high reproducibility across preparations for both fresh and frozen tissue. We did notice a statistically significant (t-test p-value = 2.2×10^{-6}) increased number of unique reads per cell in the frozen samples; however, this can be attributed to greater sequencing depth (Supplemental Figure 1,2, Supplemental Table 1), or possibly due to the freeze-thaw cycle increasing the permeability of the nucleus. Between replicates of the same preparation method no statistically significant differences were observed. Chromatin accessibility peaks were identified by the aggregation of all cells to produce an ensemble dataset containing all called peaks, resulting in a preliminary set of 93,994 high-confidence peaks, with a mean of 36.4% of reads from each cell falling within these regions. The fraction of reads in peaks for the frozen samples was greater than for the fresh samples (Supplemental Figure 2, Supplemental Table 1, p-value = 1.3×10^{-4}).

We constructed a read count matrix of our ensemble peaks and single cells from all conditions (Supplemental Data – InVivo.counts.matrix) by tallying the number of reads for each cell at each peak. We next utilized *scitools* to perform Latent Semantic Indexing (LSI), as previously described (D. A. Cusanovich, Reddington, et al., 2018; D. a Cusanovich et al., 2015), with the exclusion of cells with reads at fewer than 1,000 sites and of sites with fewer than 50 cells

exhibiting signal. The LSI matrix was projected into two-dimensional space using t-distributed Stochastic Neighbor Embedding (t-SNE) for visualization, which revealed distinct domains occupied by clusters of cells. We next used density-based clustering (Ester et al., 1996) and aggregated cells by cluster, called cluster-specific peaks and added them to a union peak set (n=98,043, 4% increase in peak count) for which all subsequent analysis was performed. We then identified nine major clusters (Figure 1C), one of which being likely barcode collisions and removed from further analysis (Methods). A comparison of the proportion of cells assigned to each cluster with respect to fresh or frozen samples did not yield a significant difference ($X^2 = 9.85$, *p*-value = 0.20; Figure 3.1B, Supplemental Table 2), though increased proportions of interneurons and microglia were observed in the frozen preparation.

To assign each of our identified clusters to a cell type, we took advantage of published single-cell RNA-seq data that produced sets of marker genes associated with cell types identified at the transcriptional level (Habib et al., 2017; A. Zeisel et al., 2015). For each set of cell-type-specific genes, we identified peaks 20 kilobasepairs (kbp) in either direction from the transcriptional start site, which were then used to calculate the enrichment for accessible chromatin for each cell within these regions. This produced a deviation z-score, similar to previously described methods (Buenrostro et al., 2015; Schep et al., 2017). We then visualized these scores on our t-SNE projections, which enabled us to clearly identify a number of neuronal and non-neuronal cell types, including astrocytes (AST), two groups of pyramidal neurons (designated Neurons 1; NR1 and Neurons 2; NR2), interneurons (INT), oligodendrocytes (OLI), microglia (MRG), and oligodendrocyte progenitor cells (OPCs) (Figure 1D). To complement this strategy we also turned to marker genes described previously in the literature that were not present in available single-cell RNA-seq datasets and assessed the chromatin accessibility at elements proximal to these genes (Cembrowski et al., 2016; Lein et al., 2004; Y. Zhang et al., 2014) (Figure 1E, Supplemental Figure 3). For example, the *Glul* gene, an established marker for astrocytes (Fages et al., 1988; Martinez-Hernandez et al., 1977) showed accessibility only in the population of cells we identified as

astrocytes (Figure 1E, left). *Prox1*, previously shown to be enriched in the dentate gyrus (Lein et al., 2004), is accessible predominantly in the dentate granule cell population (GRN, Figure 3.1E, right). Markers for particular cell types were also consistent with *in situ* hybridization data from the Allen Brain Institute (Supplemental Figure 3) and RNA-seq data from sorted cells (Cembrowski et al., 2016; Y. Zhang et al., 2014). Based on our cell type assignments, the number of cells in each population reflects the proportions seen within the intact hippocampus (Abusaad et al., 1999) (Supplemental Table 1). This includes the observation of 14 fold and 41 fold fewer astrocytes and microglia when compared to neurons, respectively, in line with previous studies (Kimoto et al., 2009).

To further confirm our cell type assignments, we utilized the recently-released function in *Seurat3* for the co-embedding of single-cell ATAC-seq and single-cell RNA-seq datasets in a shared t-SNE space (Stuart et al., 2018). We first generated gene activity scores using *Cicero* (described below), which utilizes linked distal regulatory elements and promoters to approximate the putative activity of each gene (Pliner et al., 2018a). These scores, along with transcript count matrices from Smart-seq and DroNc-seq publications (Habib et al., 2017; A. Zeisel et al., 2015), were processed using *Seurat3* to identify anchors and effectively normalize them to one another to enable PCA and then visualization in a shared t-SNE space (Figure 3.1F). Encouragingly, our cells were positioned proximal to cells with matching assignments in their respective publications. We next identified 18 distinct clusters (Figure 3.1G) using *PhenoGraph* (Levine et al., 2015). Within these clusters we quantified the percentage of cells assigned to each cell type within each of the three datasets to assess the most represented cell type that is present. This analysis further confirmed our cell type assignments with substantial concordance between the highest represented cell types across platforms (Figure 3.1H). However, cross-dataset assignment was far from perfect, with certain cell types performing better than others, *e.g.* Oligodendrocytes performed well, versus granule cells which did not. We suspect that the major driver of the discrepancies is due to the indirect nature of the gene activity scores for the single-cell ATAC-seq data.

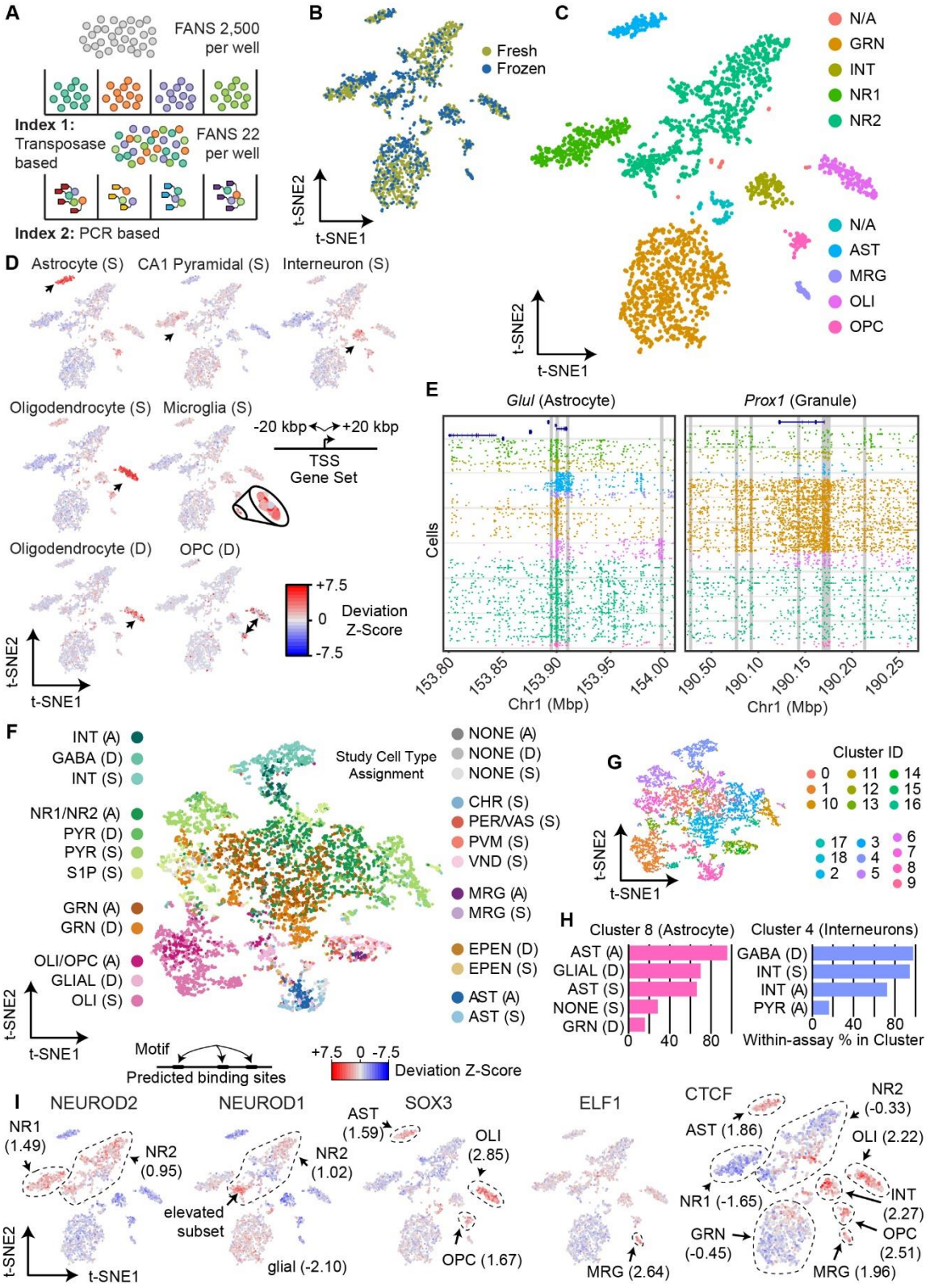

Figure 3.1 sci-ATAC-seq of the murine hippocampus. (A) sci-ATAC-seq workflow. Two indexes are incorporated into library molecules for each cell enabling single-cell discrimination. (B) LSI-t-SNE projection of single cells colored by tissue preparation method. Little variation in t-

SNE space is observed between fresh or frozen starting material. (C) LSI-t-SNE projection of cells colored by assigned cluster and cell type. (D) Enrichment of accessibility of proximal regulatory elements for marker genes as identified by single-cell RNA-seq (Smart-seq protocol, "(S)") (A. Zeisel et al., 2015) and DroNc-seq "(D)" (Habib et al., 2017) for each cell. The microglial population is enlarged for visibility. Black arrows indicate the cell cluster associated with the marker gene set. (E) sci-ATAC-seq read plots at *Glul* (astrocyte marker gene) and *Prox1* (dentate granule cell marker gene). (F) Co-embedding (t-SNE) of single-cell RNA-seq and DroNc-seq cells from (D) with our sci-ATAC-seq cells using *Seurat3*. Cells are colored by their study "(A)" designates this study, "(D)" designates cells from Habib et al. 2017, and "(S)" designates cells from Zeisel et al. 2015, and cell type desgnation from their published study (RNA) or our designations (for the sci-ATAC-seq cells). (G) PhenoGraph cluster designations on the co-embedded cells. (H) Representative cluster cell compositions. The percentage of cells within each of the three assays that were assigned to the co-embedding cluster using PhenoGraph are reported. For example (noted by an asterisk) in Cluster 8, 93.0% of the sci-ATAC-seq cells that were assigned to Cluster 8 were designated as astrocytes. (I) *ChromVAR* global motif deviation z-scores for each cell for select motifs. Dashed lines and values correspond to mean values of cell populations.

3.3.2 Global DNA binding motif accessibility

To assess the global activity of DNA binding proteins we utilized the recently-described software tool, *ChromVAR* (Schep et al., 2017), which aggregates the chromatin accessibility signal genome-wide at sites harboring a given motif, followed by the calculation of a deviation z-score for each cell. This score represents the putative activity level of the DNA binding protein that corresponds to the assessed motif, which we then visualized on our t-SNE projections (Figure 3.1I, heatmap in Supplemental Figure 4). In line with expectations, our cell type clusters showed enrichment for accessibility at DNA binding motifs concordant with the identified cell type (Figure 3.1I, Supplemental Figure 5). The analysis included the assessment of global accessibility for neuron-specific factors such NEUROD2, which associates with active chromatin marks (*e.g.* H3K27ac) in cortical tissue (Guner et al., 2017) and exhibited greater accessibility in the two pyramidal cell clusters (mean z-score (μ_z) = 1.49 and 0.95 for NR1 and NR2 respectively, all other cell types $\mu_z \leq -0.74$). We also observed increased accessibility of NEUROD1, also associated with active chromatin (Pataskar et al., 2016), in a portion of one of the pyramidal neuron clusters (NR2, $\mu_z = 1.02$) with less accessibility across glial populations ($\mu_z \leq -2.10$). While many studies have identified a role for SOX3 during neural differentiation, consistent with a previous expression study (Cheah & Thomas, 2015), we observed increased SOX3 accessibility in astrocyte ($\mu_z = 1.59$),

oligodendrocyte (μ_z = 2.85), and OPC populations (μ_z = 1.67), suggesting a glial role for this transcription factor in adulthood. ELF1, an ETS family member associated with activating interferon response in the hematopoietic lineage (Larsen et al., 2015), exhibited elevated accessibility in the microglial population (μ_z = 2.64), which also respond to interferon in the brain (*e.g.* (Goldmann et al., 2015)). Of particular interest was the strong enrichment for CTCF motif accessibility in glial cell populations (AST μ_z = 1.86, OLI μ_z = 2.22, OPC μ_z = 2.51, MRG μ_z = 1.96) and interneurons (μ = 2.27) when compared to granule cells (μ_z = -0.45) or pyramidal neurons (NR1 μ_z = -1.65, NR2 μ_z = -0.33), an observation that was reinforced by our subsequent differential accessibility analysis described below. To confirm that the observed motif accessibility increase is due to true CTCF binding sites and not just the motif presence, we also carried out a deviation analysis using peaks called from publicly available CTCF ChIP-seq data of the mouse hippocampus (Sams et al., 2016), which revealed very similar patterns of accessibility (Supplemental Figure 6, Pearson R^2 = 0.68).

3.3.3 Differential accessibility by cell type

We next sought to show that accessible regions could be identified according to cell type. To provide sufficient signal, we aggregated cells within clusters in their local neighborhoods as has been described previously (D. A. Cusanovich, Reddington, et al., 2018) and then carried out a differential accessibility analysis for each cluster compared to the rest of the cells (Methods, Figure 3.2A). Numbers of significant (q-value ≤ 0.01, Log_2 fold-change ≥ 1) loci ranged from 894 (OPCs) to 7,796 (granule cells), with substantial cell-type specific signal (Figure 3.2B, left, Supplemental Figure 7-9). To characterize these loci, we performed a motif enrichment analysis to identify DNA binding proteins that may bind within the differentially accessible regions (Figure 3.2B, right). In contrast to the prior, global accessibility analysis, where all accessible loci were utilized to detect increased signal at sites harboring a given motif in each cell; here, we are detecting enrichment of motifs in the specific subsets of loci that were determined to be differentially accessible. This

strategy revealed enrichment for binding by the SOX10 transcription factor in oligodendrocytes (Claus Stolt et al., 2002) and by NEUROG2 in the dentate granule cells (Roybon et al., 2009). Within the interneuron population, the motif with the highest enrichment was CTCF. This is consistent with our prior, global analysis of accessibility at motifs (Figure 3.1I); however, this reciprocal approach suggests that a set of sites very specific to interneurons harbor CTCF as opposed to sites that may be shared across numerous cell types with varying levels of accessibility. One of these regions was in an intron in the gene encoding actin filament associated protein 1 (*Afap1*, Supplemental Figure 10). The ChIP data revealed CTCF binding within the same intron flanking the accessible region. Previous work has suggested that CTCF may have a particular importance in this cell type (S. Kim et al., 2018). CTCF binding motifs were enriched in the accessible chromatin of affinity purified parvalbumin positive cortical interneurons but not in VIP positive interneurons or excitatory neurons (Mo et al., 2015) and mice expressing one CTCF allele only in inhibitory neurons exhibit memory impairment (S. Kim et al., 2018). Recent data has also suggested that CTCF plays a role in the generation of cortical interneurons by regulating the expression of the LIM homeodomain factor LHX6 (Elbert et al., 2019). The potential selective importance of CTCF in interneurons warrants further study.

To further determine the utility of our method in assigning regulatory elements to cell types, we tested whether we could parse enhancers that had been identified in the literature as inducers of target genes in response to neuronal activity. We focused on the *Fos* gene that has been studied previously as a general reporter of neuronal activity throughout the brain (Bullitt, 1990). Specifically, five enhancers (*E1*-*E5*) have been characterized (T. K. Kim et al., 2010) for both regulation during neuronal activity and type of stimulation (Joo et al., 2015). When we examined ATAC-seq signal at the five enhancers across cell types in hippocampus, we identified cell type specific patterns of accessibility. Notably, enhancers *E1* and *E3* were accessible only in neurons, while *E2* and *E5* were accessible in all cell types (Figure 3.2C).

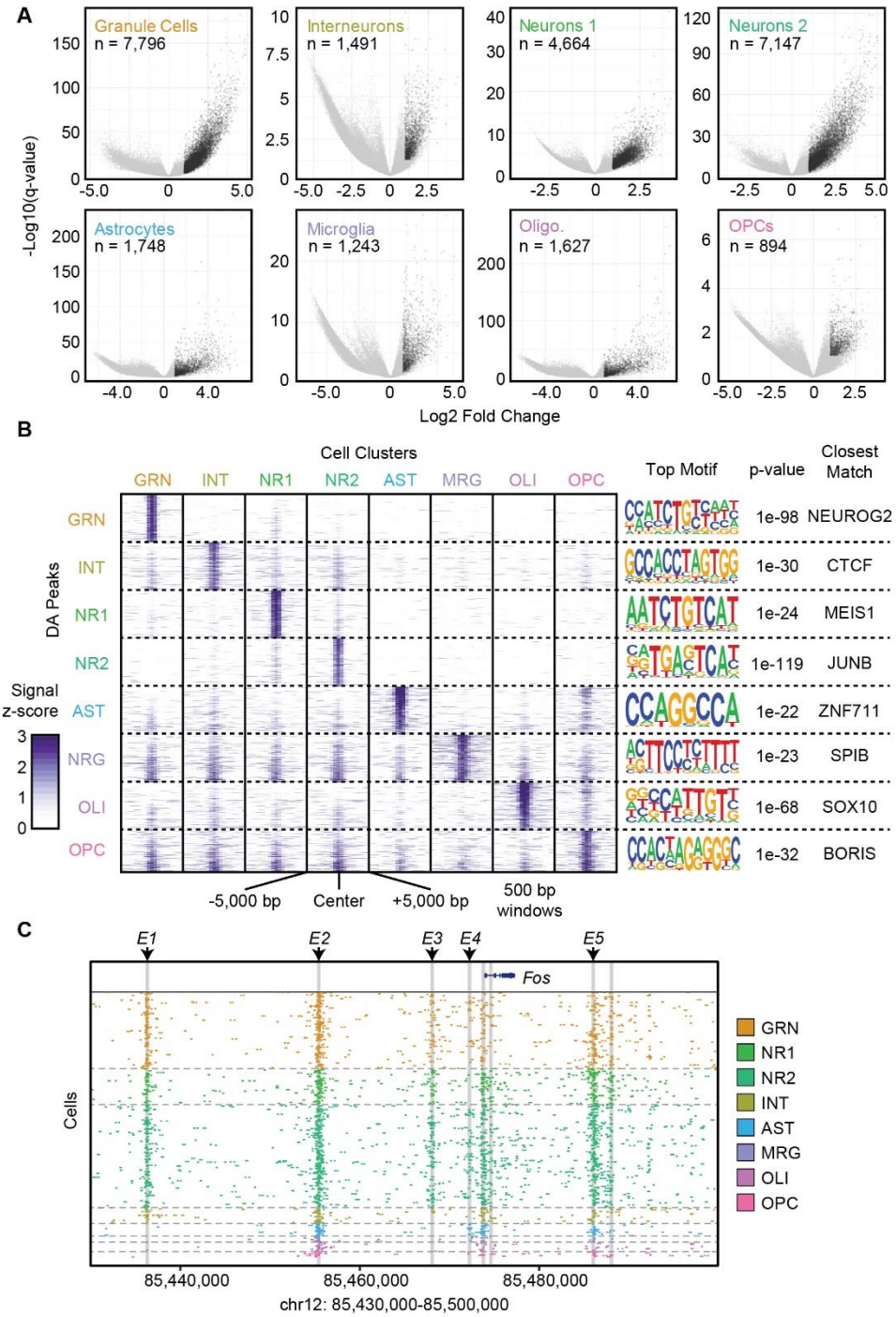

Figure 3.2 Differential accessibility analysis between cell types. (A) Volcano plots $-\log_{10}$(*q*-value) (y-axis) versus $\log_2$ accessibility signal fold change (x-axis) showing all peaks. Each comparison is for the indicated cell population versus all other cell types. Significant peaks (number indicated, *q*-value ≤ 0.01, $\log_2$ fold change ≥ 1) are in black. (B) ATAC-seq signal plots for the top differential accessible peaks for each cell type. The most significantly enriched motif for each set is shown on the right along with the corresponding *p*-value and closest matching known motif. (C) *c-Fos* locus with enhancers E1-5 highlighted to show cell-type-specific utilization.

Further, enhancer *E4* was accessible in group 2 but not group 1 pyramidal neurons and was also accessible in a small portion of dentate granule cells. Our findings suggest cell type specificity in stimuli responsiveness within the hippocampus, even between pyramidal cell subpopulations, opening the door to new studies of the basis of these signaling differences and demonstrating the utility of single-cell epigenomics over traditional bulk tissue assays.

More generally, our differential accessibility analysis was able to identify new enhancers by comparison with chromatin marks known to be associated with enhancers (Gjoneska et al., 2015). For example, when examining the most significantly differentially accessible loci for dentate granule cells, one of the top hits was a region marked by both H3K4me1 and H3K27ac, suggesting a putative enhancer upstream of the gene *Slc4a4* (Supplemental Figure 11). *Slc4a4* encodes a sodium/bicarbonate co-transporter involved in mediating both intracellular and extracellular pH (Svichar et al., 2011), and *Slc4a4* expression is elevated in dentate granule neurons. While these accessible loci were enriched only in dentate neurons, several other accessible regions were identified in dentate granule cells as well as in the two pyramidal neuron populations, suggesting this gene is expressed in multiple cell types and, like *Fos*, may exhibit variable responses in different cell types.

3.3.4 Pyramidal neuron subclustering

In our initial clustering, the two most prevalent pyramidal neuron populations, CA1 and CA3 were not able to be definitively resolved. We reasoned that analyzing these cells in isolation and using a recently-described method for discerning themes, or 'topics' of correlated signal within the data, *cisTopic* (Bravo González-Blas et al., 2018) may provide improved granularity. Based on a Latent Dirichlet Allocation framework, *cisTopic* identifies related sets of peaks that are classified as topics. On our NR1 and NR2 dataset, the optimum number of topics was determined to be 30 (Supplemental Figure 12) which were then used to project cells into two dimensional space using Uniform Manifold Approximation and Projection (Becht et al., 2018b) (UMAP; Figure 3.3A). Cells

split into four distinct groups that were identified using *PhenoGraph* (Levine et al., 2015) on the topic matrix (Figure 3.3B). One of the clusters was comprised almost exclusively of the NR1 cells (95%), with the NR2 cells split into three groups. Notably, we did not observe any bias in cluster assignment with respect to the fresh versus frozen prepared cells (Supplemental Figure 13). We next examined genes specifically associated with CA1 and CA3 neurons and were clearly able to assign two of the four clusters based on specific accessibility of promoters and/or *cis* regulatory elements at these loci (Supplemental Figure 14). We also observed some enrichment of CA2-specific genes and genes associated with mossy cells (MC) in two of the other clusters, suggesting that these cell types are likely present in the identified clusters; however, they may not make up the entirety of the population.

In addition to improved sensitivity, *cisTopic* produces sets of peaks that are associated with one another as topics (Figure 3C, Supplemental Figure 15), several of which exhibited high cluster specificity. This included CA3-specific topic 13, which was enriched for NEUROD1. These cells were within the same region of the NR2 cluster that also exhibited increased NERUROD1 accessibility (Figure 3.3B, right, Figure 3.1I). Motif enrichment files for all topics can be found in Supplemental Data 1. We additionally performed a differential accessibility analysis between the clusters (Supplemental Figure 16). While none of the significant peaks were proximal to definitive marker genes, these sites may be useful to inform future functional studies.

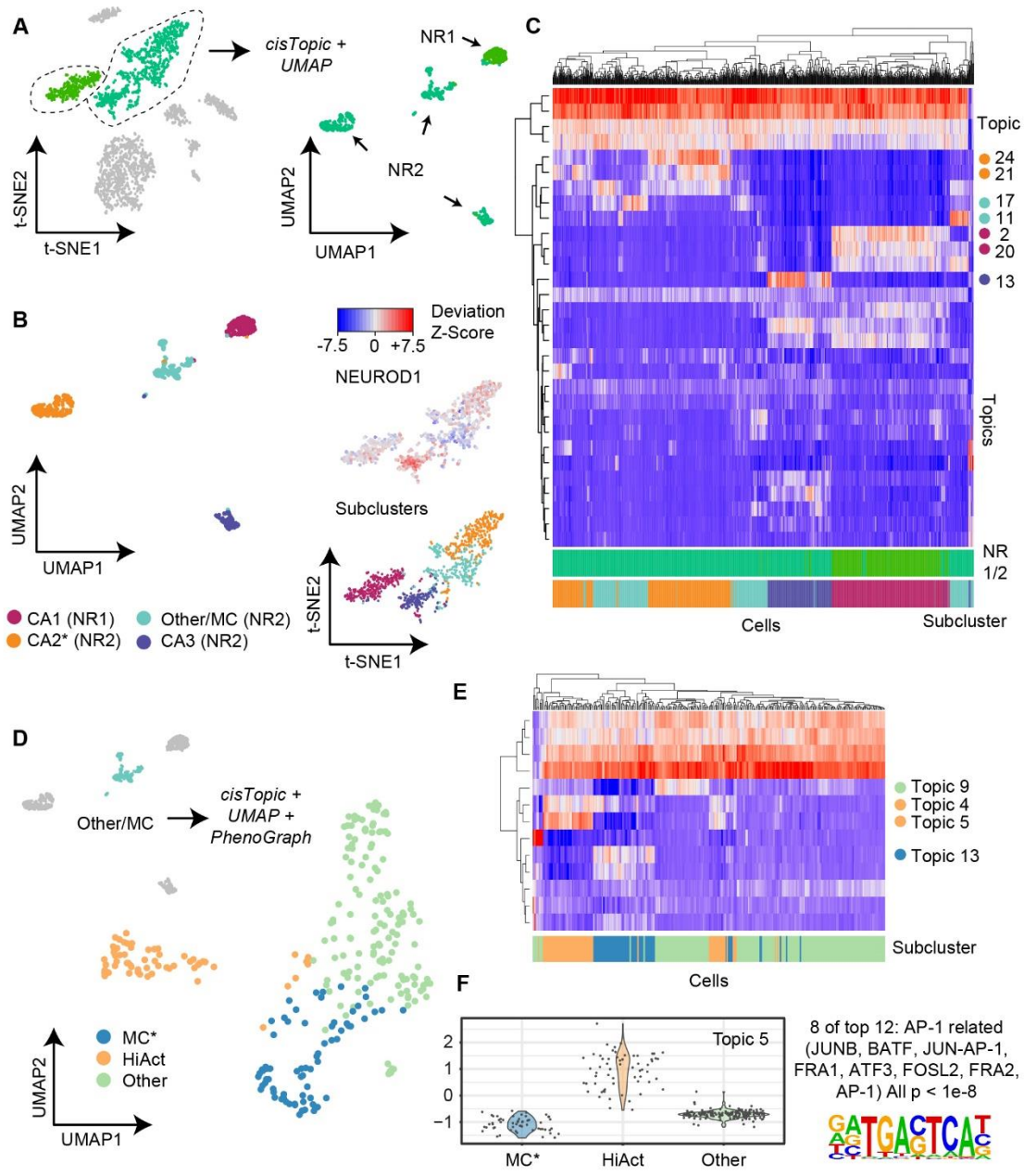

Figure 3.3 Pyramidal neuron subclustering. (A) Subclustering of the NR1 and NR2 assigned cells using *cisTopic* and *UMAP*. (B) Cluster assignments identified using *PhenoGraph*. CA1 and CA3 neuronal populations exhibited strong signal at cell-type specific marker genes. Asterisk indicates putative assignment based on modest enrichment at marker genes. Right panels show the NEUROD1 motif enrichment in the original t-SNE coordinates (top) which correspond to the region of cells assigned to CA3 cluster (bottom). (C) Biclustering of *cisTopic* topics and weights for each cell. Highlighted topics exhibit high cluster specificity. (D) Further subclustering of the Other/MC cell population produced three distinct groups, including putative mossy cells (MC). (E) Biclustering of *cisTopic* topics and weights for each cell. Highlighted topics exhibit high cluster specificity. (F) Topic 5, specific to one of the subclusters, is highly enriched for AP-1 related motifs, suggesting the cells may be in a state of heightened activity.

We suspected that the fourth cluster (teal, Other/MC) might contain additional cellular subtypes based on the hierarchical clustering of topics. We therefore carried out the same subclustering analysis as we did for the NR 1 and 2 groups specifically for these cells (optimum topic number = 13, Supplemental Figures 17-18, Supplemental Data 2), which revealed three distinct clusters (Figure 3.3D). When assessing the topics closely associated with one of the clusters, we observed a very high enrichment for AP-1 associated proteins, suggesting that they may be neurons in a heightened activity state (Figure 3.3E), though we did not observe enrichment of accessibility for any cell-type specific marker genes or DNA binding motifs, as was the case for a second cluster. We did observe increased chromatin accessibility proximal to several Mossy Cell marker genes (Cembrowski et al., 2016) which was most pronounced at *Pmp22* and *Thbs2* (Supplemental Figure 19).

3.3.5 Cis regulatory networks in the hippocampus

Many enhancer elements reside far from the transcription start sites of the genes they regulate, making enhancer-gene associations challenging. To accomplish this, we leveraged the recently-described *Cicero* algorithm (Pliner et al., 2018a), which uses an unsupervised machine-learning framework to link distal regulatory elements to their prospective genes via patterns of co-accessibility in the single-cell regulatory landscape. We applied *Cicero* to our hippocampus sci-ATAC-seq dataset to produce 487,156 links between ATAC-seq peaks at a co-accessibility score cutoff of 0.1 (Supplemental Data – InVivo.cicero_links.txt). Of these, 47,498 (10.5%) were links between two promoters, 146,818 (32.4%) linked a distal regulatory element to a promoter, and 259,236 (57.2%) were between two distal elements. We next compared our *Cicero*-linked peaks with existing chromatin conformation data that had been produced on mouse cortical tissue (Dixon et al., 2012), as no hippocampus data sets are currently available; however, a majority of topological associated domains (TADs) are conserved across cell types (Dixon et al., 2012). Consistent with expectations, we observed a 1.1 to 1.5 fold enrichment (Figure 3.4A, $p < 1x10^{-4}$ across all *Cicero*

link thresholds out to 500 kbp, Methods) for linked peaks that occur within the same TAD over equidistant peaks present in different TADs, suggesting that the identified links are associated with higher-order chromatin structure. We then identified cis-co-accessibility networks (CCANs) using *Cicero* which employs a Louvain-based clustering algorithm, which can inform us about co-regulated chromatin hubs in the genome. Using a co-accessibility score threshold of 0.15 (based on high intra-TAD enrichment, Figure 3.4A), we identified 3,243 CCANs, which incorporated 102,736 sites (mean 31.7 peaks/CCAN).

To identify the enrichment of cell-type-specificity of CCANs, we aggregated ATAC-seq signal within each CCAN for each cell type and performed a z-score normalization (Supplemental Figure 20). We then projected the CCANs in 2d space using t-SNE and visualized them based on their enrichment to their highest matching cell type (Figure 3.4B,C, Supplemental Figure 21). This revealed distinct sets of co-accessibility networks for each cell type, with common networks falling towards the center of the projection space. CCANs with greater numbers of peaks tended to be less cell type specific, likely due to the large number of genes that are encompassed by the CCAN, the majority of which are not cell type specific (Supplemental Figure 22). This observation is also consistent with chromatin conformation literature (Dixon et al., 2012) (Supplemental Figure 23). We probed our cell type specific CCANs further by assessing networks that incorporated marker gene promoters. *Prox1* (dentate granule marker), was present in a CCAN that included 89 total accessibility sites and was associated with the correct cell type (Figure 3.4D,E). While much of the CCAN did not exhibit cell type specificity, the region centered on *Prox1* (with the highest co-accessibility values) drove the assignment. To dissect out the major components of the larger CCAN, we used *Cicero* specifically on the dentate granule cells (Supplemental Figure 24A). This revealed three distinct CCANs within the region, with the *Prox1*-containing CCAN exhibiting the greatest specificity to the dentate granule cells (Supplemental Figure 24B). This suggests the possibility of larger chromatin networks with subsets of regulatory elements and genes joining or leaving the network based on cell type. Finally, we identified a number of CCANs that were

overlapping that included mutually exclusive sets of peaks, suggesting two alternative folding patterns of chromatin within the regions dependent upon the cell type (Supplemental Figure 25).

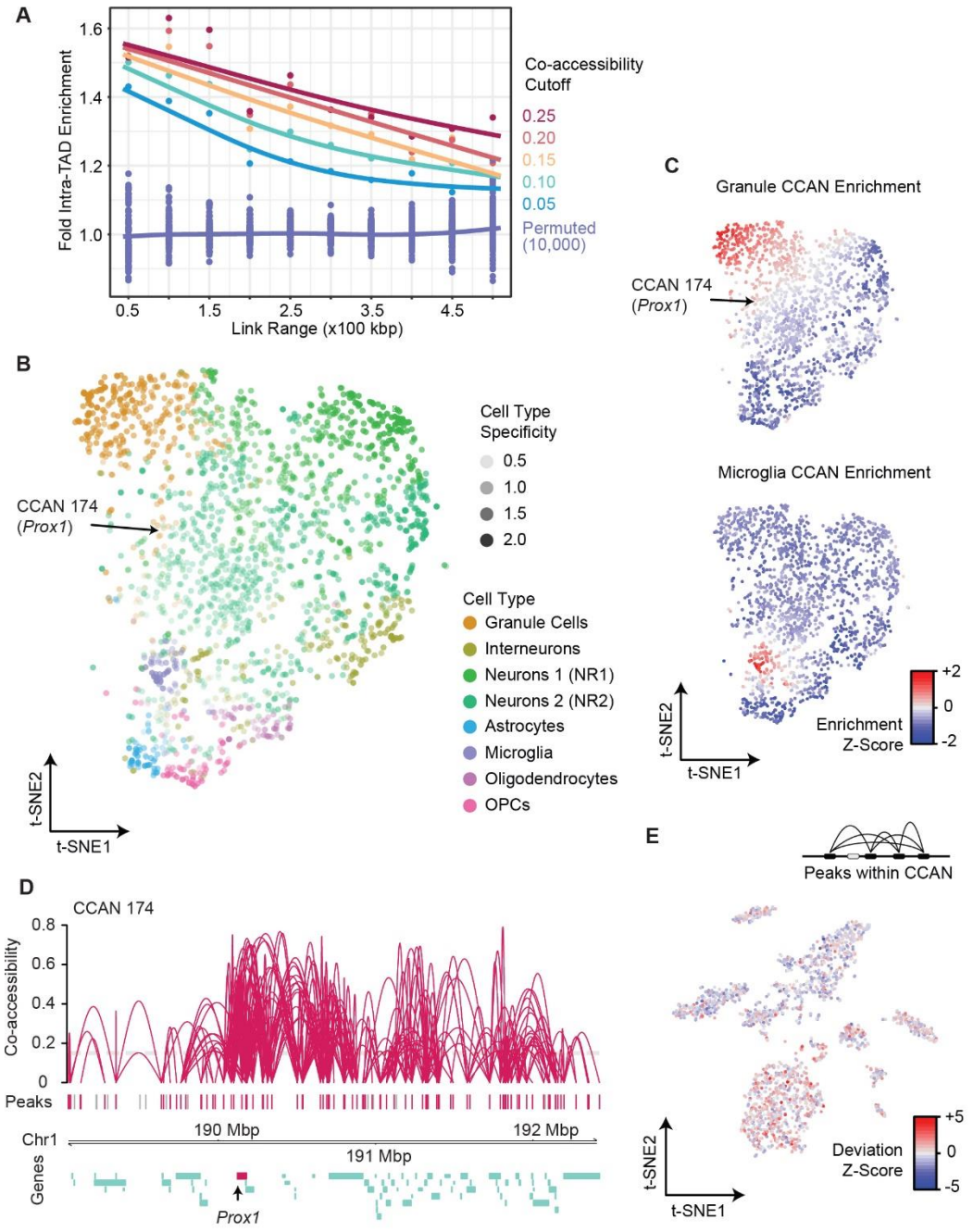

Figure 3.4 Cis co-accessibility analysis using *Cicero*. (A) *Cicero* links at several co-accessibility score thresholds are heavily enriched for links that contain peaks present in the same topological associated domain (TAD) as determined by Hi-C methods (Dixon et al., 2012). The enrichment decreases at greater distances (x-axis). (B) t-SNE projection of CCANs colored by the cell type with the greatest accessibility for the CCAN. Each point represents an individual CCAN. Networks generally group by cell type. CCAN 174 which includes the *Prox1* gene shown below in (D) is indicated with an arrow. (C) Accessibility z-scores for CCANs for granule cells and microglia. (D)

Cis co-accessibility network (CCAN) ID 174 including the *Prox1* promoter (dentate granule marker gene). (E) CCAN 174 has the greatest accessibility signal in cells identified as dentate granule cells.

3.3.6 *In vitro* neurons exhibit an altered epigenetic profile

To examine how well *in vitro* cultured hippocampal neuronal populations match their *in vivo* counterparts at the epigenetic level, we isolated hippocampal neurons from P0 pups and allowed them to mature for 16-18 days *in vitro* (DIV). At this stage, the neurons had extended long processes and expressed markers of mature neurons such as MAP2. We performed sci-ATAC-seq as described above and produced 899 high-quality single-cell chromatin accessibility profiles passing our quality thresholds (Methods). Our mean unique read count per cell was again high when compared to currently published work at 43,532. We then performed peak calling on the ensemble of *in vitro* sci-ATAC-seq profiles, resulting in 111,005 total peaks. Similar to our *in vivo* preparations, the ATAC-seq signal correlated well between the two replicates (Pearson R > 0.99). Subsequent filtering, LSI-t-SNE, and clustering, as described for the *in vivo* preparation, revealed four distinct populations (Figure 3.5A). Upon examination via marker gene and DNA binding motif accessibility enrichment, we determined one of the clusters to be the interneuron population (40.6% of cells), with the remainder being excitatory (59.4%).

We performed peak calling on the combined reads from both the *in vivo* and *in vitro* experiments and merged these peaks with those called on each set individually to produce a combined peak call set comprised of 174,503 sites. It is important to note that much of the increase over the *in vivo* peak set was due to increased coverage at sites that may not have met the calling threshold as opposed to peaks exclusive to the *in vitro* cultured neurons. We then performed LSI and t-SNE on the resulting counts matrix using cells produced in both experiments. While the *in vitro* cultured glutamatergic neurons largely formed their own grouping independent of their *in vivo* counterparts, the inhibitory neurons from the *in vitro* preparation grouped more closely with the *in vivo* population (Figure 3.5B).

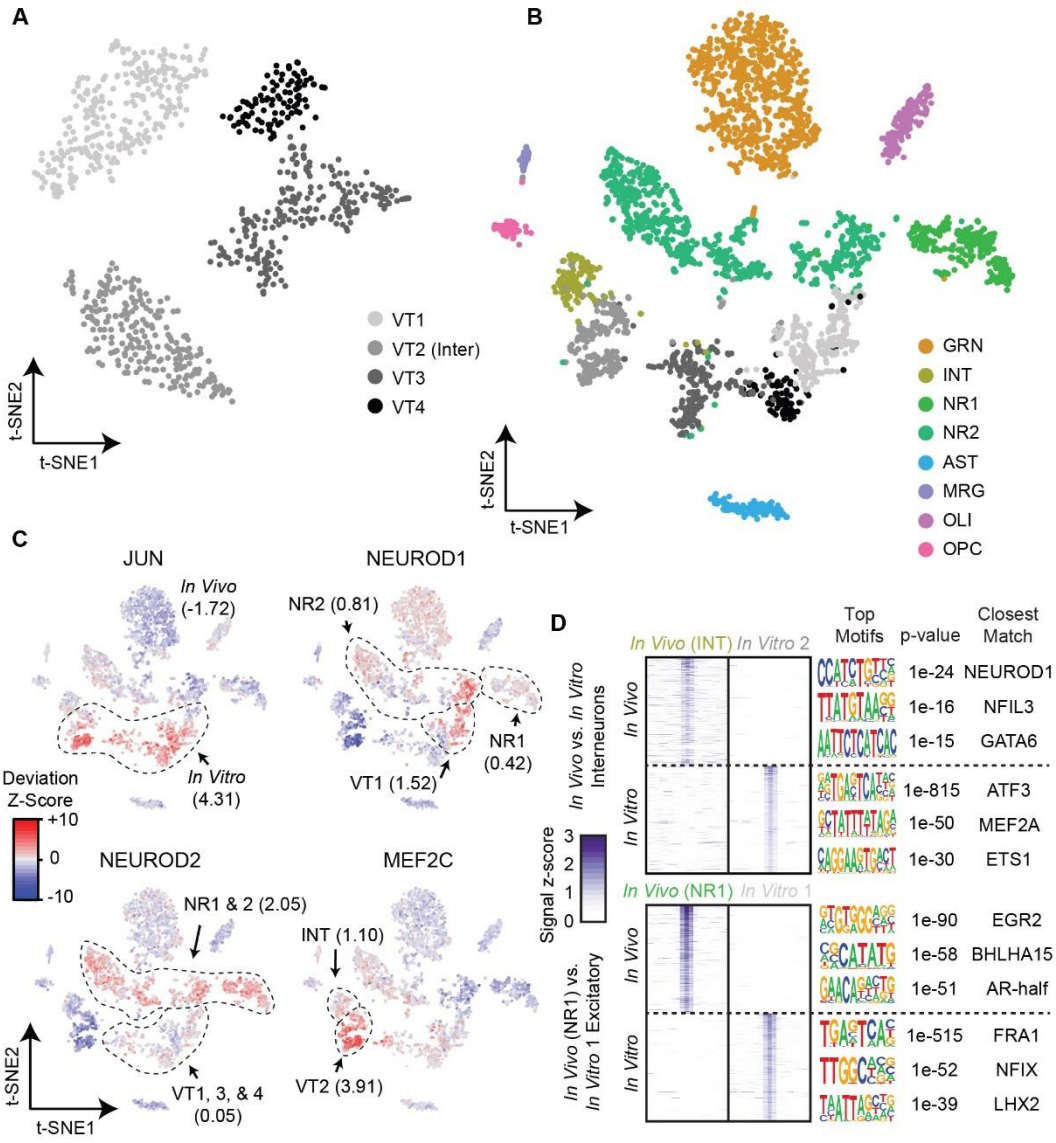

Figure 3.5 Comparison of the accessible chromatin landscape of *in vitro* cultured neurons with *in vivo* obtained profiles. (A) LSI-t-SNE projections of *in vitro* obtained cells reveals four clusters, one of which exhibits interneuron patterns (VT2) and the remaining excitatory neurons (VT1,3-4). (B) LSI-t-SNE projection of the combined *in vivo* and *in vitro* datasets colored by independently called clusters. Excitatory neurons in the two conditions generally cluster separately, with interneurons more closely associated. (C) *ChromVAR* global motif deviation z-scores for select motifs for each cell. Dashed lines and values correspond to mean values of cell populations. (D) Differential accessibility analysis between *in vivo* and *in vitro* interneurons (top, INT vs. VT2, respectively) and between two closest excitatory neuron populations between *in vivo* and *in vitro* conditions (NR1 and VT1, respectively). ATAC-seq signal is shown for the top differentially accessible loci with the top three motifs and corresponding *p*-values and matching motifs to the right.

We next examined the global DNA binding motif accessibility of the combined population (Figure 3.5C). The starkest differences between the *in vivo* and *in vitro* cell populations was in motifs associated with the AP-1 complex, *i.e.* FOS, JUN, ATF, and JDP families (μ_z = 4.32 and -1.72 for *in vitro* and *in vivo* respectively). The AP-1 complex plays a major role in stimulus response, including cell stress (Hess, 2004), which may not be surprising for neurons grown and matured *ex vivo*. It has also been shown that AP-1 modulates chromatin during neuronal activation (Su et al., 2017), suggesting the possibility of an elevated activity state in neuronal cultures compared to their *in vivo* counterparts; however, the decoupling of the many functional roles of the AP-1 complex from one another using global accessibility is not currently possible. We also examined the motifs for several other transcription factors that are relevant to neuronal development. NEUROD1, discussed above, responsible for early differentiation (E14.5 ventricular proliferative zone) (Pataskar et al., 2016) and survival of neurons, exhibited shared accessibility enrichment in a subset of cells from both the *in vivo* and *in vitro* neuronal populations. MEF2C delineates early precursors of a subset of inhibitory interneurons (Mayer et al., 2018) and we observed shared, elevated MEF2C accessibility in the interneuron populations, with greater accessibility in the *in vitro* cells (μ_z = 3.91) over that of the *in vivo* interneurons (μ_z = 1.10). In contrast to NEUROD1 and MEF2C, NEUROD2 acts later in hippocampal development than NEUROD1 (Pleasure et al., 2000), is expressed in migrating granule neurons, and binds to a number of neuron-specific promoters. The DNA binding motif for NEUROD2 was globally more accessible in the *in vivo* neurons when compared to their *in vitro* counterpart (μ_z = 2.05 and μ_z = 0.05 for *in vivo and in vitro* respectively). This finding may reflect its later developmental appearance and that the main targets of NEUROD2 are involved in layer-specific differentiation and axonal pathfinding, which are not likely to be occurring *in vitro*.

Differential accessibility analysis comparing *in vitro* and *in vivo* counterparts shed further light on the epigenetic differences between the two populations (Figure 3.5D). A comparison of the interneuron populations produced 4,356 and 7,575 peaks significantly differentially accessible in

the *in vivo* (INT) and *in vitro* (VT2) populations, respectively (q-value ≤ 0.01, Log_2 fold-change ≥ 1). A motif enrichment analysis of these peak sets revealed the most significantly enriched motifs corresponded to NEUROD1 in the *in vivo* peaks ($p = 1\times10^{-24}$), which is interesting because NEUROD1 global accessibility is low in both interneuron populations (Figure 3.4C). Interneuron peaks specific to the *in vitro* population were significantly enriched for ATF3 ($p = 1\times10^{-815}$), which is not surprising in light of the above accessibility of AP-1 in the *in vitro* cell populations and its shared role in cell stress and interaction with the AP-1 complex (Hai & Curran, 1991). We also examined differential accessibility between the most-closely grouped excitatory neuronal populations, which produced 1,761 and 2,964 for NR1 (*in vivo*) and VT1 (*in vitro*) respectively (q-value ≤ 0.01, Log_2 fold-change ≥ 1). The most significantly enriched motif in the *in vivo* peak set was EGR2 ($p = 1\times10^{-90}$), again a transcription factor expressed highly in migrating neural crest cells (Wilkinson et al., 1989) that may be absent in an *in vitro* setting where cell migration is not pertinent.

3.4 Discussion

A better understanding of the role of specific cell populations in hippocampal function is a necessary step in order to study disease processes that involve this region critical to memory and learning. Thus far, studies have used gene expression data from sorted populations (Cembrowski et al., 2016) and single cells (Habib et al., 2017; A. Zeisel et al., 2015) to identify subpopulations of cells and novel marker genes for the cells within the hippocampus. Here, we provide the most in-depth epigenetic analysis of the hippocampus at single-cell resolution to date. Our sci-ATAC-seq protocol (Methods) has been optimized for primary cell culture and both fresh or frozen tissue and produces unique read counts per cell in the tens-of-thousands, a full order-of-magnitude improvement over the initial sci-ATAC-seq publication (D. a Cusanovich et al., 2015). The data sets released with this study can be readily analyzed using *scitools* (https://github.com/adeylab). This tool suite is designed to be complementary to other single-cell ATAC-seq analysis packages,

such as *ChromVAR*, *cisTopic*, and *Cicero*, and serves as an easy framework for integrating analyses and generating plots to assess data quality and facilitate biological interpretation.

We utilized our sci-ATAC-seq maps to identify the major cell types of the hippocampus, with sufficient depth and library complexity to profile less abundant cell types, such as microglia and oligodendrocyte progenitor cells. Using the recently-described *cisTopic* analysis tool, we were able to achieve a high degree of granularity within pyramidal neuron population, enabling the definitive identification of CA1 and CA3 neurons, a population of putative CA2 neurons, and three lower-abundance populations, likely containing mossy cells and two unidentified neuronal subtypes. Our analysis of global motif accessibility revealed the expected enrichment of motifs associated with specific cell populations in addition to uncovering unanticipated findings, such as increased accessibility at CTCF motifs in interneuron and glial populations, a finding that was also observed in our differential accessibility analysis. We utilized our dataset to map cis co-accessibility networks, enabling the association of distal elements with promoters or other regulatory loci. Finally, we directly compared the accessibility profiles of neurons that were matured *in vitro* with their *in vivo* counterparts to identify altered pathways or chromatin state configurations that should be considered for future experimental design. This revealed a stark difference in the global accessibility for motifs associated with the AP-1 complex, which is involved in cell stress as well as neuronal activity. Future work to identify the cause and effect of elevated AP-1 complex activity is warranted to understand its impact on studies that utilize hippocampal neurons matured *in vitro*.

We believe that the chromatin accessibility maps we provide in this work, including the profiling of *in vitro* cultured neurons, and the software tools we are releasing are a valuable resource for any groups studying the hippocampus or those that wish to analyze single-cell chromatin accessibility data. Our maps complement existing single-cell transcriptional data, and take the field one step closer to a comprehensive atlas of the mammalian hippocampus; however, we

acknowledge that future innovation built off of the datasets we and others have produced will be required to achieve that goal.

3.5 Methods

3.5.1 Isolation of hippocampus tissue

All animal studies were approved by the Oregon Health and Science University Institutional Animal Care and Use Committee. Sixty day old C57BL/6J mice were deeply anesthetized using isofluorane. After decapitation the brain was removed and the total hippocampus was isolated and placed in ice-cold phosphate-buffered saline (pH 7.4).

3.5.2 In Vitro culturing of hippocampal neurons

Pups (P0) were killed by decapitation and the brains dissected in ice-cold Hanks Basal Salt Solution (HBSS, pH 7.4) with 25 mM Hepes buffer. Individual hippocampi were excised without the meninges and pooled by individual animal. The tissue was treated with 2% papain and 80ng/ml Dnase I in HBSS at 37 °C for 10 min. Tissue pieces were rinsed three times with Hibernate A containing 2mM Glutamax and 1x B27 supplement. Neurons were dissociated carefully and filtered with a 0.4-µm mesh. Neurons were plated at a density of $1x10^6$ cells per well of a six well dish coated with 50 µg/mL Poly-L-Lysine hydrobromide in boric acid buffer (50 mM Boric Acid, 12.5 mM Sodium Borate, decahydrate). The neurons were plated in Neurobasal A containing 1xB27 supplement and 2mM glutamax. After 2 hours, the media was changed to remove cell debris. Media half changes occurred every 3 days with fresh Neurobasal A containing 1xB27 and 2mM glutamax. Cells were maintained at 37°C with 5% CO_2 in a humidified incubator.

3.5.3 Sci-ATAC-seq assay & sequencing

Tissue was diced on ice using a sterile razor blade in freshly-prepared Nuclei Isolation Buffer (NIB: 500 µL 10 mM Tris-HCl pH .5, 100 µL 10 mM NaCl, 150 µL $MgCl_2$, 500 µL 0.1% Igepal, 0.1% Tween, 1 unit Qiagen Protease Inhibitor, nuclease-free water to 50 mL) followed by dounce homogenization. For cultured cells, nuclei were directly isolated by removing media, washing once

with ice cold PBS, and then NIB added to cover the dish followed by incubation on ice for 5 minutes, scraping using a tissue scraper, and then an additional 5-minute incubation on ice. For both tissue and cultured cells, nuclei were then pelleted and resuspended in 1 mL NIB with DAPI added to a final concentration of 5 mg/mL. Nuclei were then strained in a 35 µm strainer and sorted on a Sony SH800 Flow Sorter and deposited into 0.2 mL PCR plates containing 5 uL of 2X TD buffer and 5 uL of NIB, with 2,500 nuclei deposited per well. Plates were placed on ice until transposition. Tagmentation was performed by the addition of 1 µL of 2.5 µM barcoded transposome (EZ-Tn5 variant) (Amini et al., 2014b) and incubated at 55°C for 15 minutes followed by placing the plate on ice to stop the reaction. All wells were then pooled using wide-bore pipette tips and DAPI added to a final concentration of 5 mg/mL. Tagmented nuclei were then strained and sorted again and 22 were deposited into each new PCR well containing 0.25 µL 20 mg/mL BSA, 0.5 µL 1% SDS, 7.75 µL nuclease-free water, 2.5 µL barcoded forward primer, and 2.5 µL reverse primer. Plates were kept on ice until all sorting was completed. After sorting, plates were incubated at 55°C for 15 minutes to denature the transposase followed by placing the plate on ice and adding 12 uL of PCR mix (7.5 µL NPM, 4 µL nuclease-free water, 0.5 µL 100X SYBR Green) and then PCR amplified using the following conditions: 72°C for 5:00, 98°C for 0:30, Cycles of [98°C for 0:10, 63°C for 0:30, 72°C for 1:00, plate read, 72°C for 0:10] on a BioRad CFX real time thermocycler. Reactions were pulled when mid-exponential, typically 17-22 cycles. Post-amplification 5 µL of each reaction was pooled and cleaned up using a QIAquick PCR Purification column. Libraries were quantified using a Qubit fluorimeter, diluted to ~4 ng/µL and assessed on an Agilent Bioanalyzer HS Chip. Sequencing was carried out as previously described on a NextSeq™ 500 (research use only) using custom primers and chemistry (Vitak et al., 2017a). A detailed sci-ATAC-seq protocol is provided as a Supplemental Protocol.

For fresh replicates, nuclei were divided into two transposase plates that were processed separately. Each transposase plate was then pooled and the nuclei sorted into a full PCR plate for each preparation. The frozen hippocampi were processed using one half of a transposase plate for

each biological replicate and then all wells were pooled and sorted into a single PCR plate. We also had two biological replicates for the in vitro preparations that were processed according to the same workflow as the two frozen samples.

3.5.4 The scitools suite

All initial analysis was performed with *scitools,* a custom software package we developed to help analyze sci-ATAC-seq data and other combinatorial indexing data (sci-). The toolset is a collection of commands to perform common functions for sci- datasets, including wrappers that utilize existing tools, including: *BWA* (H. Li & Durbin, 2009), *MACS2* (Yong Zhang et al., 2008), *BEDtools* (Quinlan & Hall, 2010), *SAMtools*, as well as R (Core Team, 2019) libraries: *ggplot2* (Wickham, 2016), *chromVAR* (Schep et al., 2017), *chromVARmotifs*, *cicero* (Pliner et al., 2018a), *RtSNE*, *dbscan* (Ester et al., 1996). Usage of *scitools* for any of these functions should cite the relevant utilities. Scitools can be found at https://github.com/adeylab/scitools (an evolving tool) or as Supplemental Code for the version used at the time of this manuscript.

3.5.5 Sci-ATAC-seq data processing

BCL files were first converted to FASTQ files using bcl2fastq (2.19.0). We then demultiplexed our reads using scitools (fastq-dump, fastq-split) based on the two separate Tn5 tagmentation events on the P5 and P7 ends of the molecules and the following added unique PCR indexes on both sides. In order for a barcode to be considered a match each of these four indexes constituting a barcode had to be within two Hamming edit distances away from their expected counterpart. We aligned to the mm10 genome using the scitools fastq-align function within scitools, which mapped reads using *BWA-MEM.* Aligned reads were filtered based on a quality score cutoff of 10 and PCR duplicates removed in a barcode-aware manner using scitools bam-rmdup. We determined whether a barcode represented a cell as opposed it representing noise by using the mixed model approach previously presented (Vitak et al., 2017a). Peaks were then called using

scitools callpeak, which utilizes *MACS2* to identify peaks and then extend to 500 bp followed by peak merging and filtering of peaks that extend beyond chromosome boundaries.

3.5.6 Latent semantic indexing and 2D embedding

Count matrices were generated using scitools counts to produce a matrix of read counts at cells (columns) by called peaks (rows). This matrix was then filtered using scitools filter-matrix to exclude rows with fewer than 10 cells having reads (-R 10), and columns (cells) with fewer than 1000 rows with reads (-C 1000). The matrix was then carried through term-frequency inverse-document-frequency transformation using scitools tfidf, followed by latent semantic indexing, retaining SVD dimensions 1-15 using scitools lsi. The resulting LSI matrix was used in scitools t-SNE which makes use of the *RtSNE* R package. All t-SNE plots were generated using scitools plot-dims using an annotation file to encode cluster ID, sample ID, or other variables, including *chromVAR* motif deviation z-scores.

3.5.7 Co-embedding of single-cell RNA-seq cells with sci-ATAC-seq cells

We utilized *Cicero* (Pliner et al., 2018a) to produce gene activity scores, based on the chromatin accessibility signal at the promoter and linked distal elements to each gene. These scores were loaded into *Seurat3* (Stuart et al., 2018) along with the gene read count matrices from Zeisel et al. 2015 (Smart-seq), and Habib et al. 2017 (DroNc-seq). We then carried out anchor identification and integration of the three datasets as decribed in Stuart et al. 2018. We then performed PCA and t-SNE on the integrated data. Clusters were identified using *PhenoGraph* (Levine et al., 2015) on the PCA dimensions.

3.5.8 Identifying transcription-factor-associated changes

We applied the *chromVAR* (Schep et al., 2017) R package to our data to infer changes in global motif accessibility across our cell populations. This provides information on the putative binding of transcription-factors and consequently the possible ongoing biological processes in cell populations. The mouse_pwms_v1 motif set from the *chromVARmotifs R* package was used in this

analysis. The bias corrected motif deviation scores were plotted on the t-SNE embedded 2D coordinates with the *scitools* plot-dims -M option for visualization.

3.5.9 Cell type dependent differential accessibility

To accurately identify differentially accessible peaks we used the *make_glasso_cds* function from the *Cicero* (ver=,0.0.0.9000) package to create clusters of k=50 cells based on their the low dimensional t-SNE coordinates. We then selected clusters with 99% cell type purity and aggregated accessibility profiles. We posited that the aggregate profiles would provide the replicates required for the *DESeq2* R package, which in turn internally corrects for technical biases such as assay efficiency. With this method we tested (using the inherent *nBinomWaldTest*) for differentially accessible sites between cell types against all other cell types combined. We corrected for multiple testing at q=0.01 and further filtered differentially accessible sites by removing peaks accessible at q=0.2 in any of the other cell types. We also note that scitools aggregate-cells is also capable of aggregating cells in reduced dimensional space for purposes of differential accessibility analysis. We then applied HOMER (Heinz et al., 2010b) (http://homer.ucsd.edu/homer/motif/) to identify potential *de novo* and known regulators of chromatin accessibility within the cell type dependent differentially accessible sites. We used all accessible peaks as background and the mm10 findMotifsGenome command.

3.5.10 Subclustering of pyramidal neurons

We applied *Cistopic* ver=0.2.0 (Bravo González-Blas et al., 2018) to separate out sub populations within the in-vivo neuronal cell populations we found (NR1, NR2). We chose the optimal number of topics (30, Supplemental Figure 12) by running several models ranging from 5 to 50 topics and picking the model with the highest log-likelihood in the last iteration. We used the 250 burn-in iterations and 300 recording iterations for this analysis. We determined topic associated regions via topic binarization with GammaFit (included in *Cistopic*) on the region-topics distributions matrix (thrP=0.975). We then projected the neuronal cells into two-dimensional space

via Uniform Manifold Approximation and Projection (UMAP) (Becht et al., 2018b) on the topics-cell distributions matrix and observed four distinct cell groupings. We identified these clusters with the *Rphenograph* ver=0.99.1 (Levine et al., 2015) package on the topics-cell distributions matrix (d=4 clusters of k=150). The same processing and parameters were used to perform subclustering on the cluster exhibiting high heterogeneity.

To correctly characterize these four clusters, we called potential *de novo* and known regulators of chromatin accessibility with HOMER (run with the mm10 genome and all sites as background using the *findMotifsGenome* command) on the top associated regions of topics that were enriched in individual clusters (identified via the topics-cell distributions matrix). In addition, we called differentially accessible sites unique to each of the clusters using *DESeq2* (as in Cell type dependent differential accessibility methods section) and again applied HOMER for motif enrichment for these sites.

3.5.11 Identifying cis-regulatory networks in the hippocampus

We used the recently described *Cicero* package (Pliner et al., 2018a) to identify cis-co-accessibility networks (CCANs) according to the recommended workflow. For CCAN identification, we used a p=0.15 threshold cutoff, which identified 2,066 chromatin networks that incorporated 47,805 sites of our *in vivo* cell populations. Fold enrichment for links within annotated TADs (Dixon et al., 2012) was performed by calculating the proportion of distance-matched (±25 kbp of specified 50 kbp distance interval) intra-TAD links over inter-TAD links at a range of co-accessibility score cutoffs (0.05 to 0.25 at 0.05 intervals). 10,000 permutations were then performed for each distance bin by randomly assigning two distance-matched peaks as linked and retaining the same total number of links for each co-accessibility cutoff and then calculating the fold intra-TAD enrichment as described above.

3.5.12 Cell type specific cis-regulatory networks

To assign cis-co-accessibility networks to cell types, we first calculated the fraction of cells of each cell type that have signal at a peak and assumed that the distribution of reads per cell across cell types is close to uniform. We then z-scored the resulting matrix across the CCANs and then visualized the separation of CCANs by cell type by bi-clustering and plotting the heatmap using the *complexHeatmap* (ver=1.17.1) R package. We also visualized CCAN cell type specificity by using t-SNE on the z-scored group read fractions to embed CCANs in 2D. We assigned the cell type to each of the CCANs based on the highest z-scored value. We next identified CCANs that contain at least one of the genes (*Prox1, Dsp, Ociad2, Dkk3, Glul, Gfap, Mog, Cldn11, C1qa, Wfs1, Mobp, Pdgfra*) shown to be differentially accessible in our data. We intersected +/-80 kbp regions before and after transcription start sites of these genes with the CCANs using *BEDtools* intersect. We plotted the CCANs around genes where the cell type assigned to the CCANs matched the cell type specificity of the gene using the *Cicero plot_connections* function. We used *chromVAR* to further validate the relative enrichment of CCANs by using CCAN peaks as motif input files. We used *scitools* plot dims -M option to visualize the deviation scores for the CCANs on the t-SNE coordinates. We have to note that in order for this method to work, peaks within the CCANs had to be accessible across multiple cell types, so we decided to use only CCANs with ≥ 10 peaks for this analysis. We finally included a more in-depth analysis of CCAN 174 centered around *Prox1*. We called CCANs just within Granule cells and identified three different sub CCANs, with the core of the original CCAN 174 showing even higher specify in the *chromVAR* deviation scores plots (Supplemental Figure 24).

Chapter 4: Integrated single-cell analysis reveals treatment-induced epigenetic homogenization

Kristóf A. Törkenczy*, Ellen M. Langer*, Andrew J. Fields, Megan A. Turnidge, Andrew Nishida, Christopher Boniface, Paul T. Spellman, Emek Demir, Joe W. Gray, Rosalie C. Sears1, Andrew C. Adey

* These authors contributed equally to this work

This chapter has been reformatted for inclusion for this dissertation from the manuscript titled: "Integrated single-cell analysis reveals treatment-induced epigenetic homogenization" currently in review.

Supplementary material will be made available to the committee alongside this dissertation.

4.1 Abstract

Triple negative breast cancers (TNBC) constitute one-sixth of invasive female breast cancer cases and are the most likely to develop resistance to treatment via genetic and/or epigenetic adaptation into drug tolerant persister (DTP) states. We applied single-cell ATAC-seq and RNA-seq to characterize the dynamic regulatory and transcriptional landscape in five basal-like TNBC cell lines in response to the MEK inhibitor Trametinib. We observed surprisingly few shared changes between lines, indicating substantial heterogeneity in the emergence of DTP states. However, we identified a shift toward a common state based on the novel observation of the preferential loss of cell line-specific regulatory elements and gene expression. Integration of the two modalities enabled a granular dissection of dynamic regulatory mechanisms, which revealed highly context-dependent roles of regulatory elements. This work highlights the heterogeneity of response, yet suggests homogenization occurs in the form of the preferential loss of epigenetic configurations unique to each BCCL.

4.2 Introduction

Normal functional mammary glands require development of differentiated luminal and basal epithelial cells from a multipotent progenitor. Development of these specialized cells is mediated by epigenetic changes during embryogenesis, puberty, and pregnancy to support the functionality of the developing gland (dos Santos et al., 2015; Gascard et al., 2015a; Lien et al., 2011; Mikkelsen et al., 2010; Pellacani et al., 2016). While a multipotent progenitor is evident during embryogenesis, the multipotent nature of postnatal mammary stem cells (MaSC) in the normal gland remains controversial (Lloyd-Lewis et al., 2017). However, upon transplantation, injury, or tumor initiation, it is clear that plasticity between lineages can exist, and requires epigenetic regulation (E. Lee et al., 2019; Wahl & Spike, 2017).

Phenotypic diversity and plasticity are readily observed in tumors that arise in the mammary gland. Basal-like triple negative breast cancers, in particular, exhibit profound intratumoral cell state heterogeneity (Risom et al., 2018). The plasticity between cell states in these tumors can arise through asymmetric cell division (Almendro et al., 2013; R. Z. Granit et al., 2013) or extrinsic signals or stress, including exposure to therapies (Chaffer et al., 2011; Gupta et al., 2011a; Klevebring et al., 2014; Risom et al., 2018). Plasticity driven by chromatin remodeling in response to therapeutic treatment supports the emergence of drug tolerant persister (DTP) cells, which can survive during treatment in a quiescent or low proliferative state (Lesniak et al., 2013; Liau et al., 2017; Risom et al., 2018; Sharma et al., 2010). DTP cells retain their epigenomic plasticity, and their quiescence can be reversed upon withdrawal of drug, predisposing the patients for recurrence (Risom et al., 2018). Understanding the epigenetic changes underlying plasticity into DTP states could lead to new strategies for prevention of resistance and/or recurrence.

The high cell-state heterogeneity and propensity for cell-state switching of basal-like breast cancer cell lines necessitates single-cell profiling for further understanding of the state transitions during DTP generation. While more traditional bulk methods provide an average of the assayed

cell populations, single-cell approaches can interrogate the heterogeneity of the regulatory landscape by populating it with cells as data points. Single-cell RNA sequencing technologies have recently been at the forefront of understanding cell-state heterogeneity in development (Bach et al., 2017; Pal et al., 2017; Wuidart et al., 2018) and cancer (Brady et al., 2017; Davis et al., 2020; Karaayvaz et al., 2018; C. Kim et al., 2018b; Pervolarakis et al., 2019); however these technologies do not provide information on chromatin regulatory changes that underlie epigenetic state plasticity such as the opening or closing of distal regulatory elements linked to promoters of important cancer genes. Recent single-cell chromatin accessibility analysis of the developing mammary gland indicates that distinct chromatin states are evident in basal and luminal cell populations as early as E18 in development (Chung et al., 2019), but how chromatin states are altered in tumors upon treatment is unknown.

The Assay-for-Transposase-Accessible-Chromatin (ATAC-seq) can map regulatory landscapes through the determination of promoter and enhancer accessibility and the identification of putative DNA binding proteins through motif analysis. ATAC-seq on ensemble cell populations has shown great promise at delineating the epigenetic heterogeneity across primary human cancers and the development of chromatin accessibility profiling technologies in single-cells has helped us elucidate the complex heterogeneity of epigenomic architecture within cells of healthy and diseased tissues (Chung et al., 2019; Davis et al., 2020; Pervolarakis et al., 2019). Of these technologies, single-cell combinatorial indexing ATAC-seq (sci-ATAC-seq), in which library molecules go through two rounds of barcoding (Transposase then PCR) has been applied to a wide range of subjects, including organoid development (Mulqueen et al., 2019), myogenesis (Pliner et al., 2018b), hematopoietic differentiation (Buenrostro et al., 2018) and cell atlases of various tissues (D. A. Cusanovich, Hill, et al., 2018; Sinnamon et al., 2019a). It is uniquely positioned to investigate intratumoral phenotypic state heterogeneity and the underlying plasticity that supports the emergence of DTPs.

We previously demonstrated that the heterogeneity and plasticity of basal-like breast cancers can be modeled in breast cancer cell lines (BCCLs), and showed that targeted therapies inhibiting MEK or PI3K/mTOR pathways generated quiescent DTPs in distinct differentiation states. Importantly, plasticity between cell states in response to therapies occurred through epigenomic transitions (Risom et al., 2018). Here, we profile chromatin accessibility and transcriptional changes at the single-cell level in five basal-like breast cancer cell lines in response to MEK inhibition, which we previously characterized to alter the differentiation state of Basal-Like Breast Cancer (BLBC) (Risom et al., 2018). These lines have diverse genetic backgrounds and distinct baseline state heterogeneities and yet showed similar phenotypic cell state responses to the MEK1 inhibitor Trametinib, enriching for basoluminal cell state markers while de-enriching for mesenchymal cell state markers. We profiled individual cells from each cell line using sci-ATAC-seq and found that, despite similarities in the phenotypic state after treatment with Trametinib, adaptation to the DTP state did not appear to be mediated by shared epigenetic changes across the different cell lines. Instead, we identify a previously-undescribed process of inter-cell line homogenization, in which chromatin sites that are uniquely open or closed in specific cell lines exhibit the greatest change, which shifts in the direction of the other cell lines. These findings are further supported by single-cell analysis of the transcriptional landscape, though to a lesser degree. Use of both modalities enabled us to build an integrative analysis model to decipher adaptive interactions between enhancers and gene expression, revealing how cell line-unique chromatin landscapes homogenize toward a more common DTP state.

4.3 Results

4.3.1 Epigenetic heterogeneity across basal-like TNBC cell lines

We set out to deeply characterize the epigenetic cell states in five basal-like triple negative breast cancer cell lines (MDAMB468, HCC1806, SUM149PT, and two populations of HCC1143) at baseline and in response to MEK inhibitor treatment. The two populations of HCC1143 are distinguished throughout as HCC1143G, the commercially available line, and HCC1143S, a spontaneously generated stable line. The similar origin of these lines was confirmed by STR profiling and detection of the same driver mutations that have been previously described in HCC1143. However, HCC1143S has diverged from HCC1143G, showing a distinct genetic profile and loss of expression of many basal keratins (Figures S1A-S1C). We performed both sci-ATAC-seq and multiplexed single-cell RNA-seq on two independent experiments of the five cell lines treated for 72 hr with Trametinib or DMSO vehicle control (Figure 1A; STAR Methods). Cell line and treatment identity were maintained in both assays during sample multiplexing within the same experiment to minimize batch effects. For scRNA-seq, our samples were indexed with hashtag antibodies prior to library generation and sequencing, and for ATAC-seq, samples were multiplexed during the transposase indexing stage of combinatorial indexing. We processed sequencing data using established workflows to produce a gene × cell counts matrix for scRNA-seq and custom software to produce a peak × cell matrix of QC-passing cells for the sci-ATAC-seq dataset (Figures S1D and S1E; STAR Methods) (Sinnamon et al., 2019a).

In order to identify unique and shared epigenetic features across the cell states present in the cell lines at baseline (DMSO), we applied *cisTopic* (Bravo González-Blas et al., 2019), a probabilistic modelling technique based on Latent Dirichlet Allocation to simultaneously identify epigenetic cell states and the sets of co-regulated chromatin regions associated with them (Topics;

Figure 4.1A; STAR Methods). We determined the optimal number of topics to be 30 (Figure S1F) with a mean of 6,057 associated peaks per topic (Figure S1G).

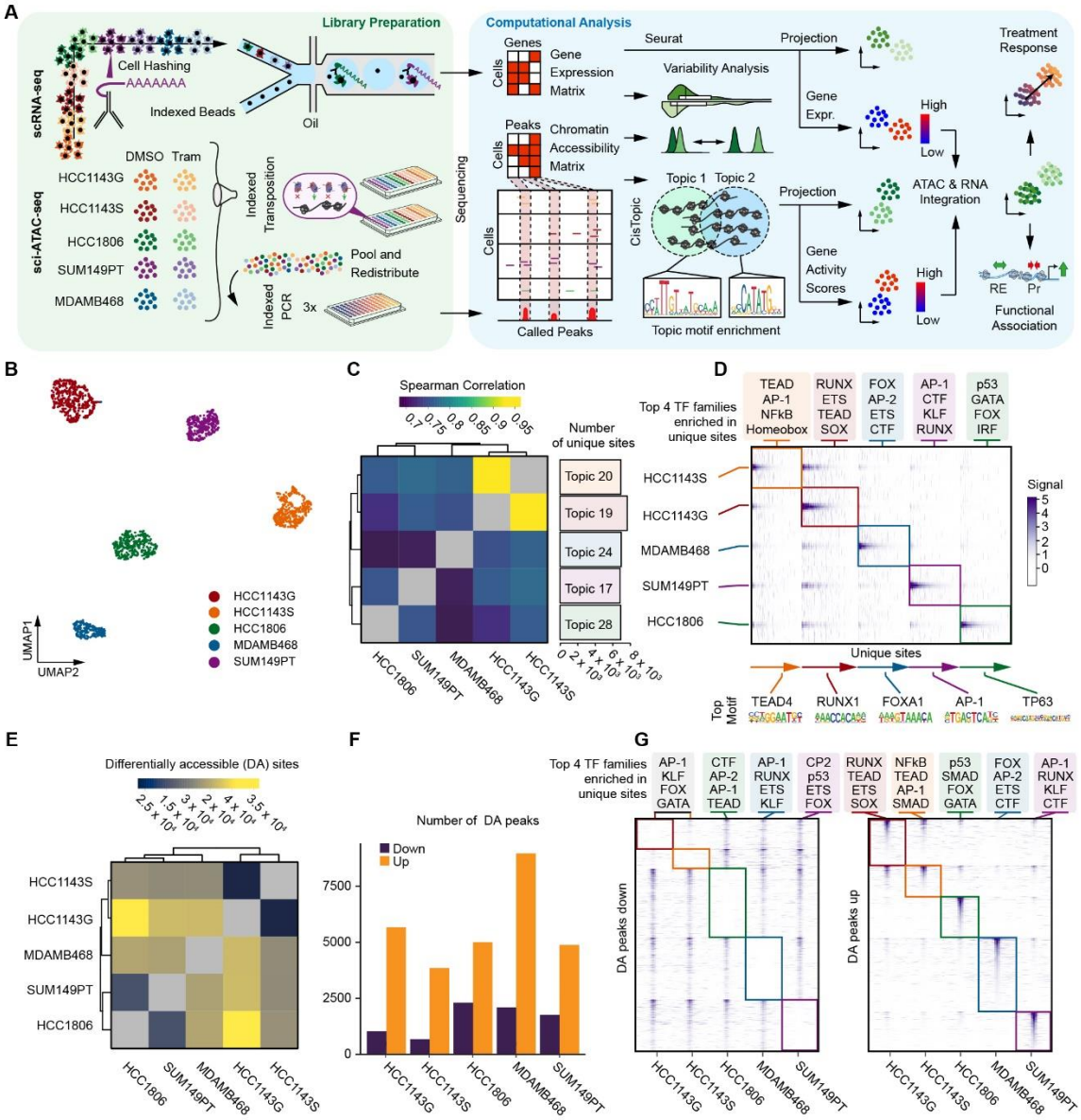

Figure 4.1 Study design and epigenetic state heterogeneity of BCCLs. (A) Overview of experimental and analysis strategy. (B) UMAP projection for the topic space of the five BCCLs assayed via sci-ATAC-seq. (C) Bi-clustering of BCCLs based on average between cell line correlation of drop out corrected site usage. The most line-specific topics are noted along with the number of sites in that topic. (D) Aggregate ATAC-seq signal at the topic-identified line-specific sites from C. Windows are centered on each peak and extend 2,500 bp in each direction. Top motifs

enriched in these sites are noted. (E) Bi-clustering of pairwise differential accessibility (DA) assessment. (F) Total numbers of DA peaks for each line relative to all others. (G) Aggregate ATAC-seq signal at line-specific DA sites. Windows are centered on each peak and extend 2,500 bp in each direction. Top motifs enriched in these sites are noted. One prominent topic (Topic 25) was shared among all cell lines, largely representing constitutive promoter elements and housekeeping genes. The remaining topics were found to primarily show cell line specificity (Figure S1H), which was emphasized when we projected cells into two-dimensional space for visualization by performing Uniform Manifold Approximation (UMAP) (Becht et al., 2018b) on the matrix of topic contributions per cell, revealing distinct separation between the BCCLs (Figure 4.1B). This separation was also observed in correlations between the average predictive probability of the distribution of sites for each cell line (Figure 4.1C; STAR methods). Consistent with their common origin, we observed the two HCC1143 subclones to be closely correlated (Spearman Corr = 0.917) in their epigenetic landscape and least correlated with HCC1806 (average Spearman Corr = 0.753) and MDAMB468 cells (average Spearman Corr = 0.767).

We next characterized the chromatin features most unique to each of the BCCLs by selecting cell line specific topics based on the average contribution score within each of the cell lines (Figure S1I; STAR Methods). We identified the set of associated peaks within each of these topics (Figure 4.1C; cutoff: thrP>0.975; STAR Methods) and verified cell line specificity by calculating the signal of accessibility for these sites across all cells grouped by cell line (Figure 4.1D). Using these cell-line specific peak sets, we then performed motif enrichment (Heinz et al., 2010a) to explore the putative DNA binding proteins driving the differences, which revealed distinct sets for each individual line, with only modest overlap, notably AP-1 and its associated factors. Together, this analysis identifies transcription factors motifs associated with chromatin features/topics that are specific to each of the cell lines (Table S1), many of which have been reported to play roles in breast development and breast cancer differentiation states.

We additionally identified differentially accessible (DA) peaks between each pair of cell lines independent of our topic-based approach (STAR methods). Consistent with the topic analysis, hierarchical clustering of the DA counts placed the HCC1143 lines closest, with HCC1806 and SUM149PT also grouping together (Figure 4.1E). We then identified sites uniquely differentially accessible in each line when compared to all other lines which produced a range of 4,517 to 11,063 differentially accessible loci per line (Figures 4.1F and 4.1G; total of 28,353 uniquely accessible and 7,835 uniquely inaccessible across all lines). These DA sites were highly enriched in the sites associated with the most specific topics to each cell line (Figure S1J, STAR Methods). Furthermore, the grouping of nearby differentially accessible loci enabled us to identify copy number alterations of genes that were previously documented from genome sequencing efforts (Barretina et al., 2012). This included *SMAD4*, deleted in MDAMB468, and *CCND1*, duplicated in the HCC1143 lines (Figure S1K), which supports the use of single-cell ATAC-seq data for the identification of putative copy number alterations, particularly deletions, consistent with a previous report (Satpathy et al., 2019), and provides a potential framework for understanding the interaction of epigenomic regulatory changes with the genomic landscape. Motif enrichment of the uniquely accessible DA loci revealed matching families to the topic-based analysis with the exception of the enrichment of the SMAD motif family in the HCC1143S and HCC1806 BCCLs (Figure 4.1D vs 4.1G, Table S1). Motif families found to be enriched in uniquely inaccessible loci were previously identified in other cell lines within their uniquely accessible loci. For example, the KLF motifs accessible in SUM149PTs were found in the inaccessible loci in HCC1143 and MDAMB468 cells. Taken together, the cell line specific topic and differential accessibility analyses are concordant and suggest distinct motif families are responsible for maintaining the separate epigenetic cell states in these BCCLs. Furthermore, the consistent response profile of these lines across multiple Trametinib exposure experiments (Risom et al., 2018) indicates that the epigenetic response observed is programmed and not stochastic.

4.3.2 Cell line specific chromatin changes upon Trametinib treatment

We previously showed that HCC1143G basal-like breast cancer cells alter their epigenetic state in response to targeted therapies, and that these changes appeared to be essential to resisting cell death (Risom et al., 2018). Here, we sought to characterize the epigenetic state shift that occurs in multiple basal-like breast cancer cell lines under MEK inhibition with Trametinib. To determine the concentration of Trametinib that would induce drug tolerant persister (DTP) cell populations in these cell lines, we evaluated growth rate inhibition metrics to determine sensitivity to Trametinib relative to the DMSO vehicle control (Hafner et al., 2016). The maximum effect of the drug (GR_{max}) revealed a high resistance of HCC1806 to Trametinib, followed by a modest resistance of MDAMB468 as compared to the more sensitive SUM149PT and HCC1143 cell lines (Figure 4.2A). Based on the GR curves, we chose to treat the five cell lines with 1μM Trametinib for 72 hours to induce DTP populations in all cell lines.

Single-cell ATAC-seq profiles for both treated (Tram.) and control (DMSO) were combined and processed using topic-based dimensionality reduction (*cisTopic*; optimal topic count of 54; 3,535 mean peaks associated with each topic; Figure S2A). Visualization in two-dimensional space with UMAP revealed separate cell line specific DTP states after Trametinib treatment (Figures 4.2B). As with the assessment of topics in untreated cells, several topics were identified that were present across all cells, representing constitutively accessible elements that did not change after Trametinib treatment (Topics 52, 54 and 35). Among topics that changed within at least one cell line between DMSO and Trametinib treatment only Topic 15 was shared between two lines, with both SUM149PT and HCC1806 exhibiting slightly decreased accessibility at elements associated with the topic. The remainder were cell line specific (Figures 4.2C and S2B). A comparison of topics identified in the combined dataset vs the control lines alone dataset revealed conserved cell line specific topics (Figure S2C). Additional topics in the combined dataset

associated with Trametinib treatment were also line-specific. As with the untreated analysis we performed differential accessibility and correlation analyses, which placed all Trametinib treated groups closest to their DMSO groups (Figure S3A). The two populations of HCC1143 lines clustered together and were most distinct from the HCC1806 cell line, with SUM149PT and MDAMB468 cells in between.

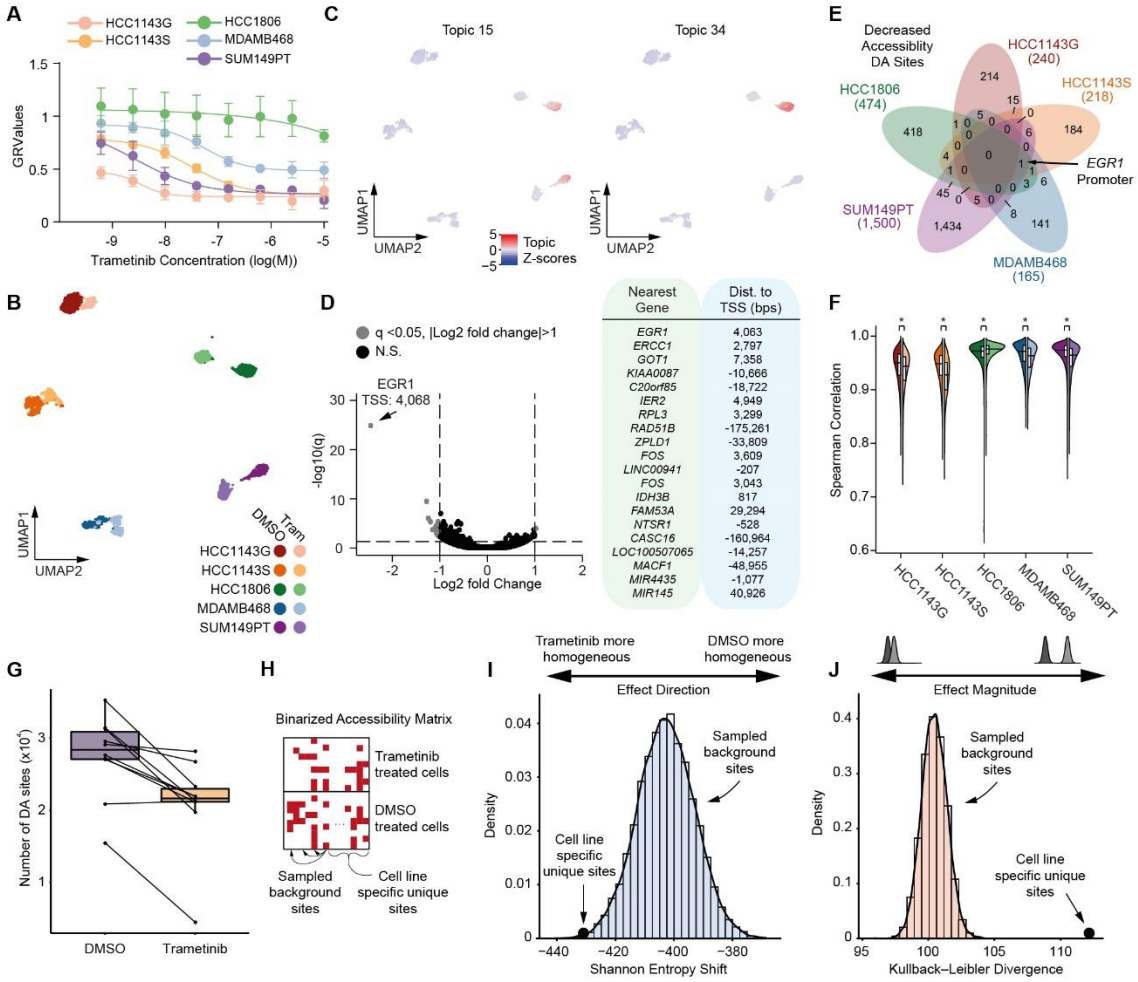

Figure 4.2 Epigenetic state shift and cross-cell line homogenization of BCCLs upon treatment. (A) Growth rate inhibition at increasing concentrations of Trametinib. In each line a DTP state emerges. (B) UMAP projection for the topic space of the five lines assayed via sci-ATAC-seq for all control (DMSO) and drug-exposed (Tram.) conditions. (C) UMAP projection as in (B) but with cells colored by the Z-scored topic probability values shown for two topics. (D) Volcano plot of DA between all control and treated cells with the significant hits listed on the right along with the distance to the nearest TSS. (E) Venn diagram of overlapping hits from within-line DA analysis. Decreasing accessibility sites are shown with the majority in only a single line. The *EGR1* promoter element is significant in all but one line. (F) All-by-all site-based Spearman correlation between cells under control (left) and Trametinib treated (right) conditions. Asterisk indicates a significant difference in distributions (Mann-Whitney U test, $p<0.05$). (G) Comparison in the number of

identified DA sites between individual lines under control (left) and Trametinib treatment (right). Lines indicate pairs of comparisons between the two conditions. (H) Entropy analysis schematic where a binarized matrix is utilized and subsampled for sets of sites to generate a background distribution for comparison with cell line specific sites. (I) Site Shannon Entropy shift for sample background sites (distribution), with the set of CLSs indicated by a point. All sets of sites exhibit a decrease in Shannon Entropy, indicating global homogenization. (J) Site Kullback-Leiber Divergence for sample background sites (distribution), with the set of CLSs indicated by a point. CLSs are an extreme outlier indicating a high magnitude shift.

As a more precise means of identifying shared changes in chromatin accessibility upon Trametinib treatment that may not manifest in the broader topic-based assessment, we next deployed a differential accessibility analysis between Trametinib treated vs control samples using all lines in aggregate. This produced very few significant hits (n=20; Figure 4.2D) and included regions near the promoters (< 5kbp from TSS) of genes that play a role in G1 cycle arrest, decreased proliferation, and apoptosis, all of which are previously demonstrated mechanisms of cellular response to Trametinib *in vitro*. These hits included the loss of accessibility in promoter regions of AP-1 complex members *FOS* and *FOSB* (Zeiser, 2014) (Figures 4.2D and S8) as well as of *NTSR1*, *LINC00941*, *IDH3b*, *GOT1*, and *IER2*, which are involved in cell proliferation, metabolism and migration (Al-Khallaf, 2017; H. Liu et al., 2019; Meléndez-Rodríguez et al., 2019; Neeb et al., 2012; Q. Wu et al., 2019; Younes et al., 2014). The most significant hit was the uniform loss of accessibility near the promoter region of *EGR1*, an important transcriptional regulator and tumor suppressor (M. Yang et al., 2016). These hits match with expectations of a reduced metabolic and proliferation state common to many drug responses; however, it is unclear if these are causal mechanisms or the consequence of the emergence of a DTP state under Trametinib treatment.

Since few common significant differences in accessibility sites were found, we next identified differentially accessible chromatin regions between Tram vs DMSO for each cell line individually, and found a total 2,597 DA peaks (1,249 up and 1,348 down; Figures 4.2E and S3B; Table S2), which varied substantially between the lines. This analysis revealed 0 shared sites between all cell lines (Figure 4.2E), and only a low number of sites that were shared between multiple cell lines (1 shared between 4 lines, 2 shared between 3 lines, and 98 between 2 lines).

These shared loci included the DA sites at *CCND1* (shared by 2 lines), a key regulator of cell proliferation, which has been shown to be directly controlled via EGR1 in glioma cell lines (D.-G. Chen et al., 2017), and *CD24* (shared by 2 lines), which is often used as a marker to determine cell line stemness (Figure S1K, Mani et al., 2008; Polyak and Weinberg, 2009). Similarly, when we examined common regulatory mechanisms, such as opening or closing of regulatory elements proximal to promoters and promoter elements of genes, we found 0 overlapping genes with regulatory elements changing in common across all cell lines (Figure S3C). Even in cases where we observed uniformly changing promoter elements between several lines, we also found multiple other elements in these promoter regions that open or close after treatment that were mutually exclusive between the lines. This was particularly pronounced in the loci proximal to *EGR1*, where dynamic elements were largely discordant between lines (Figure S3D and S3E). Further investigation of the DNA binding motifs at these loci also show little overlap, suggesting different mechanisms of promoter downregulation at the *EGR1* locus. Together, these results indicate cell line specific chromatin responses to Trametinib in the generation of DTP states.

We next prioritized dynamic topics by associating them with the set of DA peaks, which produced distinct mappings between peak sets from the two methods with substantial motif enrichment overlap (Aibar et al., 2017) (STAR Methods; Figure S3F and S3G, Table S3). When we annotated Trametinib enriched peaks associated with topics based on their distance to gene transcription start sites (TSS), only HCC1806 peaks were enriched within 1 kbp of promoter regions. MDAMB468, HCC1143, and SUM149PT regions were enriched more strongly in regions up and downstream of promoters, UTR, intronic and distal intergenic regions. This indicates that epigenetic adaptation to Trametinib is most likely enhancer-driven within the four less resistant cell lines (Figure S3G). HCC1806, in contrast, likely retains the ability to modulate pathways of resistance through promoter usage, which was further supported by pathway enrichment analysis on its Trametinib associated topic (Topic 49), (Gu Z., 2019). This revealed a significant enrichment (q-value < 0.05) of genes with roles in RNA Polymerase I promoter opening and telomere

maintenance and packaging, which could contribute to the cell line's ability to evade drug-induced growth arrest (Figure S4A). We additionally performed an analysis similar to gene set enrichment where we leveraged the deviation in accessibility across sites proximal to sets of genes associated with properties of breast cancer (Figure S4B). Results for HCC1143G were largely consistent with our previous findings using bulk RNA-seq and the same gene sets (Risom et al., 2018). A comparison between lines revealed a dearth of shared pathways that were altered in response to Trametinib exposure across the lines (Figure S4B and S4C), in line with both the topic-based and differential accessibility analysis that support very different response mechanisms for the individual lines.

Finally, we assessed the extent of internal heterogeneity within the lines at baseline and with Trametinib treatment. We correlated the predictive probability distribution of sites within each of the lines, which showed the HCC1143 lines to be significantly more heterogeneous than the other basal-like cell lines (Figure 4.2F; t-test and Mann Whitney U test; Statistics in Table S4). This agreed with the high Shannon index of the HCC1143 line when cell state heterogeneity was assessed based on differentiation state marker staining (Risom et al., 2018). Trametinib treatment was found, in all lines except for HCC1806, to increase the intra-cell line heterogeneity of site accessibility. Since HCC1806s were the most resistant to Trametinib treatment, this could indicate that the cell line is already poised for resistance in its DMSO state. Similar results were found when we assessed the Shannon entropy across all sites within cell lines, with the exception that both HCC1806 and MDAMB468 decreased heterogeneity with treatment (t-test and Mann Whitney u test; Statistics in Table S4).

4.3.3 Preferential homogenization of cell line specific accessible chromatin regions upon Trametinib treatment

Having found that the basal-like BCCLs did not share substantial common chromatin accessibility changes upon drug treatment, we further explored the cell-line specific epigenetic responses to Trametinib. We observed that the number of differentially accessible sites between all combinations of cell lines treated with DMSO was significantly higher (Mann-Whitney U test, W = 20, p= 0.023) than that of Trametinib treated populations (Figure 4.2G). We hypothesized that the decrease in the number of differentially accessible sites between Trametinib treated groups and the lack of shared changes in Trametinib response could indicate a shift towards homogenization between lines at sites unique to each cell line. We defined Cell Line-specific Sites (CLSs), as the set of sites that were significantly more accessible in one line under DMSO control conditions relative to all other lines by utilizing the union of peaks associated with topics specific to each line (mean 6,620 for each line, sum = 33,099; Figures 4.1C and S1I). The sets of sites for each individual line did not exhibit significant enrichment for any shared motifs and were not associated with any shared biological processes as assessed using ontology-based tools (Figures 4.1D and S4; STAR Methods), with the only commonality being the uniqueness to each respective line.

We tested our homogenization hypothesis by calculating the relative shift in Shannon entropy and the Kullback-Leibler (KL) divergence in the CLSs of Trametinib treated cells compared to control cells. Here, Shannon entropy informs on the direction and KL-divergence marks the magnitude of the shift in heterogeneity between BCCLs. We deployed a permutation strategy of random sampling with replacement between iterations on subsets of regulatory elements that equaled the number of CLSs (n=33,099) to establish background levels of entropy shift and KL-divergence between lines (Figure 4.2H, STAR Methods). This revealed a decrease in Shannon Entropy for all sampled sets, indicating a general shift towards homogeneity (Figure 2I). Next, we examined the Shannon Entropy shift for CLSs, which was significantly more negative than the sets

of background sites (p=0.02, given a normal distribution; Figure 4.2I, black circle). The KL divergence for CLSs, a measure of the magnitude of the effect, was a striking outlier, indicating an extreme preference for homogenization at CLSs (Figure 4.2J, p<0.0001). In addition to Topic-defined CLSs, we tested CLSs defined using differential accessibility (Figure 1F, total up regulated sites = 26,595, total down regulated sites= 7,835) and observed the same effect. Notably, these shifts upon treatment occurred in both directions: increased accessibility at CLSs that were uniquely inaccessible, and decreased accessibility at CLSs that were uniquely accessible (p<0.0001 in both directions for Shannon entropy shift and KL divergence; Figures S5A andS5B). This supports the paradigm that there is a general shift towards epigenetic homogeneity across cell lines as a response to treatment with a strong preferential loss of cell line specific sites.

To further understand this phenomenon, and understand whether it affects the accessibility of specific transcription factor binding motifs, we assessed the global change in chromatin accessibility at sets of loci harboring specific DNA binding motifs (Figure S5C; STAR Methods) (Schep et al., 2017). We then collapsed motif families based on similarity and annotations in the csBp database (Figure S5D; STAR Methods) to produce 71 groups of motif accessibility deviation. We then ordered motif groups based on the within-group variance from high to low of Trametinib treated cells and identified six that exhibited homogenization upon Trametinib treatment relative to the DMSO control (STAR Methods). This indicated that the DNA binding proteins recognizing these motifs may play an elevated role in site homogenization upon treatment between cell lines. These motifs consisted primarily of bZip domain, TEA domain and Forkhead domain transcription factors. Specifically, members of the AP-1 complex, which play important roles in regulating proliferation and apoptotic signaling in response to treatment, had a more uniform TF accessibility across Trametinib treated cells (Figure S5D).

4.3.3 Transcriptional changes in response to Trametinib

To understand the transcriptional changes occurring in the five BCCLs following Trametinib treatment and preferential inter-line site homogenization, we performed scRNA-seq using the 10x Chromium platform with antibody hashtags to multiplex the 10 sample conditions. Of note, our initial processing of the scRNA-seq data using *Seurat* (Stuart et al., 2019) produced 11 clusters of cells from our 10 conditions (STAR Methods), with two distinct clusters representing the SUM149PT cell line treated with Trametinib (Figure S6A). The additional SUM149PT cluster exhibited expression patterns that were shared with both the SUM149PT and HCC1143G cell lines (Figure S6B), suggesting that the population may represent cell collisions that were not properly eliminated in the hashtag demultiplexing process. We therefore developed a novel technique for demultiplexing that leverages the Shannon Entropy of hash barcodes associated with each cell, and this technique readily identified the additional cluster as collisions (Figures S6C and S6D, STAR Methods). After filtering to remove these, we reprocessed our dataset and identified 10 distinct clusters, two for each individual line based on treatment condition, much like our sci-ATAC-seq dataset (Figure 4.3A). We first examined the global heterogeneity within each line by computing all-by-all cell-cell distances (Figure 4.3B, Mann-Whitney U test, Table S4, STAR Methods). This revealed a consistent pattern to what was observed at the chromatin accessibility level with a slight shift to increased heterogeneity within most lines with the exception of HCC1806 which decreased in heterogeneity. We next identified differentially expressed (DE) genes between DMSO control and Trametinib treated conditions for each cell line (Table S5), which, similar to chromatin accessibility, we found to be predominantly unique to each line (Figure 4.3C; 58.8% unique to one line, 83.4% in ≤ 2 lines). Five genes were significantly differentially expressed within all five lines, all of which exhibited increased expression under Trametinib treatment. Notably, four of the five genes were basal keratins (*KRT5*, *KRT15*, *KRT16*, *KRT17*; Figure S6E), consistent with our previous work showing increased basal differentiation state marker expression upon Trametinib treatment (Risom et al., 2018). The fifth gene consistently upregulated with Trametinib treatment

was *CD24*. When we compared the cell lines on average, we found some genes downregulated upon Trametinib treatment, such as *CD44* (decreased expression in four of five lines; Figure S16A). Interestingly, the CD44(high); CD24(low) cell population has been classically used as a marker to identify more mesenchymal-like cancer stem cells (Mani et al., 2008; Polyak & Weinberg, 2009), potentially suggesting a more differentiated state with MEK inhibition.

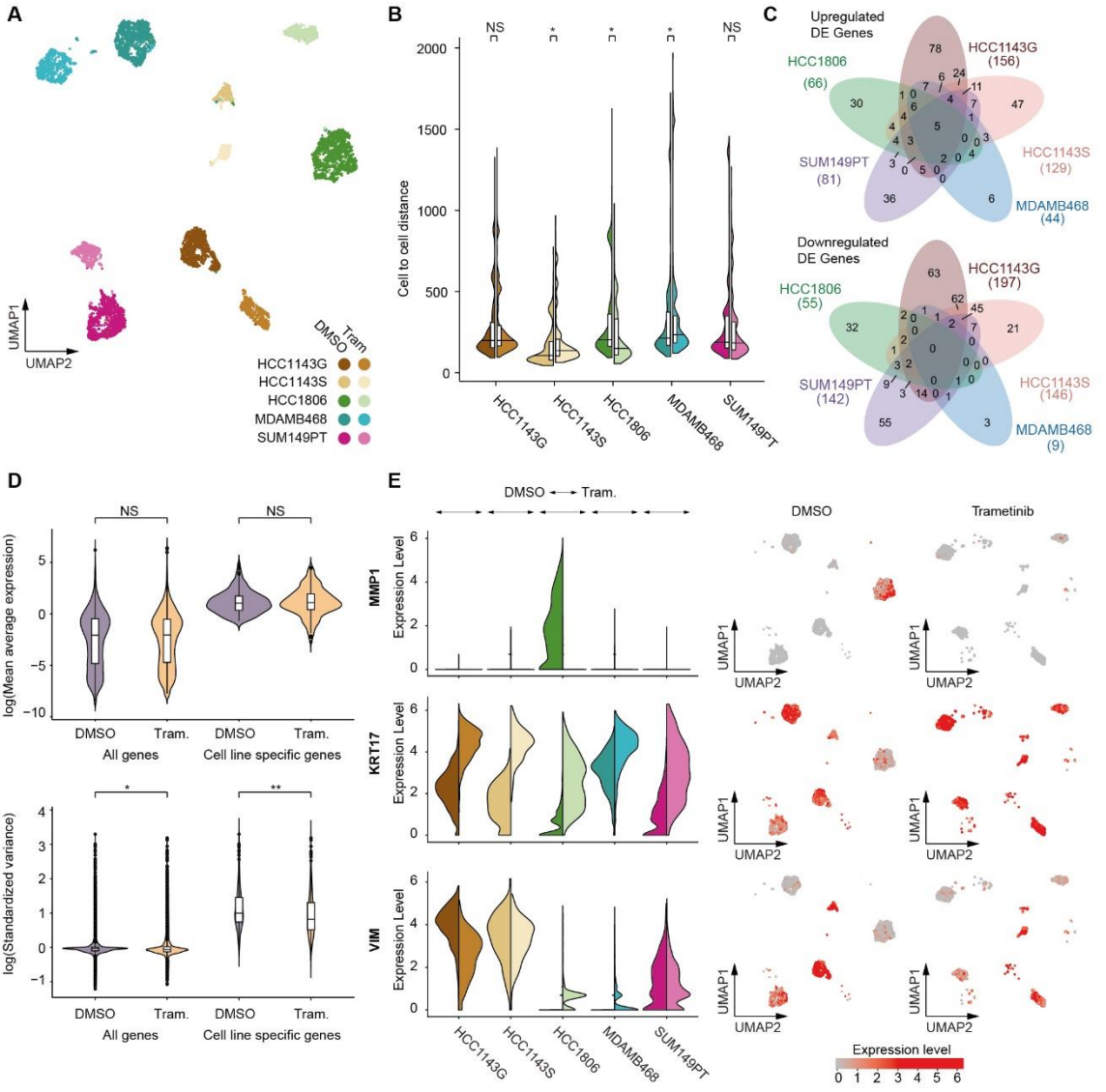

Figure 4.3 Transcriptomic state shift and cross-cell line homogenization of BCCLs upon treatment. (A) UMAP visualization of scRNA-seq for both conditions for the five BCCLs.(B) All-by-all cell-cell Euclidean distances as a measure of intra-line heterogeneity shifts between control (left) and Trametinib treated (right). Asterisk indicates a significant difference in distributions (Mann-Whitney U test, $p<0.05$). (C) Overlap of DE genes identified between conditions within

each line. (D) Distribution of mean expression for all genes and CLGs (top) indicating no significant difference between conditions. Difference in expression variability for all genes and CLGs (bottom) indicating a significant shift between both sets, with a greater shift in CLGs. (E) Example CLGs that shift to the shared state, showing the control distribution (left) and Tramtinib treated distribution (right). UMAP plots to the right show the expression level for the genes with the control cells shown on the left and Trametinib treated cells on the right.

We next explored whether our observation of a shift towards inter-line homogeneity, with an extreme preference for CLS homogenization upon Trametinib treatment was reproduced at the transcriptional level. We identified Cell Line-specific Genes (CLGs) by performing a DE analysis between cell lines under control DMSO treatment (using HCC1143G as the parent line for HCC1143), retaining any gene that either showed increased or decreased expression unique to one line when compared to each other line (n=299, 0.01< BH adj p value and |log2FoldChange| >1.5; Figure S6F, Table S6). To ensure that global expression was consistent across each condition, we confirmed that there was no difference between the mean expression for the set of all detectable genes in the experiment (Mann-Whitney-Wilcoxon test, $W = 2.2\times10^8$, *p*-value = 0.3107) as well as for the set of CLGs ($W = 4.4\times10^4$, *p*-value = 0.6546) (Figure 4.3D, upper graph). We then calculated the variance in expression across each set of genes which revealed a small but significant decrease in expression variance across all genes ($W = 2.4\times10^8$, *p*-value = <0.001, 3.4% decrease in mean variance), but a greater significant decrease in CLGs ($W = 5.3\times10^4$, *p*-value = <0.001, 12.3% decrease in mean variance; Figures 4.3D, lower graph and S6G).

These observations match what we found in the chromatin accessibility space, with a substantial shift toward inter-line homogenization of cell line specific accessible loci also reflected in homogenization of cell line specific gene expression. The effect was more muted in the transcriptional space, which may be due to the increased dynamic range of transcript abundance when compared to the near-digital measurements of chromatin accessibility. We observed the shift towards homogenization of CLGs across all possible directional combinations. This includes genes uniquely expressed in a single cell line that are downregulated upon Trametinib exposure, such as *MMP1*, which is notably involved in cancer cell migration, breast cancer progression, and poor

prognosis (Boström et al., 2011; Q. M. Wang et al., 2019), as well as genes that were uniquely significantly repressed in a line that become activated upon drug exposure, such as *KRT17* (Figure 4.3E). The category of genes that exhibited the greatest decrease in variance were those expressed at high levels in a single cell line but absent or low in other lines, and upon treatment the unique line decreased expression while expression was increased in the other lines, as was the case with *Vimentin* (Figure 4.3E).

4.3.4 Integration of single-cell chromatin accessibility and transcriptome datasets

The cell line diversity and distinct state shift in BCCLs upon Trametinib exposure provides an information-rich structure ideally-suited to cross-modality dataset integration. We first identified co-accessible loci across regulatory regions in our sci-ATAC-seq dataset using *cicero* (STAR Methods) (Pliner et al., 2018b). This produced 573,458 total co-accessibility links with 195,259 meeting a positive co-accessibility score cutoff of 0.15, a threshold we have previously shown to be highly concordant with chromatin conformation data and viable for performing integration with transcriptomic datasets (Sinnamon et al., 2019a). Using these links, we constructed a matrix of gene activity scores that leverages ATAC signal at promoter and linked co-accessible elements for each gene. We next analyzed this matrix in the same way as a typical scRNA-seq dataset using *Seurat* (Stuart et al., 2019), which produced clusters and top differentially expressed genes that were consistent with the clusters and genes identified in the scRNA-seq dataset (Figure S7A-S7C, STAR Methods).

Recent techniques (Stuart et al., 2019; Welch et al., 2019) have advanced our ability to accurately identify mutual information between modalities and enable the co-embedding of the distinct datasets into a shared manifold. Using the framework of anchoring between modalities included in *Seurat*, we harmonized our sci-ATAC-seq and scRNA-seq datasets to produce a single, integrated matrix in gene space. We visualized our integration in two-dimensions via UMAP which produced ten groups, two for each cell line representing DMSO control and Trametinib treatment

conditions similar to sci-ATAC-seq and scRNA-seq UMAP visualizations when analyzed independently (Figure 4.4A).

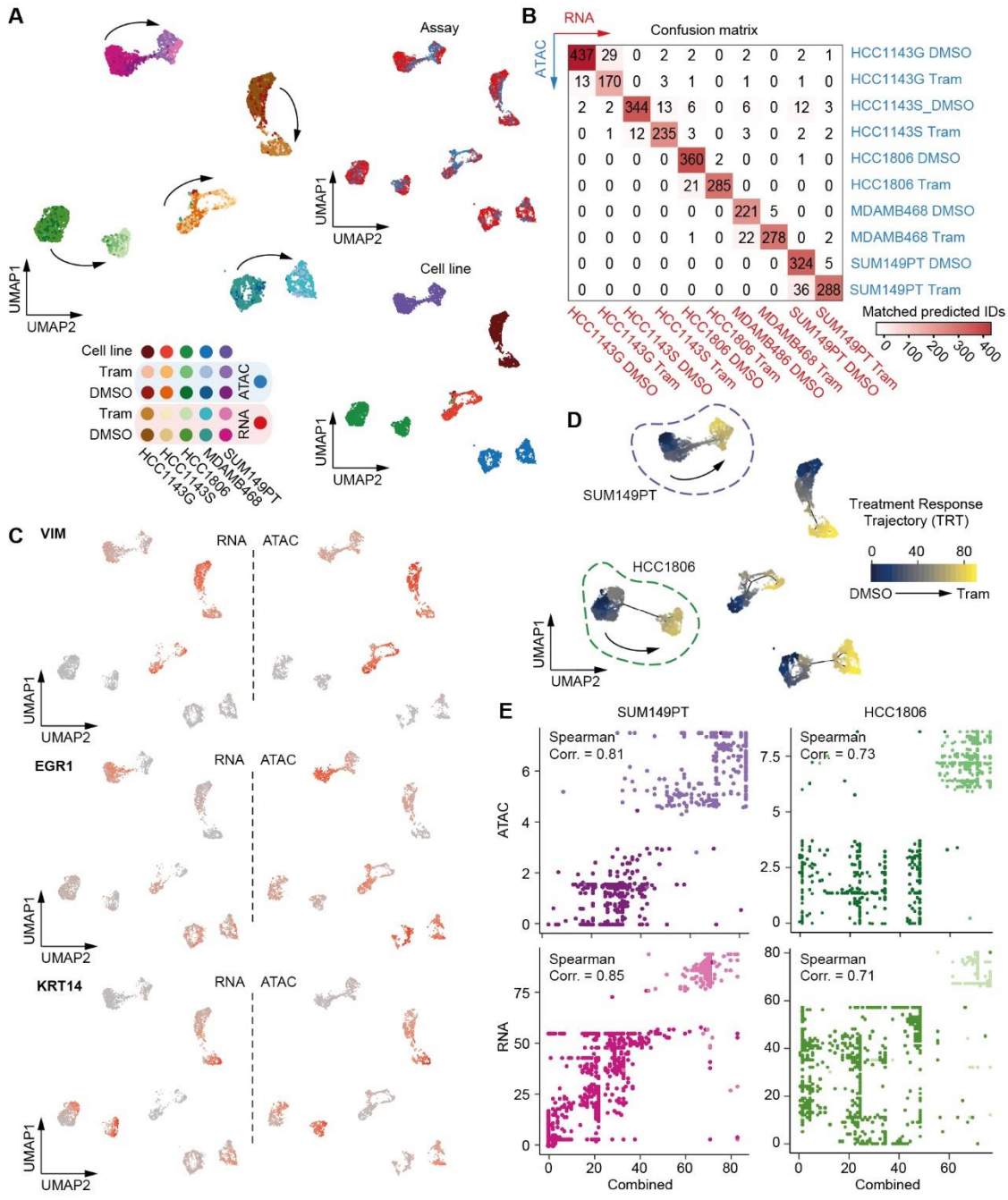

Figure 4.4 Integration of sci-ATAC-seq and scRNA-seq data and establishing a treatment response trajectory. (A) UMAP projection of integrated datasets. Arrows indicate direction from control to Trametinib treated groups of cells. Assay identity is shown in the upper right panel. Cell line identity is shown in the bottom right panel. (B) Confusion matrix of the predicted labels for sci-ATAC-seq cells based on the integrated scRNA-seq data. (C) Example genes with expression levels for the scRNA-seq cells shown on the left and gene activity scores for sci-

ATAC-seq cells shown on the right. (D) Treatment response trajectories (TRTs) for each of the lines. Cells are colored according to their position along the TRT. The two lines shown in (E) are circled in dashed lines. (E) Correlation between the TRT position of cells in the joint manifold in (D) with the position from the ATAC (top) and RNA (bottom) TRTs alone. Cells are colored by treatment condition.

The anchors also enabled the assessment of the cross-modality mutual nearest neighbors (MNNs) which we used for label transfer to assess the accuracy of the harmonized dataset. Across the MNN assignments, 98% had the same cell line label and 93% had both an accurate cell line and treatment label, which is expected to be lower due to the overlap of labels within the transition space from treated to untreated in the manifold (Figure 4.4B). The success of the label transfer was also reflected in the distribution of max prediction scores of anchors, with 98.6% of anchors scoring above the high quality 0.5 cutoff (Figure S7D). The successful integration of our datasets allowed us to observe correlated changes at genes, such as *VIM*, *EGR1*, and *KRT14* (Figure 4C), in epigenetic and transcriptome space.

Robust methods exist to determine an ordering of cells through a pseudotemporal (Cao et al., 2019a; Sinnamon et al., 2019a) or pseudodose (Srivatsan et al., 2020) space. We leveraged our harmonized scRNA-seq and sci-ATAC-seq gene activity score manifold to project Treatment Response Trajectories (TRTs) from the extreme cell state of control DMSO exposed cells to the extreme of the Trametinib treated cells for each individual line (Figure 4.4D; STAR Methods, (Cao et al., 2019a)). To confirm that the integrated dataset TRT is concordant with each individual modality alone, we also performed ordering for each modality separately and correlated ordering between the integrated and independent TRTs which produced Spearman correlations between 0.61 and 0.85 across the lines (Figures 4.4E, S7E-S7G). The correlations did not exhibit any discernible bias to either the chromatin accessibility or transcription modality. Taken together, the integrated TRTs enable the capture and analysis of the interplay between regulatory and transcriptional changes along the continuum of response to Trametinib

4.3.5 Cross-modality integration allows dissection of gene regulatory mechanisms during drug response

The combined dataset gives us the power to track the transcriptional changes and, using co-accessibility scores, identify the dynamics of the associated regulatory elements, that may contribute to the expression changes observed in each gene. This enables a high-resolution dissection of the epigenetic control of transcription during the dynamic cell state transition of each BCCL in response to Trametinib. We examined the chromatin state changes for regulatory regions associated with key genes that had altered expression after response to Trametinib.

Using our integrated single-cell chromatin accessibility and transcriptome TRT, we first assessed instances where transcriptional changes were generally shared among all five lines in response to Trametinib treatment, including *KRT17* which increased in all lines (Figure 4.3E, middle). We observed a varied pattern of distal element usage at loci linked to the *KRT17* promoter region (Figures 4.5A and 4.5B). For example, usage of a regulatory element 25.8 kbp downstream of the promoter (RE2) varied between the two HCC1143 lines with HCC1143G exhibiting accessibility at RE2 with an increase in accessibility through the TRT, whereas HCC1143S showed little to no accessibility at RE2 at any point in the TRT (Figure 4.5B). This element is also linked via co-accessibility with the promoter of *KRT14*, which increased in all lines except for HCC1143S for which expression was undetected in both conditions (Figure S8A), suggesting that in HCC1143S this element and the *KRT14* gene are likely not physically associated with the activating chromatin network that drives the expression increase of *KRT17*. In contrast, the RE1 of *KRT17* (283 kbp upstream of the promoter) was open in both HCC1143 lines, but increased its accessibility upon Trametinib treatment only in the HCC1143S line (Figure 4.5B). RE1 falls within the gene body of *ACLY* which maintained a constant expression level under control and Trametinib treatment in all lines (data not shown). The MDAMB468 line appeared to be the only line that increased *KRT17* promoter accessibility under Trametinib treatment, and this line also increased

accessibility of RE2. HCC1806 did not appear to change accessibility at any of these distal regions, which is consistent with reduced distal element changes that are exhibited by the line relative to the others (Figure S11B), and suggesting that the expression change observed for *KRT17* in HCC1806 is promoter-driven. Other dynamic genes in this region that were not linked to *KRT17* include *KRT15* and *KRT19* which increased in expression in all five lines, *KRT13* which showed a marked increase in expression only in HCC1806, and finally *FKBP10* and *P3H4* which lost expression solely in HCC1806 (Figure S8A).

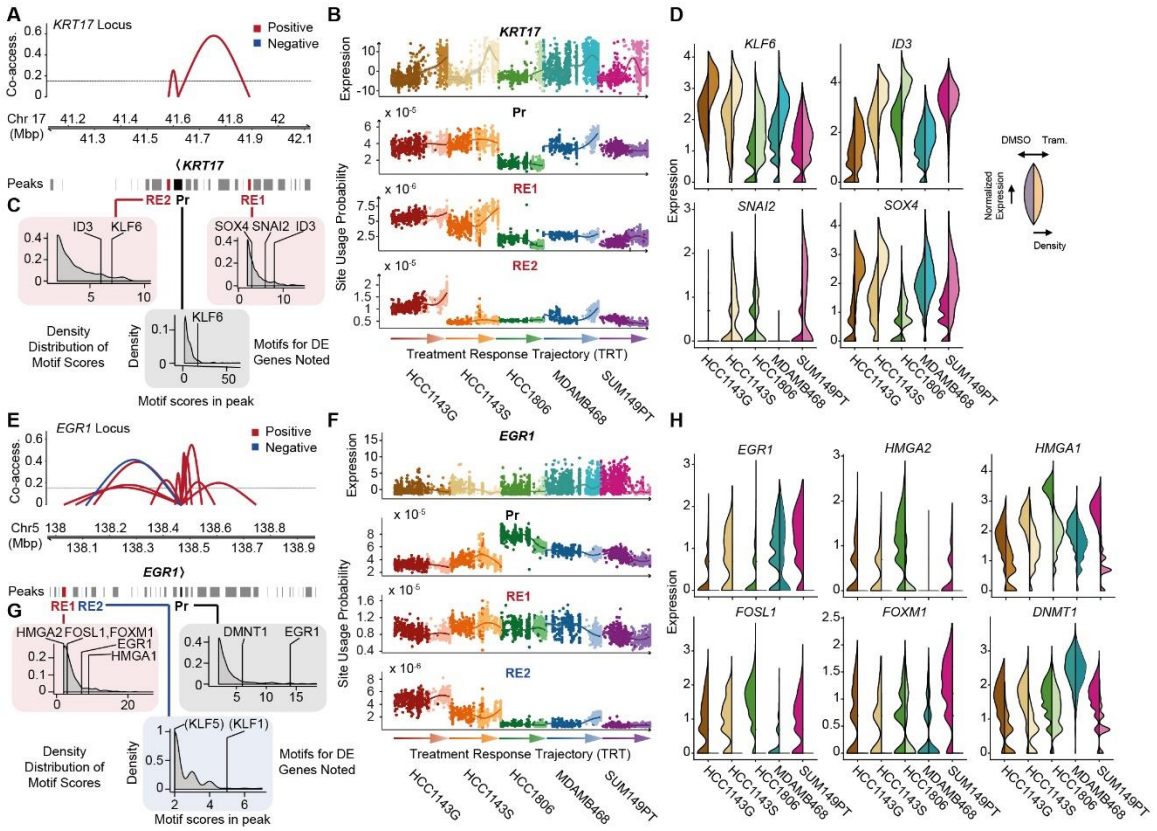

Figure 4.5 Dissection of regulatory mechanisms at dynamic genes. (A) Co-accessible loci at the *KRT17* locus. Links with a co-accessibility score greater than our cutoff of 0.15 between the *KRT17* promoter and distal elements are highlighted. (B) Expression level along the TRT for each of the five lines for *KRT17* (top), with the site usage probability in ATAC data along the TRT for the promoter and linked elements shown below. (C) Distribution of significantly enriched motif scores present at the *KRT17* promoter and linked distal elements. Highlighted motifs have corresponding gene expression changes. (D) Expression levels of transcription factors with significant enrichment of motifs present in regulatory loci associated with *KRT17*. *KRT17* expression levels are shown in Figure 4.3E.(E-H) As in (A-D) for the *EGR1* locus.

To dissect some of the driving mechanisms that underlie the variable regulation of *KRT17* across the cell lines, we examined the motifs present at the promoter and associated RE1 and RE2 distal elements. We first identified transcription factors that had a binding motif present in the queried regions with increased accessibility and that also exhibited increased expression upon Trametinib treatment in the various lines (Figures 4.5C and 4.5D). HCC1143S and SUM149PT increased accessibility at RE1 and both lines showed increased expression of *SOX4*, which has a motif present at the site. Additionally, a motif for SNAI2 is also present at RE1, which also increases in expression in HCC1143S and SUM149PT but is not expressed at all in HCC1143G, which does not exhibit accessibility changes at RE1. The repressive activity of SNAI2 is discordant with the upregulation that is observed in HCC1143S and SUM149PT; however, SNAI2 requires co-factors to impart repression, notably LSD1 coded by *KDM1A* (Figure S8A; Phillips et al., 2014), which does not exhibit altered expression in response to Trametinib. With motifs present in RE1 and RE2, *ID3* and associated DNA binding factors may also play a role (again complicated by its most common role as a transcriptional repressor), since it is increased or already high in lines showing increased accessibility in either RE1 or RE2. At RE2, both HCC1143G and MDAMB468 showed increased accessibility in response to Trametinib exposure, with MDAMB468 also showing increased accessibility at the promoter. KLF6 motifs are present in both of those regions, and *KLF6* expression increases in both of these lines. Interestingly, despite expression increases in *KLF6* and *ID3*, HCC1806 cells do not change accessibility at any of these regions; however, it has the lowest starting accessibility of any of the lines and the least increase in expression of KRT17 with treatment, suggesting that the KRT17 locus is mostly closed and remains that way with trametinib exposure.

We next examined the repression of *EGR1* (Figures 4.5E-4.5H), which all cell lines except MDAMB468 decreased in expression to undetectable levels (Figures 4.5F and 4.5H). In total, 10 REs positively co-accessible with the *EGR1* promoter were identified (including RE1) as well as 6 REs that were negatively co-accessible (including RE2; Figure 4.5E). One of the positive associated

co-accessible links is to the promoter of *HSPA9*, which is associated with proliferation, and showed a slight decrease in expression in all lines consistent with closing of this site. Other genes in this region either do not change expression, or if they do, they are not linked by co-accessibility with the promoter of *EGR1* (Figure S8B). The negative co-association with RE2 (353.5 kbp downstream) was largely driven by an increase in accessibility in MDAMB468 with subtle increases in the HCC1143 lines. In all lines, the *EGR1* promoter exhibited a decrease in accessibility, consistent with the promoter closure representing the top shared chromatin accessibility change upon Trametinib exposure (Figures 4.2D and 4.2E) and consistent repression in expression; however, this drop in accessibility was most stark in HCC1806, in line with the observation that this line has a more dynamic promoter landscape than the other lines (Figure S3G). The remaining distal REs that were associated with *EGR1* all decreased in accessibility.

In analysis similar to that described above for *KRT17*, we looked for down regulated transcription factors that bound to motifs in the REs or promoter regions of *EGR1* (Figure 4.5G). This analysis revealed motifs for EGR1 and DNMT1 at the promoter region, and expression of these factors was decreased in all cell lines except MDAMB468, consistent with our DA analysis. At RE1, HMGA1, HMGA2, FOSL, and FOXM1 motif accessibility was lost in the SUM149PT and HCC1806 lines, with a slight decrease in HCC1143G and their expression was also downregulated. Interestingly, the RE2 site that increases in accessibility in the MDAMB468 cell line had KLF1 and KLF5 motif accessibility, which shows close similarity in its motif to KLF6 which played a role in the regulation of KRT17 and had increased expression in the same cell line (Figures 4.5A-4.5D). This could indicate that KLF6 may instead be driving these accessibility increases at multiple sites genome-wide harboring KLF family motifs in this line, highlighting both the challenges of motif-based analysis as well as the benefit of coupled transcriptional data that can be used to identify the most likely candidates.

We additionally identified several instances where a gene was increasing or decreasing in expression similarly in response to Trametinib in multiple lines, with regulatory elements diverging

in their accessibility change. For example, HCC1143G and SUM149PT lines both increased expression of *S100A*; however, a linked distal regulatory element decreased in accessibility in SUM149PT and increased in the HCC1143G line (Figure S8C). This interaction was also evaluated for other genes linked to the conflicting regulatory element with no alternative gene that was likely driving this change. Our strategy for dissecting out the REs contributing to transcriptional changes also helped in separating linked regulatory elements of genes (*ISG15, HES4)* where promoters were in close proximity and difficult to resolve (Figure S8D). Together, these results highlight the distinct epigenetic mechanisms of each of these basal-like BCCLs in response to MEK inhibition with Trametinib.

4.3.6 A global view of regulatory dynamics during drug response

In addition to dissecting the regulatory control of genes that were identified as top hits by our individual analyses of the sci-ATAC-seq and scRNA-seq datasets; we sought to assess the global regulatory trends along the Trametinib TRT across the cell lines. We visualized these relationships by plotting the Spearman correlation of the accessibility of the regulatory element and gene expression through the TRT for all elements within 500 kbp versus the raw co-accessibility score of the distal Regulatory Element (RE) - Promoter (Pr) association. Notably, a majority of regulatory elements fall below our established score threshold for positive or negative co-accessibility significance ($\geq|0.15|$). This is likely driven primarily by constitutively accessible promoters that may not alter in accessibility when shifts in transcription occur, as well as challenges in promoter-gene assignment when multiple putative promoter elements are clustered together in close proximity.

Among the significant positive or negative co-accessible Pr-RE associations and significant correlations of accessibility and transcription changes ($\geq|0.5|$), we established four quartiles, each representing a distinct regulatory association for upregulated and downregulated sets of genes (Figure 4.6A and 4.6B). The first, Q1, represents REs that are correlated with the

proximal gene and have a positive co-accessibility score to the promoter. For example, *AREG* in SUM149PT cells (Figure 4.6C, top right), which has a distal element (RE2) that is co-accessible with the promoter, both of which decrease in accessibility upon Trametinib exposure and correlate with the decrease in expression of the gene. In contrast, Q2 represents REs that either increase or decrease in accessibility in concordance with the promoter, but are anti-correlated with the expression of the gene (see Figure 4.6C, lower right for examples). This quartile can be explained by the recruitment (downregulated genes) or release (upregulated genes) of repressive factors that affect the openness of chromatin at both the promoter and associated RE. Q3 represents instances where the promoter accessibility and gene expression are correlated; however the level of accessibility of the RE is anti-correlated with the gene expression and exhibits a negative co-accessibility score. This can include instances of a repressive factor modulating chromatin accessibility by binding to or releasing from the RE to decrease or increase, respectively, the promoter accessibility and expression of the gene. However, this quartile would also include loci that are not modulating the proximal gene, but instead another gene that changes expression in the opposite direction, thus driving the negative correlation with expression. Finally, Q4 represents REs that positively correlate with the expression change of the proximal gene and both are anti-correlated with the change in promoter accessibility. This quartile is one of the least populated and typically exhibits only a subtle shift in promoter accessibility change, but could represent an instance of the release of a repressive factor at the promoter facilitated by activating factor recruitment at the associated RE. Taken together, these quartiles represent a global view of distinct regulatory mechanisms linking regulatory element usage and resulting gene expression changes.

We next reasoned that a focused analysis on the elements with the highest likelihood of functional impact, as assessed in our quartile analysis, may provide insights into the global factors that drive the emergence of DTP states upon Trametinib treatment. We utilized the REs that fell into each of the quartiles (Pr-RE co-accessibility cutoff $\geq|0.15|$ and correlated with linked gene

expression ≥|0.5|) and filtered to include only those with a minimum 25% fold change in the average probability of accessibility in either direction and that are statistically significant (Figure 4.6D, t-test, q<0.01; STAR Methods). We found a significant portion of these sites to overlap with cell line specific sites playing a role in inter-line homogenization (hypergeometric test q-value < 0.001) within all lines (Figure 4.6E). This analysis is consistent with the previous observation that CLSs are most likely to change upon treatment (Figure 4.2I and 4.2J), and further associates these REs with functional impact on the transcriptional state of the linked gene.

Finally, we sought to ascertain the putative DNA binding factors driving the high-confidence functional impact on dynamic transcription. Of these sites within the four quartiles, only sets containing REs that become more accessible as the target gene increases in expression or become less accessible as the target gene decreases in expression (*i.e.* quartile Q1) had enough REs to perform motif enrichment, in line with the dearth of putative repressive factors in the quartile analysis we performed. For these positive correlation element lists, we observed little overlap between cell lines in the significant motifs within the categories (Figure 4.6F). SUM149PT and HCC1806 shared the most within the set of REs associated with transcriptional activation, including E2A, SNAI2, GABPA, and CTCF / BORIS. Within the sets of motifs enriched in REs associated with downregulated genes, SUM149PT and HCC1143 showed the greatest overlap and included Max, ELF5, ISRE, and IRF2. Taken together, the majority of distal REs that have a high-confidence functional impact on transcription are enriched for CLSs, can have a contradicting direction of accessibility change between lines with a shared transcriptional shift, and harbor little similarity with respect to motif enrichment, pointing to a complex system of regulatory control that is highly contextually dependent.

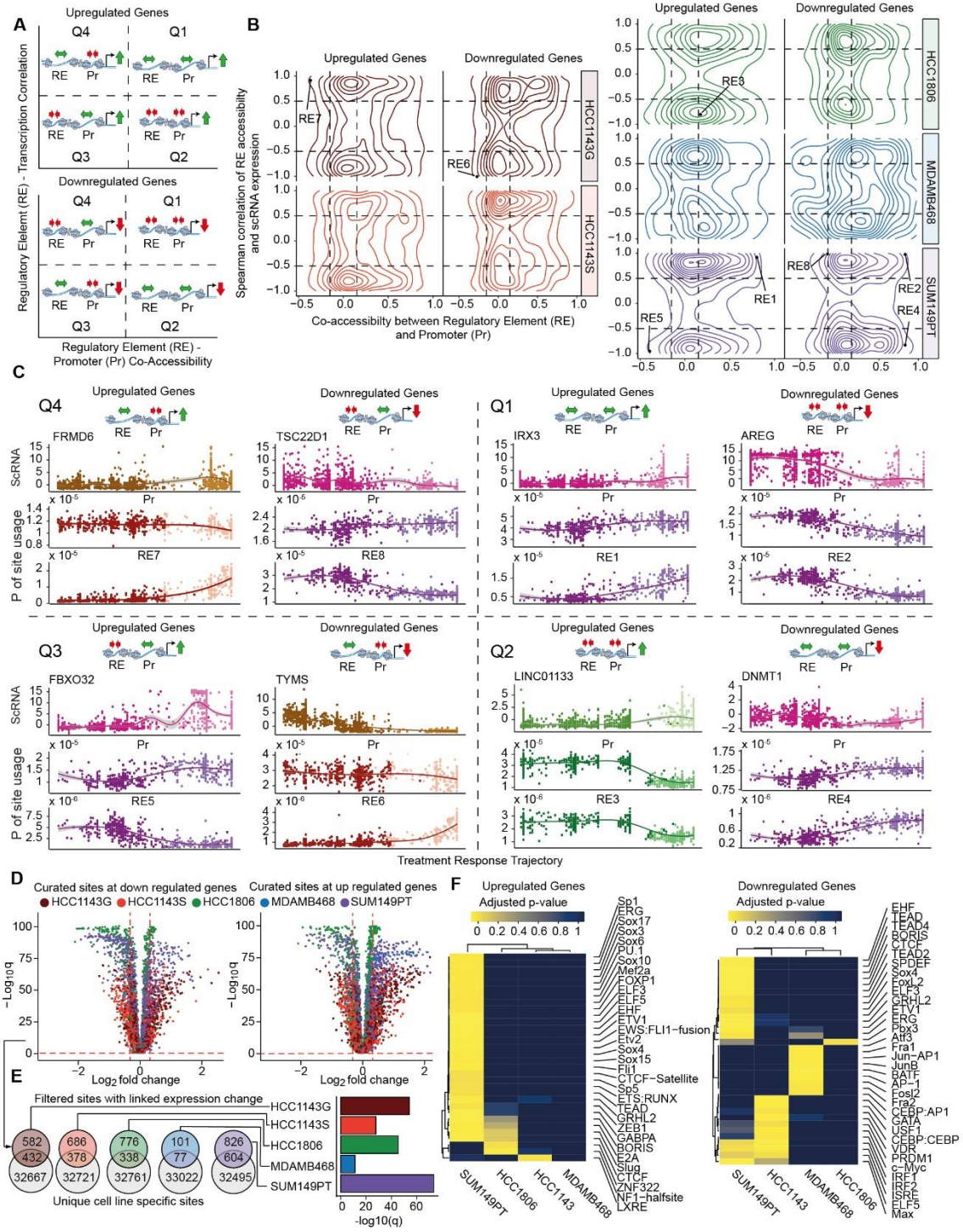

Figure 4.6. Global regulatory networks during the emergence of Trametinib DTP states.
(A) Schematic of possible regulatory element (RE) relationships with a proximal promoter and gene broken down into quartiles based on the RE-Pr interactions. (B) Density plots of RE-Promoter-Gene relationships for each BCCL. Dashed lines indicate cutoffs for assignment to the relationship quartiles portrayed in (A). Specific REs that are highlighted in (C) are indicated. (C) Examples of expression (top), promoter accessibility (Pr, middle), and regulatory element accessibility (RE, bottom) for upregulated and downregulated genes within each of the quartiles

established in (A). (D) Curated high-confidence REs (Pr-RE co-accessibility cutoff ≥|0.15|, Spearman correlation of RE with linked gene expression ≥|0.5|, |log2fold(Tram accessibility/DMSO accessibility)| ≥0.322 at RE and t-test with BH correction q value <0.01) that show significant fold change and are most likely to drive transcriptional changes. (E) High-confidence sites in (D) are significantly enriched for CLSs. Bar plots represent the -log10 q-value for a hypergeometric test. (F) Motif enrichment for high-confidence dynamic, functional REs in quartile Q1 that gain accessibility and are associated with upregulated genes and increased promoter accessibility (left), as well as those that decrease in accessibility and are associated with downregulated genes and decreased promoter accessibility.

4.4 Discussion

Triple negative breast cancers (TNBCs) constitute approximately one sixth of all invasive breast cancer cases (Dent et al., 2007). This subtype is heterogeneous and aggressive, characterized by a high rate of early relapse and residual risk upon treatment (Carey, 2011; Liedtke et al., 2008). In contrast to other subtypes for which targeted therapies against hormone receptors have improved overall survival, well-established targeted therapies for TNBC do not exist. Molecular profiling of TNBC has suggested potential targeted therapies (*e.g.* PARP inhibitors for BRCA deficient tumors), but similar to that seen with chemotherapies, TNBC tumors rapidly develop resistance to these treatments. Understanding mechanisms of therapeutic resistance is essential to developing new treatment regimens for TNBC. Here, we provide a thorough systems biology analysis at the single-cell level of the epigenetic and transcriptional response of five basal-like TNBC cell lines to MEK inhibition with Trametinib. This work provides a resource of information comparing five genetically distinct cell lines at baseline and following treatment, and develops new technologies for interrogating complex connections between chromatin accessibility and gene regulation. Importantly, we uncover a shift toward inter-line homogenization at both the epigenetic and transcriptional level that suggests convergence toward a resistance phenotype despite distinct mechanistic pathways toward that resistance.

Basal-like breast cancers exhibit high activation of the RAF-MEK-ERK pathway, and basal breast cancer cell lines have been shown to be more sensitive to MEK inhibition than luminal lines (Mirzoeva et al., 2009). MEK inhibition with Trametinib has shown antiproliferative effects in

BLBC both *in vitro* and in *in vivo* studies, however, tumor cells often acquire resistance through epigenetic adaptation (Risom et al., 2018; Saini et al., 2013; Zawistowski et al., 2017). A significant contributor to this resistance is the high intratumoral heterogeneity and plasticity associated with TNBC tumors. Previously, we showed that basal-like TNBC cell lines can reconstitute the high degree of differentiation-state phenotypic heterogeneity present in that subtype of tumors (Risom et al., 2018). Trametinib treatment drove increased expression of basal differentiation state markers (*e.g.* KRT 5, 14, 17) in BLBC cell lines and combination treatments with BET inhibitors restricted cell state switching, resulting in increased cell death. This highlights the importance of cell state plasticity in acquiring treatment resistance. Here, we've taken an approach to assess single cells to understand how intra- and inter-tumor heterogeneity affects treatment response. We present a model by which tumor cells repurpose existing developmental pathways of the mammary gland to traverse the Waddington epigenetic regulatory landscape (Waddington, 1957) to acquire drug resistance, and do so along independent paths (Figure 4.7A).

We first evaluated baseline epigenetic states via topic and differential accessibility approaches in the DMSO control cells. The HCC1143 cell lines exhibited the highest baseline intra-line heterogeneity, consistent with our previous work showing combinations of luminal, basal, and mesenchymal states within this line (Risom et al., 2018). We were also able to compare transcription factor motifs within the topics that defined each cell line to begin to understand the distinct regulatory pathways that maintain these cells at baseline. For MDAMB468, we found an enrichment for AP2, ETS, CTF and Forkhead (FOX) motifs families, with FOXA1 having the highest enrichment. FOXA1 is a transcription factor that is highly expressed, with highly accessible motifs, in mature luminal breast cells, and it is associated with repressing the basal molecular phenotype within breast cancers (Bernardo et al., 2013; Heinz et al., 2010a). FOXA1 was also shown to drive symmetric division of TNBC cells resulting in de-enrichment for KRT14 expression (Roy Z. Granit et al., 2018). This analysis may help explain the low baseline levels of KRT14 in this cell line. HCC1806 also showed enrichment for FOX motifs, along with GATA, IRF, and P53

family motifs, with the TP63 motif showing the greatest enrichment. The simultaneous enrichment of luminal (FOXA1, GATA3) and myoepithelial (TP63) motifs likely underlies the heterogeneity in this cell line. SUM149PT peaks were enriched for bZIP (AP-1), KLF, CTF and RUNX family motifs. HCC1143S-specific peaks also showed AP-1 enrichment along with NFkB, TEAD, and Homeobox motifs. And finally, HCC1143G peaks were also enriched for TEAD, and also showed enrichment of RUNX, ETS, and SOX motif families. These results indicate that these BLBC cell lines maintain distinct epigenetic states driven by different transcription factor regulators that support varying differentiation states and degrees of internal heterogeneity.

Our interrogation of the drug tolerant persister states after treatment of these BLBC cell lines, represented as the bottom of the Waddington landscape (Figure 4.7A), showed a dearth of shared changes at both the epigenetic and transcriptional level. There were no shared topics between cell lines (with the exception of the two HCC1143 lines). Analysis of the genes nearest to DA sites in the Trametinib vs. DMSO samples by sci-ATAC-seq revealed a small number of sites with consistent decreased accessibility. These genes, overall, had roles in proliferation (*e.g. EGR1, NTSR1, FOS*), cell metabolism (*e.g. IDH3B, GOT1*), or cell motility (*IER2*) (Ouyang, 2017, Al-Khallaf, 2017; Meléndez-Rodríguez et al., 2019, Shaulian and Karin, 2001; Wei, 2017). *EGR1* was the most decreased, and low *EGR1* expression has previously been correlated with unfavorable outcomes in breast cancer tumors treated with Tamoxifen (Shajahan-Haq et al., 2017). Additionally, reduced levels of *EGR1* promoted resistance to Paclitaxel (PTX) treatment in TNBC cells via the promotion of slow cell cycling (Lasham et al., 2016). Taken together, the epigenetic silencing of *EGR1* may contribute to loss of proliferation in these Trametinib treated BLBC cell lines and contribute to the emergence of a DTP state. Interrogation of DE genes in DTPs vs control cells in RNA-seq also showed few shared transcriptional changes. The limited set included a subset of keratin markers that we previously observed as increased in Trametinib treated cells, confirming similar plasticity in differentiation state (Risom et al., 2018). In addition, *CD24* expression increased in all lines, and *CD44* decreased in four of the five, indicating a potential shift away from

a more proliferative stem-like state commonly observed in basal like breast cancer (W. Li et al., 2017).

When comparing the control and Trametinib treatment conditions, we observed a slight increase in intra-line heterogeneity at the chromatin level in four of the five lines, with HCC1806, the most resistant of the lines, as the exception (Figure 4.7B). A significant difference in intra-line heterogeneity was also observed in three lines at the transcriptional level, with HCC1143S and MDAMB468 showing an increase and HCC1806 showing a decrease. The reduction in internal heterogeneity of HCC1806 may be due in part to the promoter-driven nature of the line, which tends to be less variable than their distal element counterparts and are often constitutively open. Also consistent with this observation is that the lines with increasing heterogeneity at both the chromatin and transcriptional level have a higher density of cells in the transition space between control and treated states (Figure 4.4A, 4.4D), likely contributing to the heterogeneity and indicating that the DTP state for these lines may not be as stable, even after 72 hours of drug exposure.

Despite the lack of individual sites or genes changing similarly across the cell lines in response to Trametinib treatment, we observed a striking inter-line homogenization in which the sites and genes that were unique to specific cell lines at baseline (CLSs and CLGs, respectively) were the most changed in response to treatment, and they changed in a direction that made the inter-line DTPs more similar. The decrease in distance between cell lines with respect to differentially accessible loci after treatment also supports this novel finding (Figures 4.2I-4.2G, 3D and 4.3E), as does the observation of decreased inter-line transcriptional variance that was greater at cell line specific genes. This phenomenon was observed in both directions across both modalities: *i.e.* uniquely active CLSs and CLGs shifting to an inactive state and uniquely inactive CLSs and CLGs shifting to an active state. Perhaps more surprising was the observation of instances where a uniquely active CLS or CLG stays active and the other four lines shift to a higher level of activity to match the individual line as well as the uniquely inactive reciprocal. Notably, in both the

chromatin accessibility and gene space, the features that were cell line specific did not have any similarity other than their unique presence within their respective cell line and their propensity to change in some way in response to Trametinib. These changes were in the form of either a higher probability of altering their accessibility (Figure 4.7C) or expression level and variance (Figure 4.7D) to become more similar to the other lines. Furthermore, regulatory elements identified as CLSs were also more likely to be associated with transcriptional changes (Figure 4.7E).

The consistency in observations between the chromatin and transcriptional layers prompted us to integrate our data across the two modalities in order to understand the path each cell line takes along the Waddington landscape during its adaptation to Trametinib. Recent published methods have shown successful integration of scATAC-seq and scRNA-seq data via projecting data into a shared latent space (Stuart et al., 2018; Welch et al., 2019). Our results showed successful integration between modalities with a high matching of ATAC-seq cell annotations and inferred labels after label transfer (Figure 4.4). In addition, ordering cells along the shared latent space proved to retain individual modality ordering, suggesting this as a successful alternative to other shared trajectory inference methods (Welch et al., 2017); however, we note that our dataset contains five distinct lines with two distinct states each, and other methods may be more appropriate when higher proportions of cells fall between states. This shared TRT between modalities enabled a detailed dissection of gene regulatory circuits, utilizing linked distal regulatory elements to promoters along with the transcriptional state of the corresponding gene (Figure 4.5). For genes that showed a similar expression change, we were able to break down the exact distal elements that altered their accessibility to drive the change. This revealed varied and sometimes contradicting changes between lines, including between the two HCC1143 lines. These dissections emphasize the importance of context, both locally and globally, when assessing or evaluating regulatory element function, where elements may appear to be activating in one context but either inactive or repressive in another.

To take a global approach to understanding linked epigenetic and transcriptional changes, we identified four regulatory quadrants that allow us to classify interactions between promoters and regulatory elements in a novel way (Figure 4.6A). Our results show that a large portion of REs behave classically, with an accessibility pattern that matches that of the promoter and resulting gene expression changes. Q2 and Q4 represent more unusual cases where promoter usage does not match changes in expression. Interestingly, MDAMB468 and SUM149PT cells show an excess of Q2 regulatory interactions, which will require further study. It is important to note that a significant portion of interrogated REs showed high correlation between transcriptional and accessibility changes, but no linkage between the promoter and REs. These could likely be cases where no dynamic changes occur in promoter accessibility and could expand our knowledge of regulatory hubs with constitutive promoters that have remained elusive to methods like Cicero (Pliner et al., 2018a). This approach also enables the systematic identification of regulatory elements with seemingly contradicting function in the context of different cell lines (Figure S8D). Finally, we leverage this analysis to establish a high-confidence set of REs that drive transcriptional changes. These sites were enriched for CLSs, which is consistent with the observation that CLSs as well as CLGs are more likely to change their activity level upon drug exposure; however, there were no clear consistencies between lines with respect to specific transcription factors that could be driving these changes.

By leveraging single-cell epigenetic and transcriptional changes we gained a mechanistic understanding of cellular plasticity in BLBC cells in response to Trametinib treatment. We assess these mechanisms using a novel framework for assigning transcriptional consequence to regulatory elements in a dynamic context. The resulting networks varied across lines, even where transcriptional changes may be shared, which underscores the importance of local and global context when assessing RE function, and bears particular relevance to massively parallel reporter assays, where often only a single context is evaluated. We also identified a novel paradigm of drug treatment induced homogenization within basal-like TNBC cell lines, where drug exposure drives

the preferential shift in RE accessibility and gene expression levels from a unique cell line specific state in a single line to a state shared by all lines. This may be promising for future treatment strategies, as it suggests that intratumoral heterogeneity can be successfully steered toward a more common DTP state, and that novel combination therapies targeting the common end state could contribute to a more complete remission.

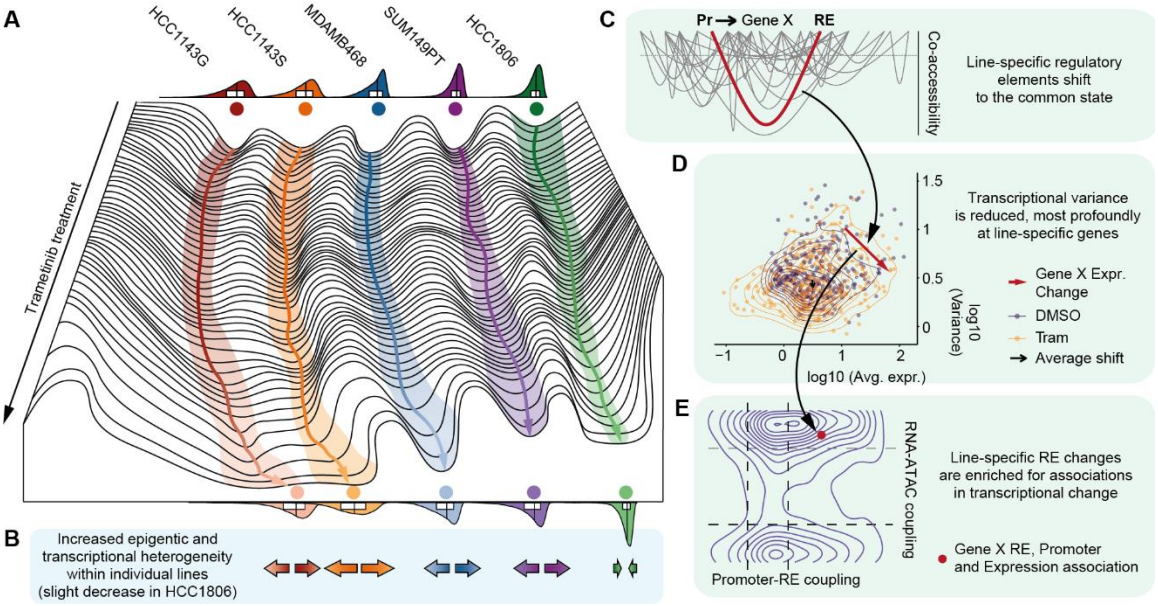

Figure 4.7 The dynamic epigenetic landscape of MEK inhibition in TNBC cell lines. (A) Waddington depiction of the treatment response to Trametinib exposure in the five BCCLs. Each line takes an independent path down through the regulatory landscape; however, the differences between the lines, represented as the 'hills' separating the 'troughs' are reduced. (B) Four of the five lines exhibit an increase in intra-line heterogeneity, with one line, HCC1806 that is the most resistant, decreasing in heterogeneity. (C) Regulatory networks were dissected with cell-line specific elements most likely to shift to a common state in DTPs. (D) The epigenetic changes extend to the transcriptional space. There is a global decrease in variance in transcription between lines that is more pronounced at line-specific genes. (E) Regulatory elements that are the most tightly linked to transcriptional changes are also most likely to be cell line specific elements.

4.5 Methods

4.5.1 BCCL cell line culture

Basal-like breast cancer cell lines HCC1143, HCC1806, SUM149PT and MDAMB468 were purchased from ATCC. The HCC1143 cell line obtained from ATCC is called HCC1143G throughout this paper to distinguish it from the HCC1143S line, a subclonal line that appears to have spontaneously drifted from the commercially available line in terms of mutational status and phenotype over years of culture. Cell lines were STR profiled to confirm identity, and were regularly screened to ensure they were free of mycoplasma contamination. HCC1143G, HCC1143S, and HCC1806 cell lines were cultured in RPMI supplemented with 10% FBS and 10 µg/mL Penicillin and Streptomycin (P/S); SUM149PT were cultured in Ham's F12 supplemented with 5% FBS, 5µg/mL Insulin, 10mM HEPES, 10 µg/mL P/S, and 1µg/mL hydrocortisone; MDAMB468 were cultured in DMEM with 10% FBS and 10 µg/mL P/S. Cells were maintained at 37°C and 5% CO2. HCC1143-BL that were used for WES were purchased from ATCC and cultured in RPMI supplemented with 10% FBS.

For treatment, cells were plated overnight in full growth media, and then treated the next day with either 1µM Trametinib or an equivalent volume of vehicle (DMSO). After three days, cells were collected for sciATAC-seq or scRNA-seq.

For the GR curves, stable pools of each cell line expressing nuclear mKate2 were generated using the NucLight Red lentiviral reagent with puromycin selection (Essen Biosciences). These cells were then plated in 96 well plates overnight in full growth media and treated the next day with Trametinib in an 8 pt. dose curve consisting of 1:4 serial dilutions with a high dose of 10µM. Red channel images were taken on the IncuCyte ZOOM (Essen Biosciences) at 0h and 72h of treatment, and images were segmented with the ZOOM software to calculate nuclei counts. GR values were calculated from these counts as previously described (Hafner et al., 2016).

4.5.2 Generating sci-ATAC-seq Libraries

Tn5 transposase was prepared and loaded with barcoded adaptor oligos using published protocols (Picelli et al., 2014; Sinnamon et al., 2019a). Nuclei were isolated from cultured cells by resuspending in 1 mL of ice-cold Nuclei Isolation Buffer (NIB) and incubated on ice for 15 min. DAPI (5 mg/ml in diH2O) was added to the nuclei in NIB at 5 µg/ml final concentration. We then followed our previously described protocol (Sinnamon et al., 2019a) for sci-ATAC-seq on these samples, with some alterations detailed below.

From the 1ml of each sample, 4,000 DAPI-stained nuclei were Fluorescence Assisted Nuclei sorted (FAN, Sony SH800) into 24 8-well strip PCR tubes (192 total wells, BioRad) containing 5µL of 2× TD buffer (Illumina) and 5 µL of NIB per well. The identity of each of 10 the conditions was conserved in the following tagmentation step where 1 µL of 8 µM dual barcoded Tn5 Transposase was added to each of the wells followed by incubation for 15 minutes at 55°C, which was ended with the plates being placed on ice to stop the reaction. After tagmentation all wells were pooled and 40 tagmented nuclei were FAN sorted again into each well of three 96 well PCR plates containing [0.25 µL 20 mg/mL BSA, 0.5 µL 1% SDS, 7.75 µL nuclease-free water, 2.5 µL indexed forward PCR primer, and 2.5 µL indexed reverse PCR primer]. Transposases were denatured by a 15-minute incubation at 55°C following sorting. Then 12 µL of PCR mix (7.5 µL Nextera PCR Mix (NPM), 4 µL nuclease-free water, 0.5 µL 100 × SYBR Green) was added to each well and then PCR amplified on a Bio-Rad CFX thermocycler via the following conditions: {5 minutes at 72°C, 30 seconds at 98°C, and cycles of [10 seconds at 98°C, 30 seconds at 63°C, 1 minute at 72°C; plate read, 10 seconds at 72°C]}. We pulled reactions mid-exponential, usually between 18-20 cycles. Post-PCR-amplification we used a QIAquick PCR purification column to clean 10 µl of each reaction for clean-up. The quality and concertation of the cleaned up libraries was determined using an Agilent Bioanalyzer, which we then diluted to and sequenced on a NextSeq 500 (research use only) using custom primers and chemistry (Vitak et al., 2017b), which produces 10 bp PCR indexes and 8 bp Tn5 indexes at each end and paired-end 50 bp reads of genomic DNA.

4.5.3 Generating sc-RNA-seq Libraries

All five cell lines were treated for 72 hr with 1μM Trametinib or DMSO, then trypsinized, and the 10 samples were each labeled independently with 0.5 μg of a TotalSeq-A anti-human Hashtag antibody (Biolegend). Cells were then pooled, with each pool containing ~10,000 cells. Libraries containing the pooled hashtagged cells were processed using the Chromium Single Cell 3' Reagent kit (v3 chemistry) (10X Genomics). The resulting libraries were profiled using the Tapestation (Agilent), followed by quantification with real time PCR (KAPA Biosystems) using a StepOne real time PCR workstation (Themo/ABI). Each library was loaded onto one lane of a HiSeq2500 (Illumina) for sequencing.

4.5.4 DNA sequencing and SNV calling

Total DNA was collected from HCC1143G, HCC1143S, and HCC1143-BL cells using the Qiagen DNAeasy Blood and Tissue kit (Qiagen). Genomic DNA (2 μg) was sonicated using Covaris E220 Focused Ultrasonicator (Covaris, Inc.) to an average size of 150bp. Whole-exome DNA sequencing libraries were prepared with 500ng of the fragmented gDNA using KAPA Hyper-Prep Kit (KAPA Biosystems) with Agilent SureSelect XT Target Enrichment System and Human All Exon V5 capture baits (Agilent Technologies), following manufacturer's protocols. Next-generation sequencing was carried out using the Illumina HiSeq 2500 platform by the OHSU Massively Parallel Sequencing Shared Resource (MPSSR). Raw paired-end sequencing reads (100bp) in FastQ format output by the HiSeq 2500 were aligned and processed using BWA MEM (0.7.12) software to the full hg19 genomic assembly (GATK, Broad Institute). Picard tools (v. 1.119), SAMtools, and GATK (v. 3.3-0) were used to sort, index, remove PCR duplicates, and locally realign bam files, as well as to generate target coverage and duplication metrics (http://broadinstitute.github.io/picard). After data processing, a mean of 94X on-target coverage was obtained for the sequencing libraries with 67% of on- target reads exceeding 50X depth.

Aligned and processed bam files were compared for calling somatic variants between samples using MuTect (v. 1.1.4, GATK, Broad Institute). Somatic variants were called between HCC1143G or HCC1143S and the HCC1143-BL cell line, which was used as a "matched normal." All resulting variants were filtered to those labeled "KEEP" by MuTect and having at least 30X coverage and 5% variant allele frequency (VAF) in the "tumor" (treatment) and at least 15X coverage (with no presence of the alternate base) in the "normal" (HCC1143-BL). Variants listed in the dbSNP database (build 137, https://www.ncbi.nlm.nih.gov) were also omitted.

Quantification and Statistical analysis

This section describes the analysis preformed on sci-ATAC-seq and sc-RNA-seq datasets and finally the integration and downstream analysis of the two modalities of data. Primary analysis on sci-ATAC-srq was performed using our previously published sci-suite (Sinnamon et al., 2019c). This is a set of tools fo ther analysis of single-cell combinatorial indexed data, including wrappers for commonly used open source software such as including BWA (Li and Durbin 2009), MACS2 (Zhang et al. 2008), BEDTools (Quinlan and Hall 2010), SAMtools, as well as R (R Core Team 2019) libraries: *ggplot2* (Wickham 2016), chromVAR (Schep et al. 2017), chromVARmotifs, Cicero (Pliner et al. 2018), RtSNE, UMAP (Becht et al., 2018b). Usage of scitools for these functions should cite relevant original source. Scitools is available at https://github.com/adeylab/scitools (continuously under development).

4.5.5 Raw processing of data for sci-ATAC-seq

We first converted our BCL to FASTQ files using bcl2fastq v. 2.19. The scitools functions fastq-dump and fastq-split were used to demultiplex reads based on the barcode identifying individual cells. These barcodes are made up of four components: two inserts of the Tn5 tagmentation events on the P5 and P7 ends of molecules (first round of indexing), and the following two unique identifying PCR indexes on both sides of the molecules (second round of indexing). We filtered molecules so each of these four components had to be within two Hamming distances

away from expected barcode sequences. Barcode matched reads were aligned to the hg38 genome using scitools fastq, which implements BWA-MEM v. 0.7.15. We filtered aligned reads by removing PCR duplicates, mitochondrial reads, and reads with a quality score less than 10 via the bam-rmdup function. We then applied our previously described mixed model approach (Vitak et al., 2017b) to the distribution unique reads per cell and a newly implemented knee-plot calling to identify reads from intact cells as opposed to debris (plot-complexity, Figure S1D). Based on the fraction of unique reads per cell and the total reads per cell we selected cells below 60% and above 0% for knee-plot calling, where we ordered cells according to their unique aligned reads and used the R package inflection to call the knee (Figure S1D). Based on these results we filtered cells with less than 5000 unique reads (bam-filter wrapper function) resulting in n cells and then used the pseudo-bulk aggregate of all passing cells to call peaks via MACS2 v. 2.1.1 (158041 peaks). These peaks were then extended to 500 bps, merged and then chromosome border corrected (atac-callpeak).

4.5.6 Topic analysis

We generated a counts matrix containing cells as columns and peaks as rows via scitools counts and filtered to exclude rows with fewer than 10 cells with reads (-R 10) and columns with fewer than 1000 (-C 1000) reads. We then binarized this counts matrix and applied Cistopic v. 0.2.0, a probabilistic topic modelling method based on Latent Dirichlet Allocation, which groups sets of accessible sites and cells simultaneously by common themes (Topics). The resulting two matrixes (topics-cell distribution and the topics-sites distribution matrixes) can then be then used for downstream projections of cell states via uniform manifold approximation and projection (UMAP, on topics-cell matrix) and topic exploration of associated sites (on topics-sites matrix). We first analyzed the DMSO treated cell lines separately to characterize initial cell line heterogeneity. We determined the optimal number of topics (30) via running a likelihood stabilization analysis (Figure S1F) where we ran multiple models with differing number of topics

and chose the model with the highest log-likelihood at the last iteration (iteration number 250). We then analyzed the combined DMSO and Trametinib treated cells. Using the same likelihood stabilization method, we first selected 50 topics (Figure S2A) and then again performed a finer scale analysis near 50 topics resulting in 54 as the optimal number of topics. In both analyses we chose the top associated sites for each topic by binarizing topic sites topics-sites distribution matrix with the GammaFit option of Cistopic (Figure S1G, thrP =0.975) and performed HOMER (Heinz et al., 2010b), http ://homer.ucsd.edu/homer/motif/) known and de novo motif enrichment (Tables S1 and S3), to identify enriched regulators of chromatin accessibility relative to all background peaks.

4.5.7 Differential accessibility

To identify differentially accessible sites between provided annotation groups, we first used scitools aggregate-cells to create clusters of k=40 cells based on their low dimensional UMAP coordinates. Assuming these cells have similar accessibility profiles we aggregated accessibility of cells within these pseudo-bulk groups to use as replicates in the down-stream analysis with the DESeq2 (Love et al., 2014). Using this R package, we corrected for technical biases such as assay efficiency and performed differential accessibility tests (using nBinomWaldTest) between all 45 combinations of cell lines and treatment groups (Figures 4.1E, right panel of S3A and S3B-S3C, Table S2), and between the joined Trametinib and DMSO treated cell lines (Figures 4.2D and S3D). Finally, in all DA analyses we corrected for multiple testing at q=0.01 with the qvalue R package. In downstream analyses we used differentially accessible sites to bi-cluster cell lines (Figures 1E, right panel of S3A and S3B-S3C) and to order topics based on importance via enrichment between DA and topic associated sites (Figures S1J and S3F).

4.5.8 Identifying unique cell line specific sites

To identify unique, cell line specific sites, we averaged the contributions of topics across cells in each of the cell lines based on the probability distribution topics-cells matrix (Figures S1H and S1I) and chose the binarized sites of the highest contributing topics uniquely present in individual cell lines. We argue that LDA inherently results in an overlap of associated sites of related topics, therefore we can assume that strongly associated cell line specific sites are represented in our chosen cell line specific topics, even when multiple topics are cell line specific. In addition, we confirmed the cell line specificity of these sites via differential accessibility analysis (Figure 4.1F and 4.1G, Star Methods). We first intersected sites that were differentially accessible between all combination of DMSO treated cell lines with either HCC1143S or HCC1143G groups held out from the comparisons, then took the union of these two cases. This removed the potential skewing of our results due to the relative closeness of the two subclones in the epigenetic landscape. We plotted the signal within these sites using scitools make-signal and scitools plot-signal (Figures 1D and 1G), which confirmed the cell line specificity of our sites. Using *AUCell*, (Aibar et al., 2017) we were able to confirm the enrichment of DA sites in our chosen topic peaks (Figure S1J), further confirming their cell line specific nature.

4.5.9 Measuring epigenetic heterogeneity of cell populations

Measuring epigenetic heterogeneity can be challenging due to the inherent high dropout rates of single cell chromatin accessibility assays. We applied two approaches to characterize cell to cell and site to site variation accurately within annotations. Dropout rates can be approximated via the use of the predictive distribution matrix, a cells by sites matrix, which is the result of the multiplication of the topics-cells and the topics-sites probability distribution matrixes. With this approach dropouts within individual cells are corrected for via the use the same site with similar topic associations. This matrix is primarily used for the ranking of cell line specific sites in standard Cistopic workflow, but can be applied for approximating cell-cell spearman correlation due the inherent dropout correction (Figure 2F). We then performed a Mann-Whitney U test (r wilcox.test)

and t-test between the treated and non-treated groups within cell lines (Table S4). We argue that the large sample number allows for a t-test to be performed even when the compared distributions do not follow normality due to the central limit theorem. In addition, we approximated site to site variation via Shannon entropy on the binarized cell-site accessibility matrix. We argue that due to the random distribution of dropout across sites, we can correct for read depth by looking at the frequency of cells being accessible at a site within annotations via Shannon entropy.

4.5.10 Cell line specific site uniformization

We observed a decrease in the number of differentially accessible sites between Trametinib treated groups and minimal number of shared changes upon treatment between cell lines (Figure 2G). We tested a potential elevated homogenization effect between lines at cell line specific sites vs random background sites. We calculated the relative shift in Shannon entropy of Trametinib treated cells relative to DMSO treated cells summed over all unique cell line specific sites ($\Delta S = \sum_{i=unique\ sites}(S_i^{DMSO} - S_i^{Tram})$). Similarly, we calculated the sum of Kullback-Leibler (KL) divergence in the Trametinib treated cells compared to control cells treated with DMSO vehicle. The Shannon entropy shift informs us of the direction and the KL divergence can tell us about the magnitude of the effect. To have an accurate comparison as a background we performed bootstrapping (10,000 iterations) where we random sampled the same number of regulatory elements as unique cell line specific sites and calculated the sum of their KL divergence and Shannon entropy. We performed these analyses on cell line specific sites identified via topic unique sites (Figures 2H-2J) and via differential accessibility (Figures S5A and S5B). For the case where unique cell line value was within the background distribution (Figure 2I) we fit a normal distribution and tested for the probability of having more extreme (lower.tail) entropy difference than the unique cell line value via the r function *pnorm*.

4.5.11 Gene set enrichment based on chromatin accessibility

We employed two strategies for gene set enrichment analysis. First, we identified Trametinib only and DMSO only enriched topics via the use of *AUCell*, (Aibar et al., 2017), where we looked for topics which had an relative enrichment in sites that were differentially accessible between Trametinib and DMSO within individual cell lines. We then performed rGREAT (McLean et al., 2010) on the top associated of the highlighted topics (Topics 6,7,11,32,42, 49,51 for Trametinib and 1,2,25,34,41,53 for DMSO) using UCSC liftover of associated sites to hg19 assembly from hg38 (Figure S4A). Second, we used a method similar to *chromVAR* (Schep et al., 2017) to calculate standardized deviation scores over sets of sites +/-10 kb from promoters of provided curated gene sets (Risom et al., 2018) relative to all background sites (atac-deviation command). We then calculated effect size each gene set by taking the difference of the mean standardized deviation in the Trametinib treated group and the DMSO treated control for each of the cell lines. We calculated significance by doing a t-test between the two groups for each of the cell line and gene set. We Bonferroni corrected for the multiple gene set comparisons within cell lines (Figure S4B and S4C).

4.5.14 Differential expression and identifying unique cell line specific sites

We ran differential expression (DE) analyses (FindMarkers command within Seurat 3) between Trametinib and DMSO treated groups within cell lines and across all cell lines with the minimum percentage of features in either compared groups set to 0.25, p adjusted < 0.05 and |log2foldchange| > 1 (Figure S6E, Table S5). We intersected individual DE genes of the cell lines using InteractiVenn (Heberle et al., 2015). We identified cell line specific genes for each of the cell lines by comparing DMSO treated groups of one cell line to all others with HCC1143S held out, due to the relative low cell numbers of this group compared to others and its relative transcriptomic closeness to the HCC1143G compared to other cell lines. For identifying CLGs for downstream analyses we used a more stringent q<0.01 and |log2foldchange|>1.5 cutoff (468 genes, Figure S6F, Table S6).

4.5.12 Identifying cis-regulatory networks and approximating gene activity

We used the R package Cicero (Pliner et al., 2018a) to link regulatory regions based on their shared accessibility across cells. We identified 573,458 total co-accessibility links with 195,259 meeting a positive co-accessibility score cutoff of 0.15. Of these 64,606 co-accessible sites, 44,756 were linked to promoter regions of genes. Using the read depth normalized variation in co-accessibility between regulatory elements and promoters across cell lines and treatment we approximated gene activity (Figure S7A and S7B). Finally, we normalized values by the total number of genes included.

4.5.13 Raw processing of data for scRNA-seq

Samples were run on two lanes of the 10X Chromium v2 system. Initial quality assessment, HTO tag counting, alignment to the hg38 human genome and transcript assignment was done with CellRanger. Median UMI counts per cell was 25,974 and 21,202, and genes detected per cell was 5,128 and 4,815 for pools 1 and 2 respectively. Downstream analysis on the output unfiltered expression matrix and the HTO count matrices was primarily done via Seurat (v. 3). We first used HTODemux provided with Seurat (v. 3) with default settings to identify cells with ambiguous sample origins. Following analyses, however, revealed 11 clusters of cells instead of the expected 10. Differential expression of these clusters revealed the 11^{th} cluster (classified as SUM149PT Trametinib treated cells) to have a gene expression profile similar to HCC1143G DMSO treated cells and SUM149PT Trametinib treated cells. This led us to develop a Shannon entropy-based method, where we used HTO tag frequency within cells to mark heterogeneity of the tags. The distribution of entropy across cells proved to be bimodal, which we fit with a mixed model similar to our sci-ATAC-seq read cutoff strategy. This analysis revealed the 11^{th} cluster to be highly heterogeneous based on its HTO-tags, which we then proceeded to filter out. Depending on the type of analysis, we applied one of the two methods of normalization available in Seurat 3 on the filtered counts matrixes. In analyses such as internal heterogeneity and cell line specific gene

expression uniformization, where removing technical heterogeneity (such as read depth) while preserving biological heterogeneity is important, we employed scTransform for normalization and scaling. Interestingly, we found that the standard Seurat 3 workflow of log-normalization, z-score transformation with the NormalizeData and ScaleData functions were better suited for integration of the sciATAC-seq and scRNA-seq datasets. This could be due to the harsher nature of the normalization on the gene activity matrix. In both methods of normalization, we corrected for mitochondial mapping percentages. We ran PCA analyses with 50 total PCAs for dimensionality reduction on the selected features and projected cells into two dimensions via UMAP (Becht et al., 2018b).

4.5.15 Measuring transcriptomic heterogeneity

We measured the transcriptomic heterogeneity of cell line and treatment groups by computing all combinations of internal Euclidian distances within annotations based on their PCA loadings (Figure 4.3, Table S4). We then performed a Mann-Whitney U test and t-test between the treated and non-treated groups within cell lines. We argue that the large sample number allows for a t-test to be performed even when the compared distributions do not follow normality due to the central limit theorem.

4.5.16 Uniformization of cell line specific expression

We subset our scTransform data into Trametinib and DMSO groups and calculated the standardized variance and average expression within these groups via the FindVariableFeatures Seurat 3 command. The standardized variance accurately corrects for average expression of the gene. We then performed a Mann-Whitney U test and t-test between these groups. Finally, we performed the same analysis on the subset of unique CLGs (stringent $q<0.01$ and $|\log2foldchange|>1.5$ cutoff, 468 genes, Figure S6F, Table S6)..

4.5.17 Integration of scRNA-seq and sci-ATAC-seq data

We applied the recently described cross-data-modality integration method based in Canonical Correlation Analysis (CCA) to co-anchor our two data sets (S7C, Stuart et al., 2018). First we performed Latent Semantic Indexing (LSI) on our the filtered chromatin accessibility matrix and calculated the normalized LSI loadings scores for anchor weighting. We then identified 1,816 co-varying features via the SelectIntegrationFeatures function between and the standard normalized scRNA-seq expression matrix and the sci-ATAC-seq normalized gene activity matrix. This method yielded better integration results when we applied the FindTransferAnchor (with the parameters dims = 1:20 and reduction = ‘‘cca’’) compared to when variable features were only selected based on the scRNA-seq (3000 features) data. Similarly we found that LSI weighting out performed Cistopic based anchor weighting in the following TransferData step (weight.reduction = atac[[‘‘lsi’’]]), where scRNA-seq data labels were transferred onto sci-ATAC-seq cells. We created a confusion count matrix based on treatment and cell line label matches of the scRNA predicted and actual labels (Figure 4.4B). In addition, we plotted the sci-ATAC-seq predicted label score distributions of the sci-ATAC-seq cells for transfer quality control (Figure S7D). These scores are computed during the TransferData step from the anchor classification scores. Using a similar method for feature imputation we transferred the scRNA-seq data onto the scATAC-seq cells and performed PCA on the combined datasets, followed by visualization via UMAP (Becht et al., 2018b).

4.5.18 Ordering of cells along treatment response

We applied the R package Monocle 3 v. alpha to the imputed feature PCAs and the projected UMAP coordinates of the combined RNA ATAC data sets to order cells along a treatment response curve (Figures 4.4D, S7E-S7F). We used the negative binomial distribution to approximate size factors, gene dispersions and clustered cells into 10 groups based on the PCAs. Using RGE_method=”SimplePPT” we learned the principle graph within cell lines by forcing partition groups to contain only cells from a given cell line. We did this by omitting cells from the ordering

where outlier cells were part of a cluster associated with a different cell line. This was primarily in the HCC1143S cluster, likely due to the lower number of RNA data for these cells. Finally, we chose the root nodes visually so that pseudotemporal ordering direction was from DMSO to Trametinib treated cells. We then separately ordered just the scRNA-seq cells based on PCA and UMAP projection of their expression data and the scATAC-seq cells based on their cistopic cell-topics matrix and the UMAP projections. For the latter we approximated the size factors using the binomialff option and used DDRTree for graphing out cells. To prove that the ordering in individual modalities was conserved in the integrated dataset we performed a Spearman correlation between ATAC and combined data and RNA and the combined data (Figure S7E- S7G). We also performed this analysis where we used the cells closest to the combined ordering as roots for the individual modalities and found slightly lower correlations. In both approaches the correlation between combined and the sci-ATAC-seq datasets were lower than the scRNA-combined data set correlations. This is likely due to the ordering ATAC cells in combined datasets based on the gene activity scores as opposed to ordering based on cistopic scores of sci-ATACseq data.

4.5.19 Characterizing linked transcriptomic and epigenetic adaptation

We created a computational framework for characterizing changes in the linked epigenetic transcriptomic landscape. We selected cicero linked (>0.001) co-accessible sites with the promoter regions (<5000 bp to TSS) of the differentially expressed genes we identified using our scRNA-seq data. We calculated the Spearman Correlation between the interpolated values of the gene expression and the dropout corrected probability of accessibility at the promoter linked sites along the treatment curve of each cell line (Figures 4.6A-4.6C). We used the loess smoothing function in R for interpolation with the smoothing parameter set as 200 and the ordering of the co-embedded cells (interval set as the maximum and minimum treatment curve order of cell line cells). For each linked site we calculated the log2fold change between the average dropout corrected probability of accessibility values of Trametinib treated and DMSO treated cells and performed a Mann-Whitney

U test between the two groups. Finally, in all DA analyses we corrected for multiple testing at q=0.01 with the qvalue R package. After removing promoters, we identified high confidence dynamically changing regulatory sites putatively involved in the regulation of DE genes by first filtering for sites with abs|co-accessibility| $\geq$0.15 and |Spearman correlation along treatment| $\geq$0.5. We finally filtered for sites with at least a 25% change in accessibility (|log2fold(Tram accessibility/DMSO accessibility)| $\geq$0.322) and a q-value<0.01 (Figure 4.6D). We then performed an enrichment of these filtered sites for the topic defined cell line specific unique sites (Figure 4.6E). We used a hypergeometric test (R command phyper) to calculate the relative enrichment of high confidence filtered sites (linked to transcriptional change) in all cell line unique sites. We corrected for false discovery rate via the Benjamini-Hochberg Procedure. Finally, we used homer to analyze the transcription factors enriched in the filtered sites of a cell line relative to the DMSO topic defined unique line specific sites of that cell line (Figure 4.6F).

Chapter 5: Conclusions and future directions

5.1 General discussion

Healthy tissue function is closely tied to genomic and epigenomic heterogeneity. Genomic somatic mutations accrue during development and aging. The resulting somatic mosaicism enables advantageous neuronal diversity during early development and can give rise to clonal populations with enhanced proliferation in continuously dividing cell populations. Interestingly, as cells aggregate mutations that increase their individual fitness in cancer-related genes, so does the chance for them to turn cancerous and form tumors. At this point, genomic heterogeneity becomes advantageous for the tumor to avoid treatment pressure (Figure 1.1).

Similarly, regulation of tissue-component cell types and cell states is heterogenous and complex (Figure 1.2). Cell lineages are founded via epigenetic changes during development, such as the deposition of methylation and acetylation marks. These poise transcriptional machinery via chromatin accessibility, opening at lineage specific gene enhancers and promoters. Cell states are the perturbations of these cell lineages and are responsible for making adaptive responses to extrinsic stimuli. Cancer can exploit the developmental epigenetic processes necessary for cell lineage formation and cell state transitions to create cell populations resistant to treatment. As a result, mapping genetic and epigenetic heterogeneity has become a major advantage of single-cell studies over bulk approaches as the latter can only provide averaged profiles of the assayed cell populations. Cell atlases produced by single-cell methods capture snapshots of cell types and states present in the studied tissue. Cell atlases can also capture the progression between cell states, which computational methods can use to decipher the regulatory ordering of cells. These can be further extended by taking samples throughout development, tumorigenesis or treatment response to observe changes across the regulatory landscape. This dissertation shows the application of high-throughput combinatorial indexed strategies to profile chromatin accessibility and whole genome copy number variation across a variety of healthy and diseased tissues.

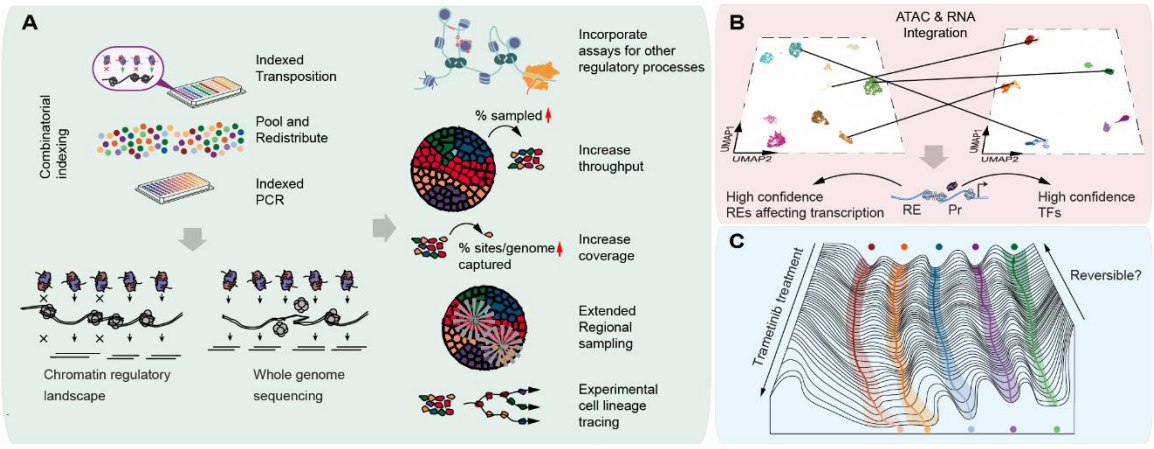

Figure 5.1 Genomic and epigenetic heterogeneity complex tissues. (A) Methodological improvements to sci-assays. (B) Computational considerations of employing integration between datasets (C) Biological questions regarding Trametinib treatment of BCCLs.

I first showed how sci-DNA-seq helped interrogate large-scale aneuploidy of the Rhesus brain, pancreatic ductal adenocarcinoma, and the cell line GM12878. Cell line and brain aneuploidy matched values reported in literature and other single-cell methods (Figures 2.2, 2.3), establishing the validity of our approach for further analysis on a tumor sample (Figure 2.4). The major advantage of sci-DNA-seq was shown on the profiled PDAC tumor, where I first identified clonal populations based on their low coverage CNV profiles, which provided the basis for aggregation within these cell populations. I developed a breakpoint-based copy number calling computational framework for this analysis which can potentially be applied for future analyses of tumor populations. This led to the detection of higher resolution (100 kbp windowed) copy number alterations at potential clone-specific driver genes, such as *IKBKB* and *PDGRFB* (Forbes et al., 2015; Perkins, 2007). Since this method was published, other single-cell DNA sequencing studies have shown how selection on preexisting clonal populations as early as pre-malignant lesions can lead to treatment resistant cells in cancer (Casasent et al., 2018; Hinohara et al., 2018; C. Kim et al., 2018a). These studies, however, outline the difficulty in determining treatment resistance on the mutational landscape alone as selection happens on the phenotypes of individual cells, which is tied to regulatory landscape on top of clonal heterogeneity. Therefore, measuring somatic heterogeneity, regulatory changes, and phenotype in parallel is important to further understand

acquired treatment resistance (Hinohara & Polyak, 2019b). Assays such as the suspended microchannel resonator can utilize scRNA-seq and growth measurements in parallel to help elucidate these processes (Kimmerling et al., 2018).

In Chapter 3, I showed how using sci-ATAC-seq can map these steady state regulatory processes by capturing the chromatin accessibility landscape of the healthy murine hippocampus (Figures 3.1-3.5). This revealed high-resolution separation of cell types, including rare populations of microglia and oligodendrocyte progenitors. The necessity of improving computational approaches of dimensionality reduction was exhibited by the application of CisTopic (Bravo González-Blas et al., 2019) to the pyramidal neuronal populations, which helped identify previously not seen CA1 and CA2 neurons. Later analysis of co-regulated chromatin hubs and TF accessibility revealed cell type specific chromatin architecture and potential transcription factors playing a role in their regulation. Finally, the separation of *in vitro* cultured neurons and *in vivo* neuronal populations revealed large regulatory differences between corresponding neuronal groups, indicating that epigenetic biases are introduced during neuron culturing conditions.

In Chapter 4, I showed how sci-ATAC-seq can map the dynamic process of epigenetic adaptation of genetically distinct basal like triple-negative breast cancer to Trametinib treatment (Figures 4.1-4.7). While our previous study (Risom, 2017) indicated a shared response drug resistant persistors across the studied five cell lines, my initial analyses showed no matching epigenetic change. However, differential accessibility indicated that the epigenetic distance between cell lines defined by chromatin accessibility decreases upon treatment. The analysis of involved sites revealed a preferential homogenization of cell-type specific sites. This phenomenon was also replicated in transcriptomic space indicating a contraction in epigenetic and transcriptomic space upon treatment. In order to understand it further, I developed a computational pipeline to match changes in chromatin regulation with shifts observed in the transcriptome upon treatment. These revealed an enrichment of cell line specific regulatory elements in sites that are linked to

changed transcription. This phenomenon is not well understood, but offers potential therapeutic benefits and opens up opportunities for later studies.

5.2 Methodological Considerations

The studies shown in this dissertation emphasize the importance of methodological improvements across assays via increasing throughput, read coverage, integration of other assays, spatial information, lineage tracing, and the development of novel computational methods.

5.2.1 Improving combinatorial indexed assays

Combinatorial indexing is a highly scalable technology which is evolving fast to incorporate assays to show genetic and epigenetic heterogeneity across single cells (Figure 5.1A). Combinatorial indexing strategies have been applied to profile the chromatin accessibility (D. A. Cusanovich et al., 2015), transcriptome (Cao et al., 2017, 2019b), whole genome (Vitak et al., 2017a), and chromosome conformation capture (Ramani et al., 2017, 2020) of single cells. While initial strategies have been 96-well plate based, recent hybrid droplet-based methods have further improved throughput in droplet based sci-ATAC-seq (Lareau et al., 2019), sciRNA-seq (Datlinger et al., 2019), and led to novel methods such as single-cell combinatorial indexed cytometry (Hwang et al., 2020). In addition, pushes have been made to profile regulatory features in parallel, such as joint parallel profiling of RNA and ATAC (Cao et al., 2018).

In addition to increasing throughput, increasing read coverage of individual cells has been central to these assays. Decreasing the sparsity of single-cell ATAC-seq data can help improve separation of cell types and states, along with capturing more fine transitions between cell states. Improvements such as the method introduced in Chapter 2 (Sinnamon et al., 2019b), omni-ATAC-seq (Corces et al., 2017), and scip-ATAC-seq (Mulqueen et al., 2019) have primarily focused on improving the effectiveness of the Tn5 transposition. This can be done by decreasing contaminating sequences, such as mitochondrial DNA and by making nuclear entry more permissible for the transposase. Another aspect of ATAC-seq data is the difficulty in distinguishing missing data from

closed chromatin. One method addresses this experimentally via the use of a GpC methyltransferase to label sites in nucleosome depleted DNA before a bisulfite sequencing. As GpC dinucleotides are abundant but are also not frequently methylated in the mammalian genome, Nucleosome Occupancy and Methylation sequencing (NOME-seq) can provide parallel endogenous methylation patters via CpG methylation and fine scale chromatin accessibility information through GpC methylation (Kelly et al., 2012). Recently, a single-cell adaptation of this approach was published (Pott, 2017) followed by an assay which profiles the transcriptome of assayed cells as well (Clark et al., 2018).

Increasing read coverage of single cells has been as important for single-cell DNA sequencing methods as in single-cell ATAC-seq. However, as opposed to chromatin accessibility, the uniformity of genome amplification is necessary for unbiased copy number calling. While sci-DNA-seq is high throughput, it does not incorporate a genome amplification method, which limits our ability to call single nucleotide polymorphisms accurately. Other existing single-cell DNA sequencing methods including whole genome amplification steps have been plagued by low throughput until recently. The incorporation of Linear amplification via transposon insertion (C. Chen et al., 2017) into a combinatorial indexed schema successfully profiled mouse sperm rare chromosome mis-segregation events in meiosis (Yin et al., 2019) .

While single-cell assays capture the granularity of complex tissues, they do not inherently record spatial information. As both complex healthy tissues and solid tumors rely on spatial organization for normal functionality or disease progression, there is a need for spatially resolved single-cell assays. Indeed, identification of CA1, CA2 and CA3 pyramidal neuronal populations in the hippocampus could have been verified with spatially resolved data. In addition, the evolution of clonal populations within a tumor can be very spatially restricted as exhibited by bulk methods (Dou et al., 2018; X. Li et al., 2018; Williams et al., 2018). Recent advances in *in situ* hybridization (ISH) technologies and *in situ* RNA sequencing show close to single-cell resolution (Eng et al., 2019; Rodriques et al., 2019; Ståhl et al., 2016; Vickovic et al., 2019). Computational methods

using the integration of ISH atlases of gene expression with scRNA-seq have also shown promise in mapping rare cell types (Achim et al., 2015; Moncada et al., 2020; Satija et al., 2015). These methods however are often limited in throughput, and cannot profile nuclear genomic and epigenetic properties. Alternative methods use targeted capture to spatially map cells from cryo-sectioned samples as exhibited in spatially resolved single-cell whole genome sequencing of *in situ* and invasive ductal carcinomas (Casasent et al., 2018).

Single-cell combinatorial indexed ATAC sequencing was recently expanded (Thornton et al., 2019) to include spatial information via the first round of transposase-based indexing (Figure 5.1D). This allowed for the spatial profiling of the mouse cortex and the pseudospatial ordering of glutamatergic neurons in the somatosensory cortex. This assay has not been extended to include single-cell whole genome sequencing but could be incorporated via the addition of nucleosome depletion strategies presented in section 2.3.1.

Understanding lineage relationships is important for deciphering the order of cell state transitions during development or tumorigenesis. Computational methods listed in section 1.2.4 can order cells or similarly behaving cell groups by assuming each to be a low-quality snapshot of a ground truth distribution of cells. Based on the statistical inference of the error and the order of these snapshots can reveal dynamic chromatin accessibility gene expression changes. These, however, are approximations of the ground truth and can often overfit data, resulting in conflicting trajectories on identical data (Weinreb et al., 2018). Therefore, we should be careful when considering the ordering of the inferred treatment trajectory in Chapter 4. One promising aspect, however, was the high correlation between the combined and single assay only ordering. The lower correlation of sci-ATAC-seq only and the combined dataset orderings was likely due the differences between ordering on gene activity scores as opposed to ordering on accessibility peak changes. This creates a possibility of developing a method where, after the initial ordering of gene activity, chromatin accessibly changes between cells with the same accessibility could be considered to fine tune within trajectory ordering.

Experimental lineage tracing is also an alternative to provide reliable information on cell lineages. These can be genetically encoded fluorescent protein; however, this method only works at low cell numbers due to the limitation in spectral resolution. By introducing prospective markers via different inducible (Sleeping Beauty transposase, Cre–loxP and CRISP–Cas9 systems) lineage relationships can be conserved. Introducing heritable genetic barcodes with the CRISP–Cas9 system can trace progeny of the initial cells, as novel induced indels become permanent additions to the original barcode (Baron & van Oudenaarden, 2019). Methods differ in how the introduction of these indels happens; scGESTALT (Raj et al., 2018) relies on a heat shock induced system, scScarTrace (Alemany et al., 2018) and LINNEUS (Spanjaard et al., 2018) rely on the injection of Cas9 RNA or protein in the one cell stage. Whereas read out of barcodes happens on the level of RNA for LINNEUS and scGESTALT, scScarTrace requires a separate DNA sequencing step (Baron & van Oudenaarden, 2019). These methods require complex genetic manipulation at the beginning of the experiment, and are model system dependent, as opposed to more easily deployable virus-based label delivery systems. However, up until recently, the latter had the limitation of not capturing progression. The CellTag system allows for progressive labeling via multiplexing with introduced indexes each round. Such a system could of benefit to understanding acquired drug resistance (Biddy et al., 2018; Guo et al., 2019; Kong et al., 2020). Retrospective tracing of somatic mutations could also reveal lineage relationships. As the mutational frequency in mitochondria is 10-100 times that of genomic DNA, it has been suggested as a natural DNA barcode (Baron & van Oudenaarden, 2019). A recent study showed that mitochondrial reads obtained from scATAC-seq or scRNA-seq can show the lineage relationship of healthy and leukaemic haematopoietic cells (Ludwig et al., 2019). This could potentially be applied in drug resistance studies; however, potential mutational selective sweeps should be taken into consideration.

5.2.2 Computational considerations

scATAC-seq provides a wealth of data which can allow for the separation of cell types not captured by scRNA-seq. This, however, often comes at the price of interpretability and highlights the importance of finding reliable markers in accessibility to annotate cell populations. While methods described in section 1.2.4 can approximate linkage to nearby genes, this does not necessarily mean changes in expression. Massively parallel enhancer reporter assays have directly contributed to our understanding of evolutionarily conserved enhancers (Inoue & Ahituv, 2015) and the GeneHancer project undertook creating a curated set of high confidence enhancer based on eQTLs, eRNA co-expression, TF co-expression, and capture Hi-C (CHi-C) data (Fishilevich et al., 2017). The method I developed for linking single-cell RNA-seq and ATAC-seq through the integration of single-cell assays adds to this by finding repressors and enhancers based on shared trends between accessibility and expression (Figure 5.1B). It also showed how a large number of regulatory elements of genes are missed based on co-accessibility using Cicero (Pliner et al., 2018b), which is likely due to promoters not changing in accessibility, such as constitutive promoters. This method still could be improved by the use of cell line specific co-accessibility, which could further refine the regulatory classification laid out in Figure 4.6. Also, next to Spearman correlation across the treatment trajectory, logistic regression could be used to predict linkage as in the latest release of SnapATAC (Fang et al., 2019).

The studies I presented in chapters 3 and 4 showed cell type- and cell state-specific putative transcription factor accessibility. Methods used in single-cell ATAC-seq, such as chromVAR and Homer, rely on motif binding as a proxy for TF activity. Verifying these TFs can be done via ChiP-seq (Barski et al., 2007; Johnson et al., 2007) and CUT&TAG seq (Kaya-Okur et al., 2019). The single-cell adaptation of ChIP-seq (Grosselin et al., 2019) is inherently low coverage per single cell, and CUT&TAG methods are still under development. As I have shown in chapter 4, the integration of single-cell accessibility assays could also contribute here as well since correlation of changes in expression and putative activity can be used to create lists of high confidence TFs at sites. One thing to note is how AP-1 complex transcription factor members presented as highly

variable across all studies, which is likely due to it representing cells under stress during the preparation of nuclei for sci-ATAC-seq library. However, most TFs associated with the AP-1 complex dropped out when I used the peaks identified as high confidence in the integrated single-cell RNA and ATAC analysis.

5.3 Future directions

The studies presented in this dissertation lay out several possible future directions of research. For methodological improvements, the integration of spatial information into sciDNA-seq could greatly enhance our ability to assay subclonal tumoral populations that are topologically restricted. Also, the potential information resulting from non-tumoral cells sampled in the vicinity of topological samples could provide information about mutational profiles in different microenvironments. Furthermore, topologically mapping metastases from the same patient could help elucidate foundational cell populations and the mutational/spatial organization necessary for successful metastasis formation. While the introduction of spatial information to the assay could be done based on the work of (Thornton et al., 2019), the preparation of tumor samples would require the development of novel techniques due to the difficulty in disaggregation (Figure 5.1D). In addition, the computational method developed for identifying linked regulatory elements with gene expression changes based on integrated scRNA-seq and scATAC-seq could be used in large single-cell atlas studies spanning multiple tissues to find high confidence tissue specific regulatory elements and the potential transcription factors controlling them. Similarly, this could be used in drug resistance trials. For instance, based on the results of Risom et al., 2018, inhibiting BET protein activity via JQ1 prevents BEZ235 drug-induced chromatin changes. Follow up computational analysis of integrated datasets could help identify the sites blocked by JQ1 that are required for drug resistance.

The results shown in chapter 4 also posit several biological questions (Figure 5.1C). Does the homogenization between cell states give us information about potential windows of opportunity for combination treatments? Are acquired resistant states reversible? Would the homogenization

between cell lines reverse as well? Sci-ATAC-seq performed after washing out of Trametinib (data not shown) will answer some of these questions after further analysis.

References

Abusaad, I., MacKay, D., Zhao, J., Stanford, P., Collier, D. A., & Everall, I. P. (1999). Stereological estimation of the total number of neurons in the murine hippocampus using the optical disector. *Journal of Comparative Neurology*, *408*(4), 560–566. https://doi.org/10.1002/(SICI)1096-9861(19990614)408:4<560::AID-CNE9>3.0.CO;2-P

Achim, K., Pettit, J. B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., & Marioni, J. C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*. https://doi.org/10.1038/nbt.3209

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., Qiu, R., Lee, C., & Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*. https://doi.org/10.1038/nature12064

Adey, A., Kitzman, J. O., Burton, J. N., Daza, R., Kumar, A., Christiansen, L., Ronaghi, M., Amini, S., Gunderson, K. L., Steemers, F. J., & Shendure, J. (2014). In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Research*. https://doi.org/10.1101/gr.178319.114

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., & Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*. https://doi.org/10.1186/gb-2010-11-12-r119

Adjei, A. A., Cohen, R. B., Franklin, W., Morris, C., Wilson, D., Molina, J. R., Hanson, L. J., Gore, L., Chow, L., Leong, S., Maloney, L., Gordon, G., Simmons, H., Marlow, A., Litwiler, K., Brown, S., Poch, G., Kane, K., Haney, J., & Eckhardt, S. G. (2008). Phase I pharmacokinetic and pharmacodynamic study of the oral, small-molecule mitogen-activated protein kinase kinase 1/2 inhibitor AZD6244 (ARRY-142886) in patients with advanced cancers. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2007.14.4956

Ahsendorf, T., Müller, F. J., Topkar, V., Gunawardena, J., & Eils, R. (2017). Transcription factors, coregulators, and epigenetic marks are linearly correlated and highly redundant. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0186324

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). {SCENIC}: single-cell regulatory network inference and clustering. *Nat. Methods*, *14*(11), 1083–1086.

Al-Khallaf, H. (2017). Isocitrate dehydrogenases in physiology and cancer: biochemical and molecular insight. *Cell Biosci.*, *7*, 37.

Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J., & Van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature*. https://doi.org/10.1038/nature25969

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H., & Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*. https://doi.org/10.1126/science.aag0299

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N.,

Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Alexandrov, L. B., … Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*. https://doi.org/10.1016/j.celrep.2012.12.008

Almendro, V., Marusyk, A., & Polyak, K. (2013). Cellular Heterogeneity and Molecular Evolution in Cancer. *Annual Review of Pathology: Mechanisms of Disease*. https://doi.org/10.1146/annurev-pathol-020712-163923

Altschuler, S. J., & Wu, L. F. (2010). Cellular Heterogeneity: Do Differences Make a Difference? In *Cell*. https://doi.org/10.1016/j.cell.2010.04.033

American Cancer Society. (2019a). Breast Cancer Facts & Figures 2019-2020. *American Cancer Society*.

American Cancer Society. (2019b). Facts & Figures 2019. *American Cancer Society*.

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K. L., & Steemers, F. J. (2014a). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature Genetics*, *46*(12), 1343–1349.

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K. L., & Steemers, F. J. (2014b). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature Genetics*. https://doi.org/10.1038/ng.3119

Audia, J. E., & Campbell, R. M. (2016). Histone modifications and cancer. *Cold Spring Harbor Perspectives in Biology*. https://doi.org/10.1101/cshperspect.a019521

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. In *Nature*. https://doi.org/10.1038/nature15393

Bach, K., Pensa, S., Grzelak, M., Hadfield, J., Adams, D. J., Marioni, J. C., & Khaled, W. T. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications*. https://doi.org/10.1038/s41467-017-02001-5

Bachmann, I. M., Halvorsen, O. J., Collett, K., Stefansson, I. M., Straume, O., Haukaas, S. A., Salvesen, H. B., Otte, A. P., & Akslen, L. A. (2006). EZH2 expression is associated with high proliferation rate and aggressive tumor subgroups in cutaneous melanoma and cancers of the endometrium, prostate, and breast. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2005.01.5180

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K. S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Karchin, R. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. https://doi.org/10.1016/j.cell.2018.02.060

Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A. M., Gingras, M. C., Miller, D. K., Christ, A. N., Bruxner, T. J. C., Quinn, M. C., Nourse, C., Murtaugh, L. C., Harliwong, I., Idrisoglu, S., Manning, S., Nourbakhsh, E., Wani, S., Fink, L., Holmes, O., … Grimmond,

S. M. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*. https://doi.org/10.1038/nature16965

Bakker, B., Taudt, A., Belderbos, M. E., Porubsky, D., Spierings, D. C. J., de Jong, T. V., Halsema, N., Kazemier, H. G., Hoekstra-Wakker, K., Bradley, A., de Bont, E. S. J. M., van den Berg, A., Guryev, V., Lansdorp, P. M., Colomé-Tatché, M., & Foijer, F. (2016). Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biology*. https://doi.org/10.1186/s13059-016-0971-7

Banyard, J., & Bielenberg, D. R. (2015). The role of EMT and MET in cancer dissemination. In *Connective Tissue Research*. https://doi.org/10.3109/03008207.2015.1060970

Baron, C. S., & van Oudenaarden, A. (2019). Unravelling cellular relationships during development and regeneration using genetic lineage tracing. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-019-0186-3

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. https://doi.org/10.1038/nature11003

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*. https://doi.org/10.1016/j.cell.2007.05.009

Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B., Wigler, M., Navin, N., & Hicks, J. (2012). Genome-wide copy number analysis of single cells. *Nature Protocols*, *7*(6), 1024–1041. https://doi.org/10.1038/nprot.2012.039

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018a). Dimensionality reduction for visualizing single-cell data using {UMAP}. *Nat. Biotechnol.*

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018b). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Beksac, A. T., Paulucci, D. J., Blum, K. A., Yadav, S. S., Sfakianos, J. P., & Badani, K. K. (2017). Heterogeneity in renal cell carcinoma. In *Urologic Oncology: Seminars and Original Investigations*. https://doi.org/10.1016/j.urolonc.2017.05.006

Benner, C., Heinz, S., & Glass, C. K. (2017). *HOMER - Software for motif discovery and next generation sequencing analysis*. Http://Homer.Ucsd.Edu/.

Berger, M. F., Lawrence, M. S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A. Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., Onofrio, R., Carter, S. L., Park, K., Habegger, L., Ambrogio, L., Fennell, T., Parkin, M., Saksena, G., Voet, D., … Garraway, L. A. (2011). The genomic complexity of primary human prostate cancer. *Nature*. https://doi.org/10.1038/nature09744

Bergmann, O., Spalding, K. L., & Frisén, J. (2015). Adult neurogenesis in humans. *Cold Spring Harbor Perspectives in Medicine*. https://doi.org/10.1101/cshperspect.a018994

Bernardo, G. M., Bebek, G., Ginther, C. L., Sizemore, S. T., Lozada, K. L., Miedler, J. D., Anderson, L. A., Godwin, A. K., Abdul-Karim, F. W., Slamon, D. J., & Keri, R. A. (2013). {FOXA1} represses the molecular phenotype of basal breast cancer cells. *Oncogene*, *32*(5),

554–563.

Bertram, J. S. (2000). The molecular biology of cancer. In *Molecular Aspects of Medicine*. https://doi.org/10.1016/S0098-2997(00)00007-8

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., & Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*. https://doi.org/10.3109/03014460.2013.807878

Bickhart, D. M., & Liu, G. E. (2014). The challenges and importance of structural variation detection in livestock. In *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2014.00037

Biddy, B. A., Kong, W., Kamimoto, K., Guo, C., Waye, S. E., Sun, T., & Morris, S. A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. *Nature*. https://doi.org/10.1038/s41586-018-0744-4

Boström, P., Söderström, M., Vahlberg, T., Söderström, K. O., Roberts, P. J., Carpén, O., & Hirsimäki, P. (2011). MMP-1 expression has an independent prognostic value in breast cancer. *BMC Cancer*. https://doi.org/10.1186/1471-2407-11-348

Brady, S. W., McQuerry, J. A., Qiao, Y., Piccolo, S. R., Shrestha, G., Jenkins, D. F., Layer, R. M., Pedersen, B. S., Miller, R. H., Esch, A., Selitsky, S. R., Parker, J. S., Anderson, L. A., Dalley, B. K., Factor, R. E., Reddy, C. B., Boltax, J. P., Li, D. Y., Moos, P. J., … Bild, A. H. (2017). Combating subclonal evolution of resistant cancer phenotypes. *Nature Communications*. https://doi.org/10.1038/s41467-017-01174-3

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., & Aerts, S. (2018). Cis-topic modelling of single cell epigenomes. *BioRxiv*, 370346. https://doi.org/10.1101/370346

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., & Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell {ATAC-seq} data. *Nat. Methods*, *16*(5), 397–400.

Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., Aryee, M. J., Majeti, R., Chang, H. Y., & Greenleaf, W. J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*, *173*(6), 1535-1548.e16. https://doi.org/10.1016/j.cell.2018.03.074

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013a). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*. https://doi.org/10.1038/nmeth.2688

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013b). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486–490. https://doi.org/10.1038/nature14590

Bullitt, E. (1990). Expression of C-fos-like protein as a marker for neuronal activity following noxious stimulation in the rat. *Journal of Comparative Neurology*, *296*(4), 517–530. https://doi.org/10.1002/cne.902960402

Bushman, D. M., & Chun, J. (2013). The genomically mosaic brain: Aneuploidy and more in neural diversity and disease. In *Seminars in Cell and Developmental Biology*. https://doi.org/10.1016/j.semcdb.2013.02.003

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*. https://doi.org/10.1038/nbt.4096

Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A., & Walsh, C. A. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Reports*. https://doi.org/10.1016/j.celrep.2014.07.043

Callaway, E. (2014). Platinum' genome takes on disease. *Nat. News*, 515, 323.

Campbell, I. M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M. E., Nagamani, S. C. S., Erez, A., Bartnik, M., Wiśniowiecka-Kowalnik, B., Plunkett, K. S., Pursley, A. N., Kang, S. H. L., Bi, W., Lalani, S. R., Bacino, C. A., Vast, M., Marks, K., Patton, M., … Stankiewicz, P. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2014.07.003

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C. L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … Zhang, J. (2020). Pan-cancer analysis of whole genomes. *Nature*. https://doi.org/10.1038/s41586-020-1969-6

Cancer, T., & Atlas, G. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia The Cancer Genome Atlas Research Network. *The New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa1301689

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. https://doi.org/10.1126/science.aau0730

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., & Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, *357*(6352), 661–667. https://doi.org/10.1126/science.aam8940

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019a). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019b). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502.

Carey, L. A. (2011). Directed Therapy of Subtypes of Triple-Negative Breast Cancer. *The Oncologist*. https://doi.org/10.1634/theoncologist.2011-s1-71

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., & Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*. https://doi.org/10.1038/nbt.2203

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2015.25

Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M. E., & Navin, N. E. (2018). Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell*. https://doi.org/10.1016/j.cell.2017.12.007

Cembrowski, M. S., Wang, L., Sugino, K., Shields, B. C., & Spruston, N. (2016). Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. *ELife*, *5*, e14997. https://doi.org/10.7554/eLife.14997

Chaffer, C. L., Brueckmann, I., Scheel, C., Kaestli, A. J., Wiggins, P. A., Rodrigues, L. O., Brooks, M., Reinhardt, F., Suc, Y., Polyak, K., Arendt, L. M., Kuperwasser, C., Bierie, B., & Weinberg, R. A. (2011). Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1102454108

Chatterjee, N., & Walker, G. C. (2017). Mechanisms of DNA damage, repair, and mutagenesis. In *Environmental and Molecular Mutagenesis*. https://doi.org/10.1002/em.22087

Cheah, P.-S., & Thomas, P. Q. (2015). SOX3 expression in the glial system of the developing and adult mouse cerebellum. *SpringerPlus*, *4*, 400. https://doi.org/10.1186/s40064-015-1194-1

Cheang, M. C. U., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P. S., Parker, J. S., Perou, C. M., Ellis, M. J., & Nielsen, T. O. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/djp082

Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., & Xie, X. S. (2017). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science*. https://doi.org/10.1126/science.aak9787

Chen, D.-G., Zhu, B., Lv, S.-Q., Zhu, H., Tang, J., Huang, C., Li, Q., Zhou, P., Wang, D.-L., & Li, G.-H. (2017). Inhibition of {EGR1} inhibits glioma proliferation by targeting {CCND1} promoter. In *Journal of Experimental & Clinical Cancer Research* (Vol. 36, Issue 1).

Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., Andrade-Navarro, M. A., Buenrostro, J. D., & Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*. https://doi.org/10.1186/s13059-019-1854-5

Chen, J., Miller, B. F., & Furano, A. V. (2014). Repair of naturally occurring mismatches can induce mutations in flanking DNA. *ELife*. https://doi.org/10.7554/eLife.02001.001

Chen, X., Miragaia, R. J., Natarajan, K. N., & Teichmann, S. A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. *Nature Communications*. https://doi.org/10.1038/s41467-018-07771-0

Chénais, B., Caruso, A., Hiard, S., & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. In *Gene*. https://doi.org/10.1016/j.gene.2012.07.042

Chittock, E. C., Latwiel, S., Miller, T. C. R., & Müller, C. W. (2017). Molecular architecture of polycomb repressive complexes. In *Biochemical Society Transactions*. https://doi.org/10.1042/BST20160173

Choukrallah, M. A., & Matthias, P. (2014). The interplay between chromatin and transcription

factor networks during B cell development: Who pulls the trigger first? In *Frontiers in Immunology*. https://doi.org/10.3389/fimmu.2014.00156

Chung, C. Y., Ma, Z., Dravis, C., Preissl, S., Poirion, O., Luna, G., Hou, X., Giraddi, R. R., Ren, B., & Wahl, G. M. (2019). Single-Cell Chromatin Analysis of Mammary Gland Development Reveals Cell-State Transcriptional Regulators and Lineage Relationships. *Cell Reports*. https://doi.org/10.1016/j.celrep.2019.08.089

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., & Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*. https://doi.org/10.1038/nbt.2514

Clapier, C. R., & Cairns, B. R. (2009). The biology of chromatin remodeling complexes. In *Annual Review of Biochemistry*. https://doi.org/10.1146/annurev.biochem.77.062706.153223

Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., & Reik, W. (2018). ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nature Communications*. https://doi.org/10.1038/s41467-018-03149-4

Claus Stolt, C., Rehberg, S., Ader, M., Lommes, P., Riethmacher, D., Schachner, M., Bartsch, U., & Wegner, M. (2002). Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10. *Genes and Development*, *16*(2), 165–170. https://doi.org/10.1101/gad.215802

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., & Yosef, N. (2019). Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*. https://doi.org/10.1016/j.cels.2019.03.010

Collings, C. K., & Anderson, J. N. (2017). Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenetics and Chromatin*. https://doi.org/10.1186/s13072-017-0125-5

Conrad, D. F., Keebler, J. E. M., Depristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., & Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*. https://doi.org/10.1038/ng.862

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R., & Chang, H. Y. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, *48*(10), 1193–1203. https://doi.org/10.1038/ng.3646

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., … Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*. https://doi.org/10.1038/nmeth.4396

Core Team, R. (2019). *R: A Language and Environment for Statistical Computing*.

Cros, J., Raffenne, J., Couvelard, A., & Poté, N. (2018). Tumor Heterogeneity in Pancreatic

Adenocarcinoma. *Pathobiology*. https://doi.org/10.1159/000477773

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. https://doi.org/10.1126/science.aab1601

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., & Shendure, J. (2018). A {Single-Cell} Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, *174*(5), 1309--1324.e18.

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H. A., Christiansen, L., Qiu, X., Steemers, F. J., Trapnell, C., Shendure, J., & Furlong, E. E. M. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, *555*, 538.

Cusanovich, D. a, Daza, R., Adey, A., Pliner, H. a, Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (New York, N.Y.)*, *348*(6237), 910–914.

Datlinger, P., Rendeiro, A. F., Boenke, T., Krausgruber, T., Barreca, D., & Bock, C. (2019). *Ultra-high throughput single-cell RNA sequencing by combinatorial fluidic indexing*. 1–27. https://doi.org/10.1101/2019.12.17.879304

Davis, R. T., Blake, K., Ma, D., Gabra, M. B. I., Hernandez, G. A., Phung, A. T., Yang, Y., Maurer, D., Lefebvre, A. E. Y. T., Alshetaiwi, H., Xiao, Z., Liu, J., Locasale, J. W., Digman, M. A., Mjolsness, E., Kong, M., Werb, Z., & Lawson, D. A. (2020). Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. *Nature Cell Biology*. https://doi.org/10.1038/s41556-020-0477-0

De Bourcy, C. F. A., De Vlaminck, I., Kanbar, J. N., Wang, J., Gawad, C., & Quake, S. R. (2014). A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0105585

De Kouchkovsky, I., & Abdul-Hay, M. (2016). 'Acute myeloid leukemia: A comprehensive review and 2016 update.' In *Blood Cancer Journal*. https://doi.org/10.1038/bcj.2016.50

Dean, F. B., Nelson, J. R., Giesler, T. L., & Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research*. https://doi.org/10.1101/gr.180501

Deleye, L., Tilleman, L., Van Der Plaetsen, A. S., Cornelis, S., Deforce, D., & Van Nieuwerburgh, F. (2017). Performance of four modern whole genome amplification methods for copy number variant detection in single cells. *Scientific Reports*. https://doi.org/10.1038/s41598-017-03711-y

Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P., & Narod, S. A. (2007). Triple-negative breast cancer: Clinical features and patterns of recurrence. *Clinical Cancer Research*. https://doi.org/10.1158/1078-0432.CCR-06-3045

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin

interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

dos Santos, C. O., Dolzhenko, E., Hodges, E., Smith, A. D., & Hannon, G. J. (2015). An Epigenetic Memory of Pregnancy in the Mouse Mammary Gland. *Cell Reports*. https://doi.org/10.1016/j.celrep.2015.04.015

Dou, Y., Gold, H. D., Luquette, L. J., & Park, P. J. (2018). Detecting Somatic Mutations in Normal Cells. In *Trends in Genetics*. https://doi.org/10.1016/j.tig.2018.04.003

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*. https://doi.org/10.1038/nature11247

Durinck, S., Ho, C., Wang, N. J., Liao, W., Jakkula, L. R., Collisson, E. A., Pons, J., Chan, S. W., Lam, E. T., Chu, C., Park, K., Hong, S. woo, Hur, J. S., Huh, N., Neuhaus, I. M., Yu, S. S., Grekin, R. C., Mauro, T. M., Cleaver, J. E., … Cho, R. J. (2011). Temporal dissection of tumorigenesis in primary cancers. *Cancer Discovery*. https://doi.org/10.1158/2159-8290.CD-11-0028

Efremova, M., & Teichmann, S. A. (2020). Computational methods for single-cell omics across modalities. In *Nature Methods*. https://doi.org/10.1038/s41592-019-0692-4

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A. J. L., Lefebvre, C., Bashashati, A., … Aparicio, S. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*. https://doi.org/10.1038/nature13952

Elbert, A., Vogt, D., Watson, A., Levy, M., Jiang, Y., Brûlé, E., Rowland, M. E., Rubenstein, J., & Bérubé, N. G. (2019). CTCF Governs the Identity and Migration of MGE-Derived Cortical Interneurons. *The Journal of Neuroscience*, *39*(1), 177 LP – 192. https://doi.org/10.1523/JNEUROSCI.3496-17.2018

Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G. C., & Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*. https://doi.org/10.1038/s41586-019-1049-y

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226–231. https://doi.org/10.1.1.71.1980

Fages, C., Khelil, M., Rolland, B., Bridoux, A., & Tardy, M. (1988). Glutamine synthetase: a marker of an astroglial subpopulation in primary cultures of defined brain areas. *Dev. Neurosci.*, *10*(1), 47–56.

Fan, X., Edrisi, M., Navin, N., & Nakhleh, L. (2019). *Methods for Copy Number Aberration Detection from Single-cell DNA Sequencing Data*. bioRxiv. https://doi.org/10.1101/696179

Fang, R., Preissl, S., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A. K., Mukamel, E. A., Zhang, Y., Behrens, M. M., Ecker, J., & Ren, B. (2019). Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *BioRxiv*. https://doi.org/10.1101/615179

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., & Cohen, D. (2017). GeneHancer: genome-wide

integration of enhancers and target genes in GeneCards. *Database : The Journal of Biological Databases and Curation*. https://doi.org/10.1093/database/bax028

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., & Campbell, P. J. (2015). COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1075

Fouad, T. M., Kogawa, T., Liu, D. D., Shen, Y., Masuda, H., El-Zein, R., Woodward, W. A., Chavez-MacGregor, M., Alvarez, R. H., Arun, B., Lucci, A., Krishnamurthy, S., Babiera, G., Buchholz, T. A., Valero, V., & Ueno, N. T. (2015). Overall survival differences between patients with inflammatory and noninflammatory breast cancer presenting with distant metastasis at diagnosis. *Breast Cancer Research and Treatment*. https://doi.org/10.1007/s10549-015-3436-x

Freed, D., Stevens, E. L., & Pevsner, J. (2014). Somatic mosaicism in the human genome. In *Genes*. https://doi.org/10.3390/genes5041064

Fu, Y., Zhang, F., Zhang, X., Yin, J., Du, M., Jiang, M., Liu, L., Li, J., Huang, Y., & Wang, J. (2019). High-throughput single-cell whole-genome amplification through centrifugal emulsification and eMDA. *Communications Biology*. https://doi.org/10.1038/s42003-019-0401-y

Fyodorov, D. V., Zhou, B. R., Skoultchi, A. I., & Bai, Y. (2018). Emerging roles of linker histones in regulating chromatin structure and function. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm.2017.94

Gallinari, P., Di Marco, S., Jones, P., Pallaoro, M., & Steinkühler, C. (2007). HDACs, histone deacetylation and gene transcription: From molecular biology to cancer therapeutics. *Cell Research*. https://doi.org/10.1038/sj.cr.7310149

Gao, R., Davis, A., McDonald, T. O., Sei, E., Shi, X., Wang, Y., Tsai, P. C., Casasent, A., Waters, J., Zhang, H., Meric-Bernstam, F., Michor, F., & Navin, N. E. (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics*. https://doi.org/10.1038/ng.3641

Gao, R., Kim, C., Sei, E., Foukakis, T., Crosetto, N., Chan, L. K., Srinivasan, M., Zhang, H., Meric-Bernstam, F., & Navin, N. (2017). Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-00244-w

García-Nieto, P. E., Morrison, A. J., & Fraser, H. B. (2019). The somatic mutation landscape of the human body. *Genome Biology*. https://doi.org/10.1186/s13059-019-1919-5

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G. S., Hicks, J., Wigler, M., & Schatz, M. C. (2015). Interactive analysis and assessment of single-cell copy-number variations. In *Nature Methods*. https://doi.org/10.1038/nmeth.3578

Gascard, P., Bilenky, M., Sigaroudinia, M., Zhao, J., Li, L., Carles, A., Delaney, A., Tam, A., Kamoh, B., Cho, S., Griffith, M., Chu, A., Robertson, G., Cheung, D., Li, I., Heravi-Moussavi, A., Moksa, M., Mingay, M., Hussainkhel, A., … Hirst, M. (2015a). Epigenetic and transcriptional determinants of the human breast. *Nature Communications*. https://doi.org/10.1038/ncomms7351

Gascard, P., Bilenky, M., Sigaroudinia, M., Zhao, J., Li, L., Carles, A., Delaney, A., Tam, A., Kamoh, B., Cho, S., Griffith, M., Chu, A., Robertson, G., Cheung, D., Li, I., Heravi-

Moussavi, A., Moksa, M., Mingay, M., Hussainkhel, A., … Hirst, M. (2015b). Epigenetic and transcriptional determinants of the human breast. *Nat. Commun.*, *6*, 6351.

Gaspar, J. M. (2018). Improved peak-calling with MACS2. *BioRxiv*. https://doi.org/10.1101/496521

Gawad, C., Koh, W., & Quake, S. R. (2014). Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1420822111

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2015.16

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vázquez-García, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., … Wedge, D. C. (2020). The evolutionary history of 2,658 cancers. *Nature*. https://doi.org/10.1038/s41586-019-1907-7

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D., & Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*. https://doi.org/10.1038/nmeth.2869

Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L. H., & Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*, *518*(7539), 365–369. https://doi.org/10.1038/nature14252

Goldhirsch, A., Winer, E. P., Coates, A. S., Gelber, R. D., Piccart-Gebhart, M., Thürlimann, B., Senn, H. J., Albain, K. S., André, F., Bergh, J., Bonnefoi, H., Bretel-Morales, D., Burstein, H., Cardoso, F., Castiglione-Gertsch, M., Coates, A. S., Colleoni, M., Costa, A., Curigliano, G., … Wood, W. C. (2013). Personalizing the treatment of women with early breast cancer: Highlights of the st gallen international expert consensus on the primary therapy of early breast Cancer 2013. *Annals of Oncology*. https://doi.org/10.1093/annonc/mdt303

Goldman, S. L., MacKay, M., Afshinnekoo, E., Melnick, A. M., Wu, S., & Mason, C. E. (2019). The impact of heterogeneity on single-cell sequencing. *Frontiers in Genetics*, *10*(MAR), 1–8. https://doi.org/10.3389/fgene.2019.00008

Goldmann, T., Zeller, N., Raasch, J., Kierdorf, K., Frenzel, K., Ketscher, L., Basters, A., Staszewski, O., Brendecke, S. M., Spiess, A., Tay, T. L., Kreutz, C., Timmer, J., Mancini, G. M., Blank, T., Fritz, G., Biber, K., Lang, R., Malo, D., … Prinz, M. (2015). USP18 lack in microglia causes destructive interferonopathy of the mouse brain. *The EMBO Journal*, *34*(12), 1612–1629. https://doi.org/10.15252/embj

Gonzalez-Angulo, A. M., Ferrer-Lozano, J., Stemke-Hale, K., Sahin, A., Liu, S., Barrera, J. A., Burgues, O., Lluch, A. M., Chen, H., Hortobagyi, G. N., Mills, G. B., & Meric-Bernstam, F. (2011). PI3K pathway mutations and PTEN levels in primary and metastatic breast cancer. *Molecular Cancer Therapeutics*. https://doi.org/10.1158/1535-7163.MCT-10-1089

Goryshin, I. Y., Miller, J. A., Kil, Y. V., Lanzov, V. A., & Reznikoff, W. S. (1998). Tn5/IS50 target recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(18), 10716–10721. https://doi.org/10.1073/pnas.95.18.10716

Granit, R. Z., Gabai, Y., Hadar, T., Karamansha, Y., Liberman, L., Waldhorn, I., Gat-Viks, I., Regev, A., Maly, B., Darash-Yahana, M., Peretz, T., & Ben-Porath, I. (2013). EZH2 promotes a bi-lineage identity in basal-like breast cancer cells. *Oncogene*.

https://doi.org/10.1038/onc.2012.390

Granit, Roy Z., Masury, H., Condiotti, R., Fixler, Y., Gabai, Y., Glikman, T., Dalin, S., Winter, E., Nevo, Y., Carmon, E., Sella, T., Sonnenblick, A., Peretz, T., Lehmann, U., Paz, K., Piccioni, F., Regev, A., Root, D. E., & Ben-Porath, I. (2018). Regulation of Cellular Heterogeneity and Rates of Symmetric and Asymmetric Divisions in Triple-Negative Breast Cancer. *Cell Reports*. https://doi.org/10.1016/j.celrep.2018.08.053

Granit, Roy Z., Slyper, M., & Ben-Porath, I. (2014). Axes of differentiation in breast cancer: Untangling stemness, lineage identity, and the epithelial to mesenchymal transition. In *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*. https://doi.org/10.1002/wsbm.1252

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyal, F., Frenoy, O., Pousse, Y., Reichen, M., Woolfe, A., Brenan, C., Griffiths, A. D., Vallot, C., & Gérard, A. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nature Genetics*. https://doi.org/10.1038/s41588-019-0424-9

Grzywa, T. M., Paskal, W., & Włodarski, P. K. (2017). Intratumor and Intertumor Heterogeneity in Melanoma. In *Translational Oncology*. https://doi.org/10.1016/j.tranon.2017.09.007

Gu Z. (2019). *rGREAT: Client for GREAT Analysis.* http://great.stanford.edu/public/html/.

Gudmundsdottir, K., & Ashworth, A. (2006). The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. In *Oncogene*. https://doi.org/10.1038/sj.onc.1209874

Guner, G., Guzelsoy, G., Isleyen, F. S., Sahin, G. S., Akkaya, C., Bayam, E., Kotan, E. I., Kabakcioglu, A., & Ince-Dunn, G. (2017). NEUROD2 Regulates Stim1 Expression and Store-Operated Calcium Entry in Cortical Neurons. *Eneuro*, *4*(February), ENEURO.0255-16.2017. https://doi.org/10.1523/ENEURO.0255-16.2017

Guo, C., Kong, W., Kamimoto, K., Rivera-Gonzalez, G. C., Yang, X., Kirita, Y., & Morris, S. A. (2019). CellTag Indexing: Genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biology*. https://doi.org/10.1186/s13059-019-1699-y

Gupta, P. B., Fillmore, C. M., Jiang, G., Shapira, S. D., Tao, K., Kuperwasser, C., & Lander, E. S. (2011a). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*. https://doi.org/10.1016/j.cell.2011.07.026

Gupta, P. B., Fillmore, C. M., Jiang, G., Shapira, S. D., Tao, K., Kuperwasser, C., & Lander, E. S. (2011b). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, *146*(4), 633–644.

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., Chin, S. F., Turashvili, G., Hirst, M., Caldas, C., Marra, M. A., Aparicio, S., & Shah, S. P. (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research*. https://doi.org/10.1101/gr.137570.112

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D. A., Rozenblatt-Rosen, O., Zhang, F., & Regev, A. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, *14*(10), 955–958. https://doi.org/10.1038/nmeth.4407

Hafner, M., Niepel, M., Chung, M., & Sorger, P. K. (2016). Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. In *Nature Methods* (Vol.

13, Issue 6, pp. 521–527).

Hai, T., & Curran, T. (1991). Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proceedings of the National Academy of Sciences, 88*(9), 3720–3724. https://doi.org/10.1073/pnas.88.9.3720

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., … Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. https://doi.org/10.1016/j.cell.2018.02.001

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Handy, D. E., Castro, R., & Loscalzo, J. (2011). Epigenetic modifications: Basic mechanisms and role in cardiovascular disease. *Circulation*. https://doi.org/10.1161/CIRCULATIONAHA.110.956839

Hardy, K. M., Booth, B. W., Hendrix, M. J. C., Salomon, D. S., & Strizzi, L. (2010). ErbB/EGF signaling and emt in mammary development and breast cancer. In *Journal of Mammary Gland Biology and Neoplasia*. https://doi.org/10.1007/s10911-010-9172-2

Heberle, H., Meirelles, V. G., da Silva, F. R., Telles, G. P., & Minghim, R. (2015). InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-015-0611-3

Heimberg, G., Bhatnagar, R., El-Samad, H., & Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems*. https://doi.org/10.1016/j.cels.2016.04.001

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010a). Simple Combinations of {Lineage-Determining} Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and {B} Cell Identities. In *Molecular Cell* (Vol. 38, Issue 4, pp. 576–589).

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010b). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Heiser, L. M., Sadanandam, A., Kuo, W.-L., Benz, S. C., Goldstein, T. C., Ng, S., Gibb, W. J., Wang, N. J., Ziyad, S., Tong, F., Bayani, N., Hu, Z., Billig, J. I., Dueregger, A., Lewis, S., Jakkula, L., Korkola, J. E., Durinck, S., Pepin, F., … Spellman, P. T. (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.*, *109*(8), 2724–2729.

Hergeth, S. P., & Schneider, R. (2015). The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO Reports*. https://doi.org/10.15252/embr.201540749

Herschkowitz, J. I., Zhao, W., Zhang, M., Usary, J., Murrow, G., Edwards, D., Knezevic, J., Greene, S. B., Darr, D., Troester, M. A., Hilsenbeck, S. G., Medina, D., Perou, C. M., & Rosen, J. M. (2012). Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1018862108

Hess, J. (2004). AP-1 subunits: quarrel and harmony among siblings. *Journal of Cell Science*, *117*(25), 5965–5973. https://doi.org/10.1242/jcs.01589

Hinohara, K., & Polyak, K. (2019a). Intratumoral Heterogeneity: More Than Just Mutations. In *Trends in Cell Biology*. https://doi.org/10.1016/j.tcb.2019.03.003

Hinohara, K., & Polyak, K. (2019b). Intratumoral Heterogeneity: More Than Just Mutations. *Trends in Cell Biology*, *29*(7), 569–579. https://doi.org/10.1016/j.tcb.2019.03.003

Hinohara, K., Wu, H. J., Vigneau, S., McDonald, T. O., Igarashi, K. J., Yamamoto, K. N., Madsen, T., Fassl, A., Egri, S. B., Papanastasiou, M., Ding, L., Peluffo, G., Cohen, O., Kales, S. C., Lal-Nag, M., Rai, G., Maloney, D. J., Jadhav, A., Simeonov, A., … Polyak, K. (2018). KDM5 Histone Demethylase Activity Links Cellular Transcriptomic Heterogeneity to Therapeutic Resistance. *Cancer Cell*. https://doi.org/10.1016/j.ccell.2018.10.014

Hoeflich, K. P., O'Brien, C., Boyd, Z., Cavet, G., Guerrero, S., Jung, K., Januario, T., Savage, H., Punnoose, E., Truong, T., Zhou, W., Berry, L., Murray, L., Amler, L., Belvin, M., Friedman, L. S., & Lackner, M. R. (2009). In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clinical Cancer Research*. https://doi.org/10.1158/1078-0432.CCR-09-0317

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., … Wang, J. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*. https://doi.org/10.1016/j.cell.2012.02.028

Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y. G., Zhu, J., Wang, P., & Xu, Y. (2013). Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. *Cell*. https://doi.org/10.1016/j.cell.2013.11.020

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusuda, J., Lane, D. P., … Wainwright, B. J. (2010). International network of cancer genome projects. In *Nature*. https://doi.org/10.1038/nature08987

Hughes, A. E. O., Magrini, V., Demeter, R., Miller, C. A., Fulton, R., Fulton, L. L., Eades, W. C., Elliott, K., Heath, S., Westervelt, P., Ding, L., Conrad, D. F., White, B. S., Shao, J., Link, D. C., DiPersio, J. F., Mardis, E. R., Wilson, R. K., Ley, T. J., … Graubert, T. A. (2014). Clonal Architecture of Secondary Acute Myeloid Leukemia Defined by Single-Cell Sequencing. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1004462

Hwang, B., Lee, D. S., Tamaki, W., Sun, Y., Ogorodnikov, A., Hartoularos, G., Winters, A., Song, Y. S., Chow, E. D., Spitzer, M. H., & Ye, C. J. (2020). SCITO-seq: single-cell combinatorial indexed cytometry sequencing. *BioRxiv*. https://doi.org/10.1101/2020.03.27.012633

Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. In *Genomics*. https://doi.org/10.1016/j.ygeno.2015.06.005

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., & Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*. https://doi.org/10.1126/science.1247651

Janes, K. A. (2016). Single-cell states versus single-cell atlases - two classes of heterogeneity that differ in meaning and method. In *Current Opinion in Biotechnology*. https://doi.org/10.1016/j.copbio.2016.03.015

Ji, Z., Zhou, W., & Ji, H. (2017). Single-cell regulome data analysis by SCRAT. *Bioinformatics*.

https://doi.org/10.1093/bioinformatics/btx315

Jin, S. G., Zhang, Z. M., Dunwell, T. L., Harter, M. R., Wu, X., Johnson, J., Li, Z., Liu, J., Szabó, P. E., Lu, Q., Xu, G. liang, Song, J., & Pfeifer, G. P. (2016). Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration. *Cell Reports*. https://doi.org/10.1016/j.celrep.2015.12.044

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*. https://doi.org/10.1126/science.1141319

Jolly, C., & Van Loo, P. (2018). Timing somatic events in the evolution of cancer. In *Genome Biology*. https://doi.org/10.1186/s13059-018-1476-3

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3230

Joo, J. Y., Schaukowitch, K., Farbiak, L., Kilaru, G., & Kim, T. K. (2015). Stimulus-specific combinatorial functionality of neuronal c-fos enhancers. *Nature Neuroscience*, *19*(1), 75–83. https://doi.org/10.1038/nn.4170

Kandel, E., & Spencer, W. (1961). Electrophysiology of hippocampal neurons. II. After-potentials and repetitive firing. *Journal of Neurophysiology*, *24*, 243–259. https://doi.org/10.1152/jn.1961.24.3.243

Kandel, E., Spencer, W., & Brinley, F. J. (1961). Electrophysiology of hippocampal neurons. I. Sequential invasion and synaptic organization. *Journal of Neurophysiology*, *24*, 225–242. https://doi.org/10.1152/jn.1961.24.3.225

Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., Specht, M. C., Bernstein, B. E., Michor, F., & Ellisen, L. W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature Communications*. https://doi.org/10.1038/s41467-018-06052-0

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*. https://doi.org/10.1038/s41467-019-09982-5

Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P., & Jones, P. A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*. https://doi.org/10.1101/gr.143008.112

Khare, S. P., Habib, F., Sharma, R., Gadewal, N., Gupta, S., & Galande, S. (2012). HIstome - A relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr1125

Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., & Navin, N. E. (2018a). Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell*, *173*(4), 879-893.e13. https://doi.org/10.1016/j.cell.2018.03.041

Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., & Navin, N. E. (2018b). Chemoresistance Evolution in {Triple-Negative} Breast Cancer Delineated by {Single-Cell} Sequencing. *Cell*, *173*(4), 879--893.e13.

Kim, S., Yu, N.-K., Shim, K.-W., Kim, J.-I., Kim, H., Han, D. H., Choi, J. E., Lee, S.-W., Choi, D. Il, Kim, M. W., Lee, D.-S., Lee, K., Galjart, N., Lee, Y.-S., Lee, J.-H., & Kaang, B.-K. (2018). Remote Memory and Cortical Synaptic Plasticity Require Neuronal CCCTC-

Binding Factor (CTCF). *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *38*(22), 5042–5052. https://doi.org/10.1523/JNEUROSCI.2738-17.2018

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., & Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. https://doi.org/10.1038/nature09033

Kimmerling, R. J., Prakadan, S. M., Gupta, A. J., Calistri, N. L., Stevens, M. M., Olcum, S., Cermak, N., Drake, R. S., Pelton, K., Smet, F. De, & Ligon, K. L. (2018). *Linking single-cell measurements of mass , growth rate , and gene expression*. 1–13.

Kimoto, H., Eto, R., Abe, M., Kato, H., & Araki, T. (2009). Alterations of glial cells in the mouse hippocampus during postnatal development. *Cellular and Molecular Neurobiology*, *29*(8), 1181–1189. https://doi.org/10.1007/s10571-009-9412-4

Kingsbury, M. a, Friedman, B., McConnell, M. J., Rehen, S. K., Yang, a H., Kaushal, D., & Chun, J. (2005). Aneuploid neurons are functionally active and integrated into brain circuitry. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(17), 6143–6147. https://doi.org/10.1073/pnas.0408171102

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. https://doi.org/10.1016/j.cell.2015.04.044

Klevebring, D., Rosin, G., Ma, R., Lindberg, J., Czene, K., Kere, J., Fredriksson, I., Bergh, J., & Hartman, J. (2014). Sequencing of breast cancer stem cell populations indicates a dynamic conversion between differentiation states in vivo. *Breast Cancer Research*. https://doi.org/10.1186/bcr3687

Knouse, K. A., Wu, J., & Amon, A. (2016). Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Research*. https://doi.org/10.1101/gr.198937.115

Knouse, K. A., Wu, J., Whittaker, C. A., & Amon, A. (2014). Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1415287111

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.68.4.820

Kobayashi, T., Iwaya, K., Moriya, T., Yamasaki, T., Tsuda, H., Yamamoto, J., & Matsubara, O. (2013). A simple immunohistochemical panel comprising 2 conventional markers, Ki67 and p53, is a powerful tool for predicting patient outcome in luminal-type breast cancer. *BMC Clinical Pathology*. https://doi.org/10.1186/1472-6890-13-5

Kong, W., Biddy, B. A., Kamimoto, K., Amrute, J. M., Butka, E. G., & Morris, S. A. (2020). CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. *Nature Protocols*. https://doi.org/10.1038/s41596-019-0247-2

Korbel, J. O., & Campbell, P. J. (2013). Criteria for inference of chromothripsis in cancer genomes. In *Cell*. https://doi.org/10.1016/j.cell.2013.02.023

Koren, S., & Bentires-Alj, M. (2015). Breast Tumor Heterogeneity: Source of Fitness, Hurdle for Therapy. In *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.10.031

Krepischi, A. C. V., Pearson, P. L., & Rosenberg, C. (2012). Germline copy number variations and cancer predisposition. *Future Oncology*. https://doi.org/10.2217/fon.12.34

Kulis, M., & Esteller, M. (2010). DNA Methylation and Cancer. In *Advances in Genetics*. https://doi.org/10.1016/B978-0-12-380866-0.60002-2

Kumagai, T., Akagi, T., Desmond, J. C., Kawamata, N., Gery, S., Imai, Y., Song, J. H., Gui, D., Said, J., & Koeffler, H. P. (2009). Epigenetic regulation and molecular characterization of C/EBPα in pancreatic cancer cells. *International Journal of Cancer*. https://doi.org/10.1002/ijc.23994

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., & Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*. https://doi.org/10.1038/35065132

Lacroix, M., Toillon, R. A., & Leclercq, G. (2006). p53 and breast cancer, an update. In *Endocrine-Related Cancer*. https://doi.org/10.1677/erc.1.01172

Ladewig, J., Koch, P., & Brüstle, O. (2013). Leveling Waddington: The emergence of direct programming and the loss of cell fate hierarchies. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3543

Lake, B., Cheng, S., Sos, B., Fan, J., Yung, Y., Kaeser, G., Duong, T., Yung, Y. C., Gao, D., Chun, J., Kharchenko, P., & Zhang, K. (2017). Integrative Single-Cell Analysis By Transcriptional And Epigenetic States In Human Adult Brain. *Nature Biotechnology*, 1–3. https://doi.org/doi:10.1038/nbt.4038

Laks, E., Zahn, H., Lai, D., McPherson, A., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T., Grewal, D., Nielsen, C., Leung, S., Bojilova, V., Smith, M., Golovko, O., Poon, S., Eirew, P., Kabeer, F., Algara, T., & Aparicio, S. (2018). *Resource: Scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires.* https://doi.org/10.1101/411058

Lamouille, S., Xu, J., & Derynck, R. (2014). Molecular mechanisms of epithelial-mesenchymal transition. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3758

Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., Pokholok, D., Aryee, M. J., Steemers, F. J., Lebofsky, R., & Buenrostro, J. D. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0147-6

Larsen, S., Kawamoto, S., Tanuma, S. I., & Uchiumi, F. (2015). The hematopoietic regulator, ELF-1, enhances the transcriptional response to Interferon-β of the OAS1 anti-viral gene. *Scientific Reports*, *5*. https://doi.org/10.1038/srep17497

Lasham, A., Mehta, S. Y., Fitzgerald, S. J., Woolley, A. G., Hearn, J. I., Hurley, D. G., Ruza, I., Algie, M., Shelling, A. N., Braithwaite, A. W., & Print, C. G. (2016). A novel {EGR-1} dependent mechanism for {YB-1} modulation of paclitaxel response in a triple negative breast cancer cell line. *Int. J. Cancer*, *139*(5), 1157–1170.

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R. J., Sanders, M. A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., Robinson, P., Coorens, T. H. H., O'Neill, L., Alder, C., Wang, J., Fitzgerald, R. C., Zilbauer, M., Coleman, N., Saeb-Parsy, K., … Stratton, M. R. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*. https://doi.org/10.1038/s41586-019-1672-7

Lee, E., Piranlioglu, R., Wicha, M. S., & Korkaya, H. (2019). Plasticity and potency of mammary stem cell subsets during mammary gland development. *International Journal of Molecular*

Sciences. https://doi.org/10.3390/ijms20092357

Lee, E. Y. H. P., & Muller, W. J. (2010). Oncogenes and tumor suppressor genes. In *Cold Spring Harbor perspectives in biology*. https://doi.org/10.1101/cshperspect.a003236

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011a). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI45014

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011b). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, *121*(7), 2750–2767.

Lein, E. S., Zhao, X., & Gage, F. H. (2004). Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *24*(15), 3879–3889. https://doi.org/10.1523/JNEUROSCI.4710-03.2004

Lesniak, D., Sabri, S., Xu, Y., Graham, K., Bhatnagar, P., Suresh, M., & Abdulkarim, B. (2013). Spontaneous epithelial-mesenchymal transition and resistance to {HER-2-targeted} therapies in {HER-2-positive} luminal breast cancer. *PLoS One*, *8*(8), e71987.

Leung, M. L., Wang, Y., Waters, J., & Navin, N. E. (2015). SNES: Single nucleus exome sequencing. *Genome Biology*. https://doi.org/10.1186/s13059-015-0616-2

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D., & Nolan, G. P. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, *162*(1), 184–197. https://doi.org/10.1016/j.cell.2015.05.047

Li, C., & Williams, S. M. (2013). Human Somatic Variation: It's Not Just for Cancer Anymore. *Current Genetic Medicine Reports*. https://doi.org/10.1007/s40142-013-0029-z

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, W., Ma, H., Zhang, J., Zhu, L., Wang, C., & Yang, Y. (2017). Unraveling the roles of CD44/CD24 and ALDH1 as cancer stem cell markers in tumorigenesis and metastasis. *Scientific Reports*. https://doi.org/10.1038/s41598-017-14364-2

Li, X., Guo, X., Li, D., Du, X., Yin, C., Chen, C., Fang, W., Bian, Z., Zhang, J., Li, B., Yang, H., & Xing, J. (2018). Multi-regional sequencing reveals intratumor heterogeneity and positive selection of somatic mtDNA mutations in hepatocellular carcinoma and colorectal cancer. *International Journal of Cancer*. https://doi.org/10.1002/ijc.31395

Li, Y., Xu, X., Song, L., Hou, Y., Li, Z., Tsang, S., Li, F., Im, K. M., Wu, K., Wu, H., Ye, X., Li, G., Wang, L., Zhang, B., Liang, J., Xie, W., Wu, R., Jiang, H., Liu, X., … Wang, J. (2012). Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaScience*. https://doi.org/10.1186/2047-217X-1-12

Liedtke, C., Mazouni, C., Hess, K. R., André, F., Tordai, A., Mejia, J. A., Symmans, W. F., Gonzalez-Angulo, A. M., Hennessy, B., Green, M., Cristofanilli, M., Hortobagyi, G. N., & Pusztai, L. (2008). Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2007.14.4147

Lien, W. H., Guo, X., Polak, L., Lawton, L. N., Young, R. A., Zheng, D., & Fuchs, E. (2011). Genome-wide maps of histone modifications unwind in vivo chromatin states of the hair follicle lineage. *Cell Stem Cell*. https://doi.org/10.1016/j.stem.2011.07.015

Liu, B., Conroy, J. M., Morrison, C. D., Odunsi, A. O., Qin, M., Wei, L., Trump, D. L., Johnson, C. S., Liu, S., & Wang, J. (2015). Structural variation discovery in the cancer genome using next generation sequencing: Computational solutions and perspectives. *Oncotarget*. https://doi.org/10.18632/oncotarget.3491

Liu, H., Wu, N., Zhang, Z., Zhong, X., Zhang, H., Guo, H., Nie, Y., & Liu, Y. (2019). Long Non-coding {RNA} {LINC00941} as a Potential Biomarker Promotes the Proliferation and Metastasis of Gastric Cancer. In *Frontiers in Genetics* (Vol. 10).

Lloyd-Lewis, B., Harris, O. B., Watson, C. J., & Davis, F. M. (2017). Mammary Stem Cells: Premise, Properties, and Perspectives. In *Trends in Cell Biology*. https://doi.org/10.1016/j.tcb.2017.04.001

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., Sherman, M. A., Vitzthum, C. M., Luquette, L. J., Yandava, C. N., Yang, P., Chittenden, T. W., Hatem, N. E., Ryu, S. C., Woodworth, M. B., Park, P. J., & Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*. https://doi.org/10.1126/science.aao4426

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D'Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J., & Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*. https://doi.org/10.1126/science.aab1785

Lorente de No, R. (1934). Studies on the structure of the cerebral cortex II. Continuation of the study of the Ammonic system. *J Psychol Neurol*, *46*, 113–177.

Love, M. I., Anders, S., & Huber, W. (2014). Differential analysis of count data - the DESeq2 package. In *Genome Biology* (Vol. 15, Issue 12). https://doi.org/110.1186/s13059-014-0550-8

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A., & Sankaran, V. G. (2019). Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell*. https://doi.org/10.1016/j.cell.2019.01.022

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*. https://doi.org/10.15252/msb.20188746

Lyko, F. (2018). The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2017.80

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2010.05.003

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt958

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*.

https://doi.org/10.1016/j.cell.2015.05.002

Mani, S. A., Guo, W., Liao, M. J., Eaton, E. N., Ayyanan, A., Zhou, A. Y., Brooks, M., Reinhard, F., Zhang, C. C., Shipitsin, M., Campbell, L. L., Polyak, K., Brisken, C., Yang, J., & Weinberg, R. A. (2008). The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell*. https://doi.org/10.1016/j.cell.2008.03.027

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., Thiruvahindrapduram, B., Fiebig, A., Schreiber, S., Friedman, J., Ketelaars, C. E. J., Vos, Y. J., Ficicioglu, C., Kirkpatrick, S., Nicolson, R., … Scherer, S. W. (2008). Structural Variation of Chromosomes in Autism Spectrum Disorder. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2007.12.009

Martelotto, L. G., Ng, C. K. Y., Piscuoglio, S., Weigelt, B., & Reis-Filho, J. S. (2014). Breast cancer intra-tumor heterogeneity. In *Breast Cancer Research*. https://doi.org/10.1186/bcr3658

Martin, C., & Zhang, Y. (2005). The diverse functions of histone lysine methylation. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm1761

Martincorena, Inigo. (2019). Somatic mutation and clonal expansions in human tissues. In *Genome Medicine*. https://doi.org/10.1186/s13073-019-0648-4

Martincorena, Iñigo, Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., & Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*. https://doi.org/10.1126/science.aaa6806

Martinez-Hernandez, A., Bell, K. P., & Norenberg, M. D. (1977). Glutamine synthetase: Glial localization in brain. *Science, 195*(4284), 1356–1358. https://doi.org/10.1126/science.14400

Mathis, R. A., Sokol, E. S., & Gupta, P. B. (2017). Cancer cells exhibit clonal diversity in phenotypic plasticity. *Open Biology*. https://doi.org/10.1098/rsob.160283

Mayer, C., Hafemeister, C., Bandler, R. C., Machold, R., Batista Brito, R., Jaglin, X., Allaway, K., Butler, A., Fishell, G., & Satija, R. (2018). Developmental diversification of cortical inhibitory interneurons. *Nature, 555*(7697), 457–462. https://doi.org/10.1038/nature25999

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., & Wills, Q. F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw777

McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R. S., Vermeesch, J. R., Hall, I. M., & Gage, F. H. (2013). Mosaic copy number variation in human neurons. *Science*. https://doi.org/10.1126/science.1243472

McGranahan, N., & Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. In *Cell*. https://doi.org/10.1016/j.cell.2017.01.018

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology, 28*(5), 495–501. https://doi.org/10.1038/nbt.1630

Meléndez-Rodríguez, F., Urrutia, A. A., Lorendeau, D., Rinaldi, G., Roche, O., Böğürcü-Seidel, N., Ortega Muelas, M., Mesa-Ciller, C., Turiel, G., Bouthelier, A., Hernansanz-Agustín, P., Elorza, A., Escasany, E., Li, Q. O. Y., Torres-Capelli, M., Tello, D., Fuertes, E., Fraga, E., Martínez-Ruiz, A., … Aragonés, J. (2019). HIF1α Suppresses Tumor Cell Proliferation

through Inhibition of Aspartate Biosynthesis. *Cell Rep.*, *26*(9), 2257-2265.e4. https://doi.org/10.1016/j.celrep.2019.01.106

Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., Farmer, A., Fordyce, P., Linnarsson, S., & Greenleaf, W. (2018). High-throughput chromatin accessibility profiling at single-cell resolution. *Nature Communications*. https://doi.org/10.1038/s41467-018-05887-x

Micalizzi, D. S., Farabaugh, S. M., & Ford, H. L. (2010). Epithelial-mesenchymal transition in cancer: Parallels between normal development and tumor progression. In *Journal of Mammary Gland Biology and Neoplasia*. https://doi.org/10.1007/s10911-010-9178-9

Mikkelsen, T. S., Xu, Z., Zhang, X., Wang, L., Gimble, J. M., Lander, E. S., & Rosen, E. D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. *Cell*. https://doi.org/10.1016/j.cell.2010.09.006

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*. https://doi.org/10.1038/ncomms15183

Minami, A., Nakanishi, A., Ogura, Y., Kitagishi, Y., & Matsuda, S. (2014). Connection between tumor suppressor BRCA1 and PTEN in damaged DNA repair. In *Frontiers in Oncology*. https://doi.org/10.3389/fonc.2014.00318

Mirzoeva, O. K., Das, D., Heiser, L. M., Bhattacharya, S., Siwak, D., Gendelman, R., Bayani, N., Wang, N. J., Neve, R. M., Guan, Y., Hu, Z., Knight, Z., Feiler, H. S., Gascard, P., Parvin, B., Spellman, P. T., Shokat, K. M., Wyrobek, A. J., Bissell, M. J., … Korn, W. M. (2009). Basal subtype and MAPK/ERK kinase (MEK)-phosphoinositide 3-kinase feedback signaling determine susceptibility of breast cancer cells to MEK inhibition. *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-08-3389

Mitelman, F., Johansson, B., & Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. In *Nature Reviews Cancer*. https://doi.org/10.1038/nrc2091

Mo, A., Mukamel, E. A., Davis, F. P., Luo, C., Henry, G. L., Picard, S., Urich, M. A., Nery, J. R., Sejnowski, T. J., Lister, R., Eddy, S. R., Ecker, J. R., & Nathans, J. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*, *86*(6), 1369–1384. https://doi.org/10.1016/j.neuron.2015.05.018

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., & Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0392-8

Morris, G. J., Naidu, S., Topham, A. K., Guiles, F., Xu, Y., McCue, P., Schwartz, G. F., Park, P. K., Rosenberg, A. L., Brill, K., & Mitchell, E. P. (2007). Differences in breast carcinoma characteristics in newly diagnosed African-American and Caucasian patients: A single-institution compilation compared with the national cancer institute's surveillance, epidemiology, and end results database. *Cancer*. https://doi.org/10.1002/cncr.22836

Muller, P. A. J., & Vousden, K. H. (2013). P53 mutations in cancer. In *Nature Cell Biology*. https://doi.org/10.1038/ncb2641

Mulqueen, R. M., DeRosa, B. A., Thornton, C. A., Sayar, Z., Torkenczy, K. A., Fields, A. J., Wright, K. M., Nan, X., Ramji, R., Steemers, F. J., O'Roak, B. J., & Adey, A. C. (2019). Improved single-cell {ATAC-seq} reveals chromatin dynamics of in vitro corticogenesis. In *bioRxiv*.

Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., Sinnamon, J. R., Shendure, J., Trapnell, C., O'Roak, B. J., Xia, Z., Steemers, F. J., & Adey, A. C. (2018). Highly scalable generation of DNA methylation profiles in single cells. *Nature Biotechnology, February*. https://doi.org/10.1038/nbt.4112

Navin, N. E. (2015). The first five years of single-cell cancer genomics and beyond. In *Genome Research*. https://doi.org/10.1101/gr.191098.115

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., & Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*. https://doi.org/10.1038/nature09807

Neeb, A., Wallbaum, S., Novac, N., Dukovic-Schulze, S., Scholl, I., Schreiber, C., Schlag, P., Moll, J., Stein, U., & Sleeman, J. P. (2012). The immediate early gene Ier2 promotes tumor cell motility and metastasis, and predicts poor survival of colorectal cancer patients. *Oncogene*, *31*(33), 3796–3806.

Ni, X., Zhuo, M., Su, Z., Duan, J., Gao, Y., Wang, Z., Zong, C., Bai, H., Chapman, A. R., Zhao, J., Xu, L., An, T., Ma, Q., Wang, Y., Wu, M., Sun, Y., Wang, S., Li, Z., Yang, X., … Wang, J. (2013). Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1320659110

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., … Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*. https://doi.org/10.1016/j.cell.2012.04.024

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., … Campbell, P. J. (2012). The life history of 21 breast cancers. *Cell*. https://doi.org/10.1016/j.cell.2012.04.023

Nilsen, G., Liestøl, K., Loo, P. Van, Moen Vollan, H. K., Eide, M. B., Rueda, O. M., Chin, S. F., Russell, R., Baumbusch, L. O., Caldas, C., Børresen-Dale, A. L., & Lingjærde, O. C. (2012). Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics*. https://doi.org/10.1186/1471-2164-13-591

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*. https://doi.org/10.1126/science.959840

O'Connor, M. J. (2015). Targeting the DNA Damage Response in Cancer. In *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.10.040

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175.

Okano, M., Xie, S., & Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases [1]. In *Nature Genetics*. https://doi.org/10.1038/890

Okano, Masaki, Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*. https://doi.org/10.1016/S0092-8674(00)81656-6

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004a). Circular binary

segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxh008

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004b). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, *5*(4), 557–572. https://doi.org/10.1093/biostatistics/kxh008

Osborne, C., Wilson, P., & Tripathy, D. (2004). Oncogenes and Tumor Suppressor Genes in Breast Cancer: Potential Diagnostic and Therapeutic Applications. *The Oncologist*. https://doi.org/10.1634/theoncologist.9-4-361

Pal, B., Chen, Y., Vaillant, F., Jamieson, P., Gordon, L., Rios, A. C., Wilcox, S., Fu, N., Liu, K. H., Jackling, F. C., Davis, M. J., Lindeman, G. J., Smyth, G. K., & Visvader, J. E. (2017). Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nature Communications*. https://doi.org/10.1038/s41467-017-01560-x

Park, H. S., Jang, M. H., Kim, E. J., Kim, H. J., Lee, H. J., Kim, Y. J., Kim, J. H., Kang, E., Kim, S. W., Kim, I. A., & Park, S. Y. (2014). High EGFR gene copy number predicts poor outcome in triple-negative breast cancer. *Modern Pathology*. https://doi.org/10.1038/modpathol.2013.251

Pataskar, A., Jung, J., Smialowski, P., Noack, F., Calegari, F., Straub, T., & Tiwari, V. K. (2016). NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. *The EMBO Journal*, *35*(1), 24–45. https://doi.org/10.15252/embj.201591206

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. https://doi.org/10.1080/14786440109462720

Pellacani, D., Bilenky, M., Kannan, N., Heravi-Moussavi, A., Knapp, D. J. H. F., Gakkhar, S., Moksa, M., Carles, A., Moore, R., Mungall, A. J., Marra, M. A., Jones, S. J. M., Aparicio, S., Hirst, M., & Eaves, C. J. (2016). Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Reports*. https://doi.org/10.1016/j.celrep.2016.10.058

Perkins, N. D. (2007). Integrating cell-signalling pathways with NF-κB and IKK function. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2083

Pervolarakis, N., Sun, P., Gutierrez, G., Nguyen, Q. H., Jhutty, D., XY Zheng, G., Nemec, C. M., Dai, X., Watanabe, K., & Kessenbrock, K. (2019). Integrated Single-Cell Transcriptomics and Chromatin Accessibility Analysis Reveals Novel Regulators of Mammary Epithelial Cell Identity. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3443688

Peterson, S. E., Yang, A. H., Bushman, D. M., Westra, J. W., Yung, Y. C., Barral, S., Mutoh, T., Rehen, S. K., & Chun, J. (2012). Aneuploid cells are differentially susceptible to caspase-mediated death during embryonic cerebral cortical development. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.3706-12.2012

Phillips, S., Prat, A., Sedic, M., Proia, T., Wronski, A., Mazumdar, S., Skibinski, A., Shirley, S. H., Perou, C. M., Gill, G., Gupta, P. B., & Kuperwasser, C. (2014). Cell-state transitions regulated by SLUG are critical for tissue regeneration and tumor initiation. *Stem Cell Reports*. https://doi.org/10.1016/j.stemcr.2014.03.008

Picelli, S., Björklund, A. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, *24*(12), 2033–2040. https://doi.org/10.1101/gr.177881.114

Pleasure, S. J., Collins, A. E., & Lowenstein, D. H. (2000). Unique expression patterns of cell fate molecules delineate sequential stages of dentate gyrus development. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *20*(16), 6095–6105. https://doi.org/20/16/6095 [pii]

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., & Trapnell, C. (2018a). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2018.06.044

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., & Trapnell, C. (2018b). Cicero Predicts cis-Regulatory {DNA} Interactions from {Single-Cell} Chromatin Accessibility Data. *Mol. Cell*, *71*(5), 858--871.e8.

Polyak, K., & Weinberg, R. A. (2009). Transitions between epithelial and mesenchymal states: Acquisition of malignant and stem cell traits. In *Nature Reviews Cancer*. https://doi.org/10.1038/nrc2620

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *ELife*. https://doi.org/10.7554/eLife.23203

Potts, S. J., Krueger, J. S., Landis, N. D., Eberhard, D. A., David Young, G., Schmechel, S. C., & Lange, H. (2012). Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. *Laboratory Investigation*. https://doi.org/10.1038/labinvest.2012.91

Poynter, S. T., & Kadoch, C. (2016). Polycomb and trithorax opposition in development and disease. In *Wiley Interdisciplinary Reviews: Developmental Biology*. https://doi.org/10.1002/wdev.244

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., & Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research*. https://doi.org/10.1186/bcr2635

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., Díez, M., Viladot, M., Arance, A., & Muñoz, M. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast*. https://doi.org/10.1016/j.breast.2015.07.008

Pratilas, C. A., Taylor, B. S., Ye, Q., Viale, A., Sander, C., Solit, D. B., & Rosen, N. (2009). V600EBRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0900780106

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D. U., Zhang, Y., Sos, B. C., Afzal, V., Dickel, D. E., Kuan, S., Visel, A., Pennacchio, L. A., Zhang, K., & Ren, B. (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience*, 1–8. https://doi.org/10.1038/s41593-018-0079-3

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell developmental trajectories. *BioRxiv*. https://doi.org/10.1101/110668

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A., & Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*. https://doi.org/10.1038/nbt.4103

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., Noble, W. S., Duan, Z., & Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nature Methods*, *14*(3), 263–266. https://doi.org/10.1038/nmeth.4155

Ramani, V., Deng, X., Qiu, R., Lee, C., Disteche, C. M., Noble, W. S., Shendure, J., & Duan, Z. (2020). Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods*. https://doi.org/10.1016/j.ymeth.2019.09.012

Ramaswamy, B., Mrozek, E., Lustberg, M., Wesolowski, R., Layman, R., Abdel-Rasoul, M., Timmers, C., Patrick, R., Sexton, J., Macrae, E., Shapiro, C., Budd, T., Harris, L., Isaacs, C., Ismail-Khan, R., Dees, C., Poklepovic, A., Grever, M., Chen, H., … Carson, W. (2016). *Abstract LB-216: NCI 9455: Phase II study of trametinib followed by trametinib plus AKT inhibitor,GSK2141795 in patients with advanced triple negative breast cancer*. https://doi.org/10.1158/1538-7445.am2016-lb-216

Ramon, Y., & Cajal, S. (1911). *Histologie du Syste`me Nerveux de L'Homme et des Verte´bre´s, vol II*.

Rasmussen, K. D., & Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. In *Genes and Development*. https://doi.org/10.1101/gad.276568.115

Rehen, S. K., McConnell, M. J., Kaushal, D., Kingsbury, M. a, Yang, a H., & Chun, J. (2001). Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13361–13366. https://doi.org/10.1073/pnas.231487398

Reuben, A., Spencer, C. N., Prieto, P. A., Gopalakrishnan, V., Reddy, S. M., Miller, J. P., Mao, X., De Macedo, M. P., Chen, J., Song, X., Jiang, H., Chen, P. L., Beird, H. C., Garber, H. R., Roh, W., Wani, K., Chen, E., Haymaker, C., Forget, M. A., … Wargo, J. A. (2017). Genomic and immune heterogeneity are associated with differential responses to therapy in melanoma. *Npj Genomic Medicine*. https://doi.org/10.1038/s41525-017-0013-8

Rinehart, J., Adjei, A. A., LoRusso, P. M., Waterhouse, D., Hecht, J. R., Natale, R. B., Hamid, O., Varterasian, M., Asbury, P., Kaldjian, E. P., Gulyas, S., Mitchell, D. Y., Herrera, R., Sebolt-Leopold, J. S., & Meyer, M. B. (2004). Multicenter phase II study of the oral MEK inhibitor, CI-1040, in patients with advanced non-small-cell lung, breast, colon, and pancreatic cancer. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2004.01.185

Risom, T. (2017). *Measuring And Managing Phenotypic Heterogeneity And Plasticity In Breast Cancer To Improve Therapeutic Control*. *July*.

Risom, T., Langer, E. M., Chapman, M. P., Rantala, J., Fields, A. J., Boniface, C., Alvarez, M. J., Kendsersky, N. D., Pelz, C. R., Johnson-Camacho, K., Dobrolecki, L. E., Chin, K., Aswani, A. J., Wang, N. J., Califano, A., Lewis, M. T., Tomlin, C. J., Spellman, P. T., Adey, A., … Sears, R. C. (2018). Differentiation-state plasticity is a targetable resistance mechanism in basal-like breast cancer. *Nat. Commun.*, *9*(1), 3815.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*. https://doi.org/10.1038/nature14248

Rodgers, K., & Mcvey, M. (2016). Error-Prone Repair of DNA Double-Strand Breaks. In *Journal of Cellular Physiology*. https://doi.org/10.1002/jcp.25053

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., & Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. https://doi.org/10.1126/science.aaw1219

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., & Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. https://doi.org/10.1126/science.aam8999

Rosenkrantz, J. L., & Carbone, L. (2017). Investigating somatic aneuploidy in the brain: why we need a new model. In *Chromosoma*. https://doi.org/10.1007/s00412-016-0615-4

Roybon, L., Hjalt, T., Stott, S., Guillemot, F., Li, J. Y., & Brundin, P. (2009). Neurogenin2 directs granule neuroblast production and amplification while neuroD1 specifies neuronal fate during hippocampal neurogenesis. *PLoS ONE*, *4*(3). https://doi.org/10.1371/journal.pone.0004779

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J. E., Ashenberg, O., Cerami, E., Coffey, R. J., Demir, E., Ding, L., Esplin, E. D., Ford, J. M., Goecks, J., Ghosh, S., Gray, J. W., Guinney, J., Hanlon, S. E., Hughes, S. K., … Zhuang, X. (2020). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. In *Cell*. https://doi.org/10.1016/j.cell.2020.03.053

Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A., & Teichmann, S. A. (2017). The Human Cell Atlas: From vision to reality. In *Nature*. https://doi.org/10.1038/550451a

Rubin, C. M. (1998). The Genetic Basis of Human Cancer. *Annals of Internal Medicine*. https://doi.org/10.7326/0003-4819-129-9-199811010-00045

Ryland, G. L., Doyle, M. A., Goode, D., Boyle, S. E., Choong, D. Y. H., Rowley, S. M., Li, J., Bowtell, D. D., Tothill, R. W., Campbell, I. G., & Gorringe, K. L. (2015). Loss of heterozygosity: What is it good for? *BMC Medical Genomics*, *8*(1), 1–12. https://doi.org/10.1186/s12920-015-0123-z

Saini, K. S., Loi, S., de Azambuja, E., Metzger-Filho, O., Saini, M. L., Ignatiadis, M., Dancey, J. E., & Piccart-Gebhart, M. J. (2013). Targeting the PI3K/AKT/mTOR and Raf/MEK/ERK pathways in the treatment of breast cancer. In *Cancer Treatment Reviews*. https://doi.org/10.1016/j.ctrv.2013.03.009

Sams, D. S., Nardone, S., Getselter, D., Raz, D., Tal, M., Rayi, P. R., Kaphzan, H., Hakim, O., & Elliott, E. (2016). Neuronal CTCF Is Necessary for Basal and Experience-Dependent Gene Regulation, Memory Formation, and Genomic Structure of BDNF and Arc. *Cell Reports*, *17*(9), 2418–2430. https://doi.org/10.1016/j.celrep.2016.11.004

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., Chakravarty, D., Daian, F., Gao, Q., Bailey, M. H., Liang, W. W., Foltz, S. M., Shmulevich, I., Ding, L., Heins, Z., … Schultz, N. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*. https://doi.org/10.1016/j.cell.2018.03.035

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*. https://doi.org/10.1038/nbt.3192

Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., Olsen, B. N.,

Mumbach, M. R., Pierce, S. E., Corces, M. R., Shah, P., Bell, J. C., Jhutty, D., Nemec, C. M., Wang, J., Wang, L., Yin, Y., Giresi, P. G., Chang, A. L. S., … Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral {T} cell exhaustion. *Nat. Biotechnol.*, *37*(8), 925–936.

Saunders, A., Macosko, E. Z., Wysoker, A., Goldman, M., Krienen, F. M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., Goeva, A., Nemesh, J., Kamitaki, N., Brumbaugh, S., Kulp, D., & McCarroll, S. A. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell*, *174*(4), 1015-1030.e16. https://doi.org/10.1016/J.CELL.2018.07.028

Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, *14*(10), 975–978. https://doi.org/10.1038/nmeth.4401

Schier, A. F. (2020). Single-cell biology: beyond the sum of its parts. *Nature Methods*, *17*(1), 17–20. https://doi.org/10.1038/s41592-019-0693-3

Schübeler, D. (2015). Function and information content of DNA methylation. In *Nature*. https://doi.org/10.1038/nature14192

Score, J., Hidalgo-Curtis, C., Jones, A. V., Winkelmann, N., Skinner, A., Ward, D., Zoi, K., Ernst, T., Stegelmann, F., Döhner, K., Chase, A., & Cross, N. C. P. (2012). Inactivation of polycomb repressive complex 2 components in myeloproliferative and myelodysplastic/myeloproliferative neoplasms. *Blood*. https://doi.org/10.1182/blood-2011-07-367243

Scoville, W., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21.

Sen, G. L., Reuter, J. A., Webster, D. E., Zhu, L., & Khavari, P. A. (2010). DNMT1 maintains progenitor function in self-renewing somatic tissue. *Nature*. https://doi.org/10.1038/nature08683

Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1002533

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., Bashashati, A., Prentice, L. M., Khattra, J., Burleigh, A., Yap, D., Bernard, V., McPherson, A., Shumansky, K., Crisan, A., … Aparicio, S. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*. https://doi.org/10.1038/nature10933

Shah, S. P., Xuan, X., DeLeeuw, R. J., Khojasteh, M., Lam, W. L., Ng, R., & Murphy, K. P. (2006). Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl238

Shajahan-Haq, A. N., Boca, S. M., Jin, L., Bhuvaneshwar, K., Gusev, Y., Cheema, A. K., Demas, D. D., Raghavan, K. S., Michalek, R., Madhavan, S., & Clarke, R. (2017). {EGR1} regulates cellular metabolism and survival in endocrine resistant breast cancer. *Oncotarget*, *8*(57), 96865–96884.

Sharma, S., Javadekar, S. M., Pandey, M., Srivastava, M., Kumari, R., & Raghavan, S. C. (2015). Homology and enzymatic requirements of microhomology-dependent alternative end joining. *Cell Death & Disease*. https://doi.org/10.1038/cddis.2015.58

Sharma, Sreenath V, Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M. A., Wong, K.-K., Brandstetter, K.,

Wittner, B., Ramaswamy, S., Classon, M., & Settleman, J. (2010). A {Chromatin-Mediated} Reversible {Drug-Tolerant} State in Cancer Cell Subpopulations. In *Cell* (Vol. 141, Issue 1, pp. 69–80).

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., Levin, J. Z., Nemesh, J., Goldman, M., McCarroll, S. A., Cepko, C. L., Regev, A., & Sanes, J. R. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*. https://doi.org/10.1016/j.cell.2016.07.054

Shlien, A., & Malkin, D. (2009). Copy number variations and cancer. In *Genome Medicine*. https://doi.org/10.1186/gm62

Sinn, H. P., & Kreipe, H. (2013). A brief overview of the WHO classification of breast tumors, 4th edition, focusing on issues and updates from the 3rd edition. In *Breast Care*. https://doi.org/10.1159/000350774

Sinnamon, J. R., Torkenczy, K. A., Linhoff, M. W., Vitak, S. A., Mulqueen, R. M., Pliner, H. A., Trapnell, C., Steemers, F. J., Mandel, G., & Adey, A. C. (2019a). The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Research*, *29*(5), 857–869. https://doi.org/10.1101/gr.243725.118

Sinnamon, J. R., Torkenczy, K. A., Linhoff, M. W., Vitak, S. A., Mulqueen, R. M., Pliner, H. A., Trapnell, C., Steemers, F. J., Mandel, G., & Adey, A. C. (2019b). The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Research*, *29*(5), 857–869. https://doi.org/10.1101/gr.243725.118

Sinnamon, J. R., Torkenczy, K. A., Linhoff, M. W., Vitak, S. A., Mulqueen, R. M., Pliner, H. A., Trapnell, C., Steemers, F. J., Mandel, G., & Adey, A. C. (2019c). The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Res.*, *29*(5), 857–869.

Smith, M. L., & Milner, B. (1981). The role of the right hippocampus in the recall of spatial location. *Neuropsychologia*, *19*(6), 781–793.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A. L., & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0932692100

Sos, B. C., Fung, H. L., Gao, D. R., Osothprarop, T. F., Kia, A., He, M. M., & Zhang, K. (2016). Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biology*. https://doi.org/10.1186/s13059-016-0882-7

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., & Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CrIsPr-Cas9-induced genetic scars. *Nature Biotechnology*. https://doi.org/10.1038/nbt.4124

Spencer, W. A., & Kandel, E. R. (1961a). Electrophysiology of hippocampal neurons: III. Firing level and time constant. *Journal of Neurophysiology*, *24*(3), 260–271. https://doi.org/10.1152/jn.1961.24.3.260

Spencer, W. A., & Kandel, E. R. (1961b). Electrophysiology of hippocampal neurons: IV. Fast prepotentials. *Journal of Neurophysiology*, *24*(3), 272–285. https://doi.org/10.1152/jn.1961.24.3.272

Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., Zhang, F., Steemers, F., Shendure, J., &

Trapnell, C. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, *367*(6473), 45–51. https://doi.org/10.1126/science.aax6234

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., … Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. In *Science*. https://doi.org/10.1126/science.aaf2403

Stahley, S. N., & Kowalczyk, A. P. (2015). Desmosomes in acquired disease. In *Cell and Tissue Research*. https://doi.org/10.1007/s00441-015-2155-2

Stone, J. L., O'Donovan, M. C., Gurling, H., Kirov, G. K., Blackwood, D. H. R., Corvin, A., Craddock, N. J., Gill, M., Hultman, C. M., Lichtenstein, P., McQuillin, A., Pato, C. N., Ruderfer, D. M., Owen, M. J., St Clair, D., Sullivan, P. F., Sklar, P., Purcell, S. M., Korn, J., … Ardlie, K. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*. https://doi.org/10.1038/nature07239

Stracquadanio, G., Wang, X., Wallace, M. D., Grawenda, A. M., Zhang, P., Hewitt, J., Zeron-Medina, J., Castro-Giner, F., Tomlinson, I. P., Goding, C. R., Cygan, K. J., Fairbrother, W. G., F. Thomas, L., Sætrom, P., Gemignani, F., Landi, S., Schuster-Böckler, B., Bell, D. A., & Bond, G. L. (2016). The importance of p53 pathway genetics in inherited and somatic cancer genomes. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc.2016.15

Streets, A. M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., & Huang, Y. (2014). Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1402030111

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck 3rd, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of {Single-Cell} Data. *Cell*, *177*(7), 1888--1902.e21.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Stoeckius, M., Smibert, P., & Satija, R. (2018). Comprehensive integration of single cell data. *BioRxiv*, 460147. https://doi.org/10.1101/460147

Su, Y., Shin, J., Zhong, C., Wang, S., Roychowdhury, P., Lim, J., Kim, D., Ming, G. L., & Song, H. (2017). Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nature Neuroscience*, *20*(3), 476–483. https://doi.org/10.1038/nn.4494

Svichar, N., Esquenazi, S., Chen, H.-Y., & Chesler, M. (2011). Preemptive Regulation of Intracellular pH in Hippocampal Neurons by a Dual Mechanism of Depolarization-Induced Alkalinization. *Journal of Neuroscience*, *31*(19), 6997–7004. https://doi.org/10.1523/JNEUROSCI.6088-10.2011

Takahashi, K., & Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm.2016.8

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., Levi, B., Gray, L. T., Sorensen, S. A., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S. M., … Zeng, H. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*. https://doi.org/10.1038/nn.4216

Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjöld, M., Ponder, B. A. J., & Tunnacliffe, A.

(1992). Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics*. https://doi.org/10.1016/0888-7543(92)90147-K

Tessarz, P., & Kouzarides, T. (2014). Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3890

Thornton, C. A., Mulqueen, R. M., Torkenczy, K. A., Lowenstein, E. G., Fields, A. J., Steemers, F. J., Wright, K. M., & Adey, A. C. (2019). Spatially-mapped single-cell chromatin accessibility. *BioRxiv*, 815720. https://doi.org/10.1101/815720

Tischfield, J. A. (1997). Loss of heterozygosity or: How I learned to stop worrying and love mitotic recombination. *American Journal of Human Genetics*. https://doi.org/10.1086/301617

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. In *Genome Research* (Vol. 25, Issue 10, pp. 1491–1498). https://doi.org/10.1101/gr.190595.115

Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., Miura, I., Wakana, S., Nishino, J., & Yagi, T. (2015). Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Research*. https://doi.org/10.1101/gr.186148.114

Valencia, A. M., & Kadoch, C. (2019). Chromatin regulatory mechanisms and therapeutic opportunities in cancer. In *Nature Cell Biology*. https://doi.org/10.1038/s41556-018-0258-1

Van Der Maarel, S. M., Deidda, G., Lemmers, R. J. L. F., Van Overveld, P. G. M., Van Der Wielen, M., Hewitt, J. E., Sandkuijl, L., Bakker, B., Van Ommen, C. J. B., Padberg, G. W., & Frants, R. R. (2000). De novo facioscapulohumeral muscular dystrophy: Frequent somatic mosaicism, sex-dependent phenotype, and the role of mitotic transchromosomal repeat interaction between chromosomes 4 and 10. *American Journal of Human Genetics*. https://doi.org/10.1086/302730

Van Gent, D. C., & Kanaar, R. (2016). Exploiting DNA repair defects for novel cancer therapies. In *Molecular Biology of the Cell*. https://doi.org/10.1091/mbc.E15-10-0698

Venkatesh, S., & Workman, J. L. (2015). Histone exchange, chromatin structure and the regulation of transcription. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3941

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J. F., Gould, J., Griffin, G. K., Borg, Å., Ronaghi, M., Frisén, J., Lundeberg, J., Regev, A., & Ståhl, P. L. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*. https://doi.org/10.1038/s41592-019-0548-y

Visvader, J. E., & Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: Current status and perspectives. In *Genes and Development*. https://doi.org/10.1101/gad.242511.114

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., Carbone, L., Steemers, F. J., & Adey, A. (2017a). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, *14*(3), 302–308. https://doi.org/10.1038/nmeth.4154

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., Carbone, L., Steemers, F. J., & Adey, A. (2017b). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, *14*(3), 302–308. https://doi.org/10.1038/nmeth.4154

Waddell, N., Pajic, M., Patch, A. M., Chang, D. K., Kassahn, K. S., Bailey, P., Johns, A. L., Miller, D., Nones, K., Quek, K., Quinn, M. C. J., Robertson, A. J., Fadlullah, M. Z. H., Bruxner, T. J. C., Christ, A. N., Harliwong, I., Idrisoglu, S., Manning, S., Nourse, C., … Grimmond, S. M. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*. https://doi.org/10.1038/nature14169

Waddington, C. H. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *The Strategy of the Genes A Discussion of Some …*. https://doi.org/10.1007/3-540-32786-X_7

Wahl, G. M., & Spike, B. T. (2017). Cell state plasticity, stem cells, EMT, and the generation of intra-tumoral heterogeneity. In *npj Breast Cancer*. https://doi.org/10.1038/s41523-017-0012-z

Wallace, M. L., Saunders, A., Huang, K. W., Philson, A. C., Goldman, M., Macosko, E. Z., McCarroll, S. A., & Sabatini, B. L. (2017). Genetically Distinct Parallel Pathways in the Entopeduncular Nucleus for Limbic and Sensorimotor Output of the Basal Ganglia. *Neuron*. https://doi.org/10.1016/j.neuron.2017.03.017

Wang, Q. M., Lv, L. I., Tang, Y., Zhang, L. I., & Wang, L. F. (2019). MMP-1 is overexpressed in triple-negative breast cancer tissues and the knockdown of MMP-1 expression inhibits tumor cell malignant behaviors in vitro. *Oncology Letters*. https://doi.org/10.3892/ol.2018.9779

Wang, X., Chen, H., & Zhang, N. R. (2018). DNA copy number profiling using single-cell sequencing. *Briefings in Bioinformatics*, *19*(5), 731–736. https://doi.org/10.1093/bib/bbx004

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F., & Navin, N. E. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*. https://doi.org/10.1038/nature13600

Wattenberg, M., Viégas, F., & Johnson, I. (2017). How to Use t-SNE Effectively. *Distill*. https://doi.org/10.23915/distill.00002

Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M., & Klein, A. M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1714723115

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., … Kling, T. (2013). The cancer genome atlas pan-cancer analysis project. In *Nature Genetics*. https://doi.org/10.1038/ng.2764

Welch, J. D., Hartemink, A. J., & Prins, J. F. (2017). MATCHER: Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology*. https://doi.org/10.1186/s13059-017-1269-0

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., & Macosko, E. Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, *177*(7), 1873-1887.e17. https://doi.org/10.1016/j.cell.2019.05.006

West, A. C., & Johnstone, R. W. (2014). New and emerging HDAC inhibitors for cancer treatment. In *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI69738

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson, D. G., Bhatt, S., Chavrier, P., Bravo, R., & Charnay, P. (1989). Segment-specific expression of a zinc-finger gene in the developing nervous system of the mouse. *Nature*, *337*(6206), 461–464. https://doi.org/10.1038/337461a0

Williams, M. J., Werner, B., Heide, T., Curtis, C., Barnes, C. P., Sottoriva, A., & Graham, T. A. (2018). Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*. https://doi.org/10.1038/s41588-018-0128-6

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*. https://doi.org/10.1186/s13059-017-1382-0

Wu, Q., Zhang, W., Xue, L., Wang, Y., Fu, M., Ma, L., Song, Y., & Zhan, Q. M. (2019). APC/C-CDH1-regulated IDH3β coordinates with the cell cycle to promote cell proliferation. *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-18-2341

Wu, Y. E., Pan, L., Zuo, Y., Li, X., & Hong, W. (2017). Detecting Activated Cell Populations Using Single-Cell RNA-Seq. *Neuron*. https://doi.org/10.1016/j.neuron.2017.09.026

Wu, Y., Sarkissyan, M., & Vadgama, J. (2016). Epithelial-Mesenchymal Transition and Breast Cancer. *Journal of Clinical Medicine*. https://doi.org/10.3390/jcm5020013

Wuidart, A., Sifrim, A., Fioramonti, M., Matsumura, S., Brisebarre, A., Brown, D., Centonze, A., Dannau, A., Dubois, C., Van Keymeulen, A., Voet, T., & Blanpain, C. (2018). Early lineage segregation of multipotent embryonic mammary gland progenitors. *Nature Cell Biology*. https://doi.org/10.1038/s41556-018-0095-2

Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T., & Zhang, Q. C. (2019). SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nature Communications*. https://doi.org/10.1038/s41467-019-12630-7

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., … Wang, J. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*. https://doi.org/10.1016/j.cell.2012.02.025

Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A. J., Zheng, L., Zhang, H., Huang, S., Min, J., Nicholson, T., Chen, T., Xu, G., Shi, Y., Zhang, K., & Shi, Y. G. (2011). Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2011.04.005

Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. In *Genome biology*. https://doi.org/10.1186/s13059-020-1929-3

Yang, Liubin, Rau, R., & Goodell, M. A. (2015). DNMT3A in haematological malignancies. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc3895

Yang, Lixing, Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C. H., Zhang, C., Ren, X., Protopopov, A., Chin, L., Kucherlapati, R., Lee, C., & Park, P. J. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*. https://doi.org/10.1016/j.cell.2013.04.010

Yang, M., Teng, W., Qu, Y., Wang, H., & Yuan, Q. (2016). Sulforaphene inhibits triple negative breast cancer through activating tumor suppressor Egr1. *Breast Cancer Res. Treat.*, *158*(2), 277–286.

Yang, P., Du, C. W., Kwan, M., Liang, S. X., & Zhang, G. J. (2013). The impact of p53 in predicting clinical outcome of breast cancer patients with visceral metastasis. *Scientific*

Reports. https://doi.org/10.1038/srep02246

Yang, X. J. (2004). Lysine acetylation and the bromodomain: A new partnership for signaling. In *BioEssays*. https://doi.org/10.1002/bies.20104

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., … Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine*. https://doi.org/10.1038/nm.3886

Yersal, O., & Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. In *World Journal of Clinical Oncology*. https://doi.org/10.5306/wjco.v5.i3.412

Yin, Y., Jiang, Y., Berletch, J. B., Disteche, C. M., Noble, W. S., Steemers, F. J., Adey, A. C., & Shendure, J. A. (2018). High-throughput mapping of meiotic crossover and chromosome mis-segregation events in interspecific hybrid mice. *BioRxiv*, 338053. https://doi.org/10.1101/338053

Yin, Y., Jiang, Y., Lam, K. W. G., Berletch, J. B., Disteche, C. M., Noble, W. S., Steemers, F. J., Camerini-Otero, R. D., Adey, A. C., & Shendure, J. (2019). High-Throughput Single-Cell Sequencing with Linear Amplification. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2019.08.002

Younes, M., Wu, Z., Dupouy, S., Lupo, A. M., Mourra, N., Takahashi, T., Fléjou, J. F., Trédaniel, J., Régnard, J. F., Damotte, D., Alifano, M., & Forgez, P. (2014). Neurotensin ({NTS}) and its receptor ({NTSR1}) causes {EGFR}, {HER2} and {HER3} over-expression and their autocrine/paracrine activation in lung tumors, confirming responsiveness to erlotinib. *Oncotarget*, *5*(18), 8252–8269.

Yu, C., Yu, J., Yao, X., Wu, W. K. K., Lu, Y., Tang, S., Li, X., Bao, L., Li, X., Hou, Y., Wu, R., Jian, M., Chen, R., Zhang, F., Xu, L., Fan, F., He, J., Liang, Q., Wang, H., … Wang, J. (2014). Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Research*. https://doi.org/10.1038/cr.2014.43

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv145

Zawistowski, J. S., Bevill, S. M., Goulet, D. R., Stuhlmiller, T. J., Beltran, A. S., Olivares-Quintero, J. F., Singh, D., Sciaky, N., Parker, J. S., Rashid, N. U., Chen, X., Duncan, J. S., Whittle, M. C., Angus, S. P., Velarde, S. H., Golitz, B. T., He, X., Santos, C., Darr, D. B., … Johnson, G. L. (2017). Enhancer remodeling during adaptive bypass to MEK inhibition is attenuated by pharmacologic targeting of the P-TEFb complex. *Cancer Discovery*. https://doi.org/10.1158/2159-8290.CD-16-0653

Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., & Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, *347*(6226), 1138–1142. https://doi.org/10.1126/science.aaa1934

Zeisel, Amit, Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., & Linnarsson, S. (2018). Molecular Architecture of the Mouse Nervous System. *Cell*, *174*(4), 999-1014.e22. https://doi.org/10.1016/J.CELL.2018.06.021

Zeiser, R. (2014). Trametinib. *Recent Results Cancer Res.*, *201*, 241–248.

Zhang, K. (2017). Stratifying tissue heterogeneity with scalable single-cell assays. In *Nature Methods*. https://doi.org/10.1038/nmeth.4209

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., Phatnani, H. P., Guarnieri, P., Caneda, C., Ruderisch, N., Deng, S., Liddelow, S. A., Zhang, C., Daneman, R., Maniatis, T., Barres, B. A., & Wu, J. Q. (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *Journal of Neuroscience*, *34*(36), 11929–11947. https://doi.org/10.1523/JNEUROSCI.1860-14.2014

Zhang, Yong, Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhu, C. (2020). *Single-cell multimodal omics: the power of many*. *17*(January).

Zola-Morgan, S., Squire, L. R., & Amaral, D. G. (1986). Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *6*(10), 2950–2967.

Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*. https://doi.org/10.1126/science.1229164

Appendix

Additional published papers

Peer reviewed:

Mulqueen RM, Pokholok D, Norberg SJ., **Torkenczy K. A**, Fields AJ, Sun D, Sinnamon JR, Shendure J, Trapnell C, O'Roak BJ, Xia Z, Steemers FJ, Adey AC. (2018). Highly scalable generation of DNA methylation profiles in single cells. *Nature biotechnology*, **36**(5):428-431.

Su Y, Pelz, C, Huang, T, **Torkenczy K. A**, Wang X, Cherry A, Daniel CJ, Liang J, Nan X, Dai MS, Adey AC, Impey S, Sears RC. Post-translational modification localizes MYC to the nuclear pore basket to regulate a subset of target genes involved in cellular responses to environmental signals. *Genes & development*, **32**:21-22 (2018), 1398-1419.

Daughtry BL, Rosenkrantz JL, Lazar NH, Fei SS, Redmayne N, **Torkenczy K. A**, Adey A, Yan M, Gao L, Park B, Nevonen KA, Carbone L and Chavez SL. Single-cell sequencing of primate preimplantation embryos reveals chromosome elimination via cellular fragmentation and blastomere exclusion. Genome Res. **25** (2019), doi:10.1101/gr.239830.118.

Preprint:

Mulqueen, R. M., DeRosa, B. A., Thornton, C. A., Sayar, Z., **Torkenczy, K. A.**, Fields, A. J., Wright, K. M., Nan, X., Ramji, R., Steemers, F. J., O'Roak, B. J., & Adey, A. C. (2019). Improved single-cell {ATAC-seq} reveals chromatin dynamics of in vitro corticogenesis. In *bioRxiv*.

Thornton, C. A., Mulqueen, R. M., **Torkenczy, K. A.**, Lowenstein, E. G., Fields, A. J., Steemers, F. J., Wright, K. M., & Adey, A. C. (2019). Spatially-mapped single-cell chromatin accessibility. BioRxiv, 815720. https://doi.org/10.1101/815720