Cancer treatment and relapse : single-cell chromatin profiling of rare persister cells using droplet-based microfluidics

Kevin Grosselin

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à ESPCI Paris

Cancer treatment and relapse: single-cell chromatin profiling of rare persister cells using droplet-based microfluidics

École doctorale n°388

CHIMIE PHYSIQUE ET CHIMIE ANALYTIQUE DE PARIS CENTRE

Spécialité CHIMIE PHYSIQUE

COMPOSITION DU JURY :

Mr Christian Muchardt
Institut Pasteur, Président

Mme Virginie Pénard-Lacronique
Institut Gustave Roussy, Rapporteur

Mme Anne-Marie Gué
LAAS, Rapporteur

Mme Angela Taddei
Institut Curie, Examinateur

Mr Joshua Waterfall
Institut Curie, Examinateur

Mr Andrew Griffiths
ESPCI Paris, Directeur de thèse

Mme Annabelle Gérard
HiFiBiO Therapeutics, Membre invité

Soutenue par **Kevin GROSSELIN**
le 26 Septembre 2018

Dirigée par **Prof. Andrew Griffiths**

CONFIDENTIEL
date de fin de confidentialité : 30/09/2020

PSL RESEARCH UNIVERSITY PARIS

Remerciements

Mes remerciements s'adressent en premier lieu à Annabelle Gérard et Andrew Griffiths. Annabelle, tu n'as pas choisi la facilité il y a 4 ans en faisant confiance à un petit chimiste qui n'avait jamais touché une pipette ! Tu as su me transmettre les connaissances et la rigueur nécessaire pour la réussite de ce travail de thèse. Si j'en suis à rédiger ce manuscrit aujourd'hui c'est en grande partie grâce à toi, merci pour tout. Andrew, merci de m'avoir accueilli au sein du Laboratoire de Biochimie. J'ai bénéficié d'un environnement de travail remarquable qui m'a permis de m'épanouir scientifiquement et humainement. Je te remercie pour ta disponibilité et ta confiance aux cours de ces années de thèse.

Je remercie Fred Dom, directeur d'HiFiBiO France, d'avoir financé mes recherches au cours de ces 4 dernières années. Tu m'as toujours considéré comme un membre à part entière d'HiFiBiO bien que tu n'aies pas toujours été tendre avec moi (2 défaites sévères en autant de matchs de squash !). J'ai eu la chance de pouvoir bénéficier de toutes les ressources nécessaires pour mener à bien mon projet scientifique.

Je remercie Mme Virginie Pénard-Lacronique, Mme Anne-Marie Gué, Mme Angela Taddei, Mr Christian Muchardt & Mr Joshua Waterfall d'avoir accepté de faire partie du jury de soutenance de ma thèse.

Je remercie chaleureusement Céline Vallot de l'Institut Curie pour notre intense collaboration au cours de cette dernière année. A ce titre, je remercie également Virginie Pénard-Lacronique de l'Institut Gustave Roussy, Jan Zylicz de l'Institut Curie et les plateformes de séquençage de l'Institut Curie et de l'ICM qui ont accepté mes demandes les plus farfelues et toujours traité mes échantillons dans les plus brefs délais.

Ce travail de thèse n'était pas un cavalier seul et de nombreuses personnes ont contribué à son avancement. Marcel, merci pour ton aide précieuse lors du développement microfluidique. Tu as toujours refusé l'idée du Kevinator mais avec

du recul je crois que c'était un mal pour un bien ! Adam, merci pour les analyses des données de séquençage et pour m'avoir encouragé à apprendre la programmation. Baptiste, merci d'avoir partagé toute ta rigueur expérimentale, mais je l'avoue je ne découpe pas encore les dents de crocodile des petits bouts de Scotch !

Merci, Yannick, pour m'avoir fourni des billes de grande qualité ; Adeline, pour ton aide précieuse lors de notre fameuse "Hell Week" ; Vera, même si tu ne m'as jamais dit que filtrer les cellules à la décongélation n'était pas forcément utile ; les membres anonymes de la team Mickey, avec qui les discussions étaient toujours productives (en deux mots : *Yellow Submarine*) ; le B230, dont j'ai pris beaucoup de plaisir à diriger ! Merci Raph, pour m'avoir transmis ton savoir microfluidique ; Antoine, même si tu n'as toujours pas compris qu'une bonne ligne de commande vaut mieux qu'une interface graphique ; Sophie, bien que tu aies brisé notre binôme ménage SCC ; Steph, "the umr manager!" partie trop tôt en télétravail.

Merci ma chère Marina, mes chers Pablo, Stéphane et Marco. Vous m'avez toujours donné d'excellents conseils et étiez toujours présents pour répondre à mes questions, même les moins pertinentes ! Merci Isa, Hélène et Céline, vous êtes indispensables et votre aide a grandement facilité mon quotidien.

Je sais pertinemment que ce travail de thèse n'aurait pas été de même qualité sans vous les amis. Je pense notamment à tous les Sages et à toi, Le Van. Que ce soit en Tunisie, en Ariège, en Bretagne, en Corse, en Ardèche ou lors de l'Appartathon, des TBT et du Marathon du Médoc, ces moments passés ensemble resteront des souvenirs impérissables de cette période de thèse.

Merci Maman, Papa, Groli, Poup et Bikou pour votre soutien tout au long de la thèse. Vous avez toujours été et vous resterez une formidable source d'inspiration.

Enfin, je profite de ces dernières lignes pour remercier la personne dont la contribution à ce travail a été la plus grande. Tu es la première personne à subir mes états d'âme mais tu as eu la patience et le courage d'y faire face. Je ne pense pas me tromper en écrivant ici que tout ceux qui ont travaillé avec moi au cours de ces dernières années se joignent également à ces remerciements. S'ils ont eu la chance de me connaitre souriant et agréable au quotidien, c'est grâce à toi ! Elodie, merci.

Contents

List of Figures

List of Tables

Cancer treatment and relapse: single-cell chromatin profiling of rare persister cells using droplet-based microfluidics

Abstract

Epigenetic mechanisms including DNA methylation and histone modifications modulate chromatin structure and cell-type specific functions. Epigenomic profiling have revealed that chromatin landscapes are widely altered in cancer cells. Genome-wide maps of histone modifications are conventionally obtained by Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq). However, this method only yields an average snapshot of the modification status and doesn't provide insight about intratumoral epigenetic heterogeneity. Rare subpopulations including drug-tolerant persister cells remain undetectable.

Droplet-based microfluidics allow to use micro-metric monodisperse droplets as reaction vessels to perform high-throughput single-cell assays. In this thesis, I describe a single-cell ChIP-seq system combining droplet microfluidics with DNA barcoding technology that enables histone modifications mapping at single-cell resolution, from thousands of cells. Chromatin from individual cells is fragmented and barcoded in droplets prior immunoprecipitation and sequencing library preparation. Sequencing reads deconvoluted by their barcode sequence attribute each sequence to their originating cells allowing reconstruction of single-cell chromatin profiles with an unprecedent coverage of up to 10^4 unique loci per cell.

Applied to profile chromatin marks associated with active transcription and repressed gene expression (H3K4me3 & H3K27me3) in mixed population of human B and T lymphocytes, scChIP-seq showed that >99% of the cells were correctly identified, defining distinct chromatin states of immune cells with high accuracy. In patient-derived xenograft (PDX) models of breast cancer with acquired drug resistance, the method identified rare populations of cells in the untreated, drug-sensitive tumors with a chromatin landscape similar to resistant cells after treatment. These results highlight the potential selection of cells with chromatin marks in response and resistance to cancer therapy.

Keywords: single-cell epigenomics, droplet-based microfluidics, drug resistance.

Introduction

Chapter 1

From chromatin modifications to cell-type identity

A fundamental research question in biology is to understand how hundreds of distinct cell types arise from identical genetic material in multicellular organisms. The many different cell types can't be explained solely by genetics but rather by an additional information that can bridge phenotype to genotype. In 1942, Conrad H. Waddington coined the term EPIGENETICS as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" [Waddington 42].

Since then, studies have shown that differential patterns of gene expression underlie the diversity of phenotypes, with the two being closely related to the organizational structure of DNA in the cell. In many instances, patterns of gene expression and associated phenotypes are preserved through cellular division. Hence, epigenetics may be nowadays defined as the study of stable and ideally heritable changes in gene expression (or phenotype) without changes in the underlying DNA sequence.

Eukariotic genomes are organized within nucleus into chromatin, a complex nucleoprotein structure discovered as early as 1882 [Flemming 82]. The functions of the chromatin extend well beyond simple DNA compaction. An additional layer of "epigenetic information" is stored in the form of chemical modifications impacting both DNA and histone proteins that constitute the chromatin. Epigenetic mechanisms through chromatin modifications regulate gene expression and shape specific chromatin landscapes, which allow predictions to be made about cell type and tissue identity [Bernstein 05, Barski 07].

1.1 Packaging DNA

1.1.1 The nucleosome: subunit of chromatin

> *"A eukaryotic chromosome made out of self-assembling 70 Å units, which could perhaps be made to crystallize, would necessitate rewriting our textbooks on cytology and genetics! I have never read such a naive paper purporting to be of such fundamental significance. Definitely it should not be published anywhere!"* [a]

[a] Anecdote reported by Donald E. Olins and Ada L. Olins in their review "Chromatin history: our view from the bridge" [Olins 03]

In these words, a reviewer from *Nature* rejected this new view of chromatin structure hypothesized as early as 1973 by Christopher L. Woodcock. Yet the partial digestion of nuclear DNA by micrococcal nuclease already revealed DNA fragments of a size multiple of 200 base pairs corresponding to the nucleosomal DNA [Hewish 73], but their first observation as particles was made by Donald and Ada Olins in 1974 using electron microscopy (see Fig. 1.1.1a) [Olins 74]. They showed that nucleosomes form the fundamental and repeated units of chromatin as "beads on a string".

Nucleosome structure

The same year, Kornberg and Thomas isolated and characterized the nucleosomes as being composed of an octamer of histone proteins around which DNA is wrapped (also called "nucleosome core particle", see Fig. 1.1.1b) [Kornberg 74]. Adjacent nucleosomes are separated by, on average, ~50 base pairs of linker DNA, whose length varies among species and cell types [Kornberg 77].

The structure of the nucleosome core particle comprises 147 base pairs of DNA coiled in a left-handed 1.65 turns around an octamer of histone proteins. This octamer is composed of equimolar amounts of four core histones (H3, H4, H2A and H2B) structured as follows: a central tetramer of H3:H4 heterodimers flanked by two H2A:H2B heterodimers [Luger 97].

Histone proteins

The core histones (H3, H4, H2A and H2B) are basic and relatively small proteins (11-15 kDa) but highly conserved in eukariotes. Electron microscopy and high-resolution cristal structures have shown a disk-shape arrangement of the core histones in the nucleosome core particle, except for their N- and C-terminal ends,

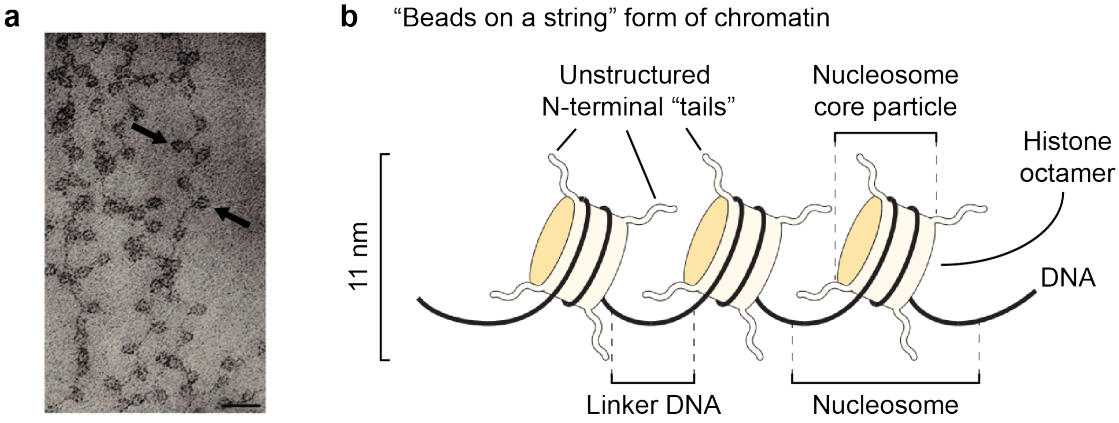


Figure 1.1.1: **"Beads on a string" structure of chromatin.**
(a) Electron micrograph revealing the "beads on a string" structure of chromatin. Nucleosomes are indicated by arrows. Scale bar is 30 nm. Reprinted from [Olins 74]. (b) Schematic representation of the elements forming the 11 nm fiber. Adapted from [Schones 08]

protruding from the surface (also referred to as "tails"). The histone tails account for 25-30% of the mass of the core histones and are largely unstructured [Luger 97]. In particular, N-terminal tails composed of ~30 amino acid residues are subject to post-translational modifications which play important roles in chromatin organization as well as in the regulation of many biological processes (histone post-translational modifications are introduced in sub-section 1.2.2 and their functions in section 1.3).

Linker histones complete the nucleosome structure. They bind to the exterior of the nucleosome core particles at the entry/exit sites of DNA, but their exact position is still not well defined [Hergeth 15]. This family of histone proteins is much less conserved than the core histones, they exist in multiple subtypes and they are also subject to post-translational modifications. Linker histones are known to be involved in enhancing structural stability of nucleosomes and, to a larger extent, facilitate the folding of chromatin into higher-order structures [Fan 03]. Interestingly, linker histones are also suspected to be involved in gene regulation by preventing / recruiting transcriptional activators or repressors [Kim 13].

1.1.2 Higher order of chromatin organization

Genomic DNA in eukariotic cells is packaged into chromatin and forms higher order structure to compact DNA within nucleus (see Fig. 1.1.2).

Nucleosomes are the first packaging elements of DNA. As previously men-

tionned, DNA is wound around nucleosomes to form the chromatin primary structure in the appearance of "beads on a string" (also referred as the 11 nm fiber) [Olins 74, van Holde 89]. This structure shortens DNA sevenfold compared to naked DNA.

The chromatin secondary structure is formed by interactions between nucleosomes and is stabilized by linker histones. Under physiological conditions, chromatin compacts into its higher folded structure and forms a superhelical fiber of 30 nm in diameter [Hansen 89]. This fiber is further folded into higher order of chromatin organization until the final mitotic chromosome structure.

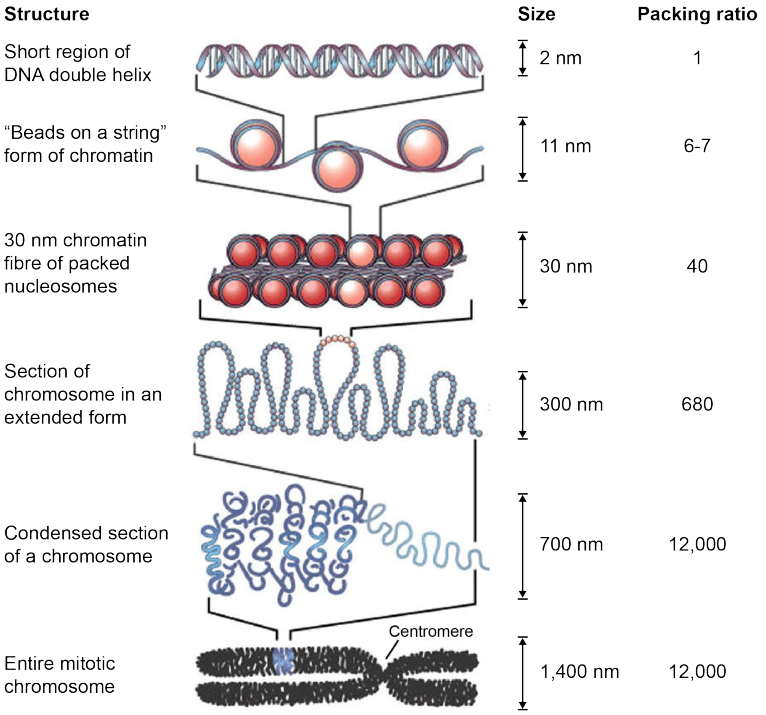


Figure 1.1.2: **Packaging DNA**.
Size and packing ratio obtained from [Pienta 84]. Reprinted from [Felsenfeld 03]

1.2 Chromatin modifications and their machineries

Chromatin remodeling, DNA methylation and histones post-translational modifications are the main contributors to the epigenetic mechanisms. In the following introductory chapter, the discussion is limited to covalent chromatin modifications, with a main focus on histone modifications.

1.2.1 DNA methylation

Chemical modifications of DNA bases were identified in 1948 by chromatography [Hotchkiss 48]. Cytosine methylation (5-methylcytosine) is one of the most widely studied epigenetic modification in human and is sometimes referred as the "fifth DNA base". Methylation of cytosine occurs principally in repetitive genomic regions and within CpG dinucleotides (Cytosine followed by Guanine separated by a phosphate group). The latter tend to group together to form regions with high density of CpG dinucleotides called CpG islands [Bird 85]. In human, CpG are rare (~1%) but about 60% of gene promoters are composed of CpG islands [Bernstein 07].

DNA methylation is mediated by DNA methyltransferases family (DNMTs). It has been established that DNMT1 is responsible for maintaining DNA methylation pattern through cellular division [Li 92]. By contrast, DNMT3a and DNMT3b are known to establish *de novo* methylation sites, both enzymes being highly expressed in embryonic stem cells as opposed to differentiated cells.

Methylation of cytosine was long assumed as an irreversible and stable epigenetic mark. Recent work identified a new family of enzymes referred as TET (Ten-Eleven Translocation), which are able to modify 5-methylcytosine into reaction intermediates leading to the removal of the methyl group from the cytosine [Tahiliani 09, Ito 11].

One functional consequence of DNA methylation is the regulation of gene expression [Razin 80, Bird 02]. In general, methylated CpG islands are correlated with gene repression. Two complementary modes of action have been described: (i) the methyl group prevents the binding of transcription factors [Watt 88] and (ii) the methyl group promotes the recruitment of repressor complexes through the binding of highly conserved proteins from the MBD family (Methyl-CpG Binding Domain) [Bird 99, Sarraf 04]. DNA methylation plays an important role in maintaining genome stability, regulating the transcription as well as in the inactivation of X-chromosome in female mammals [Reik 05, Hellman 07] and in the parent-of-origin allelic imprinting [Li 93].

1.2.2 Histone post-translational modifications

A pioneering work by Allfrey in 1964 led to the identification of the first histone post-translational modification [Allfrey 64]. Since then, more than hundreds different histone post-translational modifications grouped in at least eight classes have been reported [Kouzarides 07]. All histones are subject to post-translational modifications occuring primarily within the N-terminal tails of the core histones.

Acetylation, methylation and phosphorylation of histone tail residues are among the most widely studied histone modifications and have important roles in DNA-based biological processes (see Table 1.1 and section 1.3). A nomenclature is used to describe the large variety of histone modifications in order to unambiguously identify the histone, the modified amino acid (symbol and position) and the level of modification [Turner 05]. For example, the histone H3 trimethylated on its lysine 4 is noted H3K4me3. Here, the discussion is limited to acetylation and methylation, which have been studied in this thesis.

Modifications	**Residues modified**	**Functions regulated**
Acetylation	**K**-ac	Transcription, Repair, Replication, Condensation
Methylation	**K**-me1, **K**-me2, **K**-me3	Transcription, Repair
Methylation	**R**-me1, **R**-me2a, **R**-me2s	Transcription
Phosphorylation	**S**-ph, **T**-ph	Transcription, Repair, Condensation
Ubiquitylation	**K**-ub	Transcription, Repair
Sumoylation	**K**-su	Transcription
ADP ribosylation	**E**-ar	Transcription
Deimination	**R** > **Cit**	Transcription
Proline isomerization	**P-cis** > **P-trans**	Transcription

Table 1.1: **Classes of histone modifications.**
Reprinted from [Kouzarides 07]

Histone-modifying enzymes

Histone post-translational modifications are mediated by enzymes that are well characterized in the literature. Additions of modification are catalyzed by histone-modifying enzymes also referred to as *writers*. Chromatin marks being dynamic and reversible, *erasers* enzymes have been identified to catalyze the removal of histone modifications.

Writers

Acetylation has long been considered as a hallmark of transcriptionally active chromatin but the direct link was only established with the identification of the first Histone AcetylTransferases (HATs) in 1996 [Brownell 96b]. The acetylation

involves the transfer of an acetyl group from the cofactor Acetyl-CoA to the lysine residues (K) of histone amino tails [Allfrey 64]. Since then, numerous HATs enzymes have been identified and they are now classified with respect to their cellular localization and substrate specificity. Type-A HATs are localized in the nucleus and catalyze the acetylation of nucleosomal histones in the context of transcription regulation. Conversely, Type-B HATs are localized in the cytoplasm and catalize the acetylation of newly synthesized histones but don't acetylate nucleosomal substrates already deposited into chromatin [Brownell 96a].

Histone methylation mainly occurs on lysine (K) and arginine (R) residues of histone amino-tails. Lysine residues may be mono-, di- or trimethylated whereas arginine residues may be mono- and dimethylated in a symmetric or asymmetric configuration [Bedford 09, Ng 09]. The first histone lysine methyltransferase (KMT) identified in 2002 targets H3K9 [Rea 00] but numerous KMTs have been identified so far. Almost all of them contain the so-called SET domain, a highly evolutionary conserved domain of ~120 amino acid residues, catalyzing the transfer of a methyl group from the cofactor S-adenosylmethionine (SAM) to the lysine residues (K) of histone amino tails [Tschiersch 94, Bannister 11]. Interestingly, the lysine position and the level of methylation have different impact on the regulation of the gene expression: methylation of H3K4 and H3K36 is generally associated with gene activation, whereas methylation of H3K9, H3K27 and H4K20 are common sites for gene repression [Bernstein 07, Barski 07]. Methylation of arginine residues is catalyzed by enzymes belonging to the Protein Arginine MethylTransferases family (PRMTs). Like the methylation of lysine, the arginine position and the level of methylation have different impact on the regulation of the gene expression: H4R3me2a, H3R2me2s, H3R17me2a and H3R26me2a are associated with gene activation, whereas H3R2me2a, H3R8me2a, H3R8me2s and H4R3me2s are associated with gene repression [Bedford 09, Blanc 17].

Erasers

Soon after the identification of HATs enzymes, Histone DeACetylases (HDACs) were reported in the literature. As expected, their activity was related to transcriptional repression [Taunton 96]. There are four classes of HDACs based on their function and homology with previously identified deacetylase complexes in yeast [Holbert 05].

For many years, erasers of methylated histones were not known and this modification was considered stable and static. A first class of lysine demethylase (Lysine-Specific Demethylase LSD) was identified but their mechanism of demethylation was only compatible with mono- and dimethylated lysines [Shi 04]. Soon after,

a second important class of demethylases was discovered, having certain enzymes capable of demethylating trimethylated lysines. The latter possess all a highly conserved catalytic Jumonji domain [Tsukada 06].

1.3 Functional consequences of histone modifications

Two major modes of action have been described to explain the functional consequences of histone post-translational modifications: (i) histone modifications influence the overall structure of the chromatin and organize chromatin environments, and (ii) histone modifications promote and stabilize the binding of chromatin factors (also referred to as *"readers"*) and orchestrate DNA-based biological processes.

1.3.1 Establishing global chromatin environments

From a chromatin structure point of view, eukariotic genomes can be roughly divided into two conformation states: euchromatin and heterochromatin. At the cytological level, euchromatin is only condensed during mitosis, whereas heterochromatin remains condensed throughout the cell cycle. Both regions have also been characterized at the molecular level: euchromatin is relatively relaxed and comprises most of the active portion of the genome. Conversely, because of its condensed organization, the heterochromatin is less accessible to the transcription machinery and therefore considered inactive.

Heterochromatin

Heterochromatin can be in turn divided into two distinct environments referred to as constitutive heterochromatin and facultative heterochromatin.

The first one defines genomic regions such as telomeres or centromeres and contains permanently silenced genes. For example, constitutive heterochromatin has been characterized by high level of trimethylated H3K9 and Heterochromatin Protein 1 (HP1) [Li 07].

On the other hand, facultative heterochromatin is constituted of genes that are expressed during the development and cellular differentiation and which then become silenced. The best example illustrating facultative heterochromatin is the inactivation of the X-chromosome in female mammals. The inactivated X-chromosome has been characterized by high level of H3K27me3 and Polycomb-group proteins (PcG). Polycomb repressive complex 2 (PRC2) can methylate H3K27 through its catalytic subunit EZH2 domain. Interestingly, H3K27me3 also

mediate PRC2 recruitment during DNA replication, thus maintaining facultative heterochromatin and contributing to the inheritance of chromatin modifications [Hansen 08].

Euchromatin

Euchromatin represents the majority of the genome regions (e.g. ~92% of the human genome). Euchromatin is much more relaxed than heterochromatin so DNA has flexibility in the biological output (e.g. active or repressive gene expression). Transcriptionally active euchromatin has been characterized with high levels of acetylation and methylation at H3K4, H3K36 and H3K79. Conversely, inactive euchromatin has been characterized with low level of acetylation and methylation. However, it is important to note that one histone modification is not necessary specific to one chromatin state. For example, the entire body of an active gene is highly enriched in trimethylated H3K36, whereas H3K36me3 enrichment at the promoter is a common feature among repressed genes. A similar observation was made for methylation at H3K9 [Vakoc 05].

Histone modifications promote local and global structural perturbations

Interactions between adjacent nucleosomes or between histones and DNA may be altered by histone modifications.

Acetylation of the lysine residues reduces the positive charge of the histones, which "unfolds" chromatin and results in a less compact structure accessible for binding of transcription factors. Particularly, numerous lysine residues have the potential to be acetylated (e.g. H3K9, H3K14, H3K18, H4K5, H4K8, H4K12...), suggesting important effects on chromatin structure. Enhancers and promoters of active genes are hyper-acetylated, indicating again that acetylation facilitates accessibility [Wang 08]. This is difficult to observe *in-vivo* but, for example, it has been shown *in-vitro* that acetylation of H4K16 prevents the formation of the 30 nm fiber and higher order of chromatin organization [Shogren-Knaak 06]. Conversely, hypoacetylation leads to a more compact chromatin structure, thus reducing the DNA accessibility for transcription factors [Strahl 00].

Histone phosphorylation may also affect chromatin compaction by charge changes. There are fewer potential sites for phosphorylation compared to acetylation, but for example it has been established that genome-wide phosphorylation of H3S10 promotes chromatin condensation during mitosis [Wei 98].

1.3.2 Regulation of DNA-based processes

Histone modifications are involved in various DNA-based processes by serving as recognition sites for effector proteins capable of reading information and by stabilizing their binding to the chromatin. Numerous evolutionary highly conserved proteins (*"readers"*) have been identified and characterized to specifically interact with modified histones.

Bromodomain proteins recognized specifically acetylated lysines and such motif is mainly found in HATs and chromatin remodeling complexes [Hassan 02].

On the other hand, methylated lysines are bound by chromodomain proteins which can be associated with either active and repressive chromatin states. For example, ATP-dependent remodeling proteins from the Chromo Helicase DNA binding family (CHD) have been found to recruit transcription activating complexes through its binding to methylated H3K4 [Pray-Grant 05]. Conversely, chromodomains are also linked to inactive gene expression. For example, transcription repression and condensed chromatin structure are associated with high levels of H3K27me3 and di-, trimethylated H3K9. The latter histone modifications are bound by Polycomb Group proteins (PcG) and Heterochromatin Protein 1 (HP1) respectively, which mediate the maintenance of the overall structure of heterochromatin [Lachner 01, Cao 02].

1.3.3 Histone modifications interactions

As discussed in section 1.2.2, the large abundance of histone modifications enables a tight control of chromatin structure and a great flexibility in the regulation of DNA-based processes. However, this diversity leads to crosstalk between histones that can be modified at different sites simultaneously. Histone modifications can positively or negatively affect each other (Fig. 1.3.1). In addition, communication between histone modifications also exists with other chromatin modifications such as DNA methylation, which all participate to fine-tune the overall regulation of the biological functions (Fig. 1.3.2) [Du 15].

Histone modifications crosstalk

Histones can be modified at different sites simultaneously. Communications between histone modifications may occur at different level: among different histones, among different tails of the same histone, among the same histone tail or even among the same site [Wang 08]. A set of positive and negative relations between modifications among the same histone is depicted in Fig. 1.3.1.

Histone modifications can interact in various ways. The binding of a protein

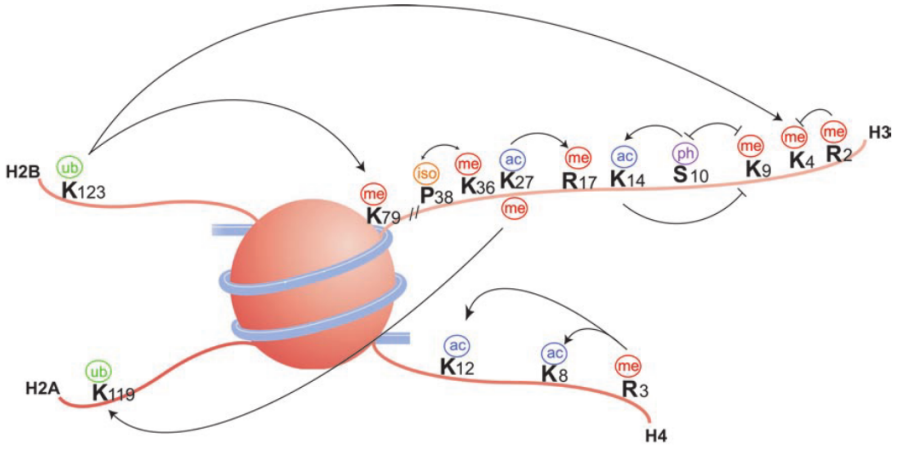


Figure 1.3.1: **Histone modifications interactions.**
Histone modifications can influence each other in a positive (arrowhead) or negative (flat head) effect. Reprinted from [Bannister 11]

might be impacted by adjacent modifications: for example phosphorylated H3S10 disturb the binding of HP1 to methylated H3K9 [Fischle 05], but on the other hand, enhance the recognition of H3 lysines by acetyltransferases [Clements 03]. An antagonistic competition might also occur if mutually exclusive modifications target the same site. The best example is the lysine residue, which might be acetylated, methylated, ubiquitylated or even sumoylated [Kouzarides 07].

Interplay between histone modifications and DNA methylation

Histone modifications are not only influencing each other, they may also cooperate with DNA methylation. Examples of interplay between DNA methylation and histone modifications are shown in Fig. 1.3.2.

As previously discussed in sub-section 1.2.1, DNA methylation correlates with gene repression. Methylated CpG dinucleotides can recruit MBD proteins (e.g. MeCP2), which in turn recruit histone deacetylase complexes (HDACs), resulting in chromatin remodeling and potential methylation of H3K9 by histone lysine methyltransferases [Nan 98]. An alternative pathway for direct methylation of H3K9 involving a lysine methyltransferase associated with MeCP2 was also described in the literature [Fuks 03] (Fig. 1.3.2a). On the other hand, histone modifications can direct DNA methylation. For instance, chromodomain proteins can bind to methylated H3K9 and mediate DNA methylation through DNMTs recruitment (Fig. 1.3.2b).

Interactions between histone modifications, as well as their interplay with DNA

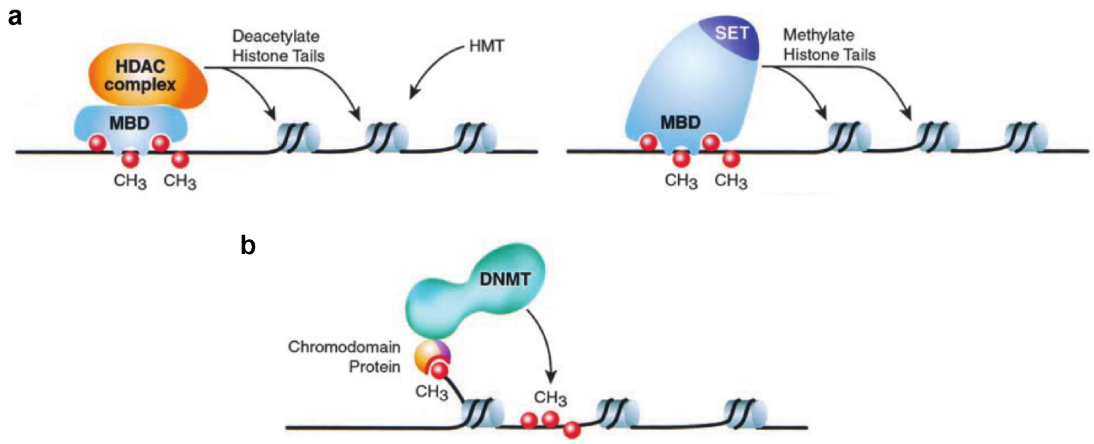


Figure 1.3.2: **Interplay between DNA methylation and histone modifications.**
(a) Proteins from the MBD family may recruit HDACs complexes to deactylate histones and indirectly favor histone methylation via HMTs. Alternatively, HMTs containing a MBD domain may also directly methylate histone tails. (b) Methylated histones may recruit DNMTs via a chromodomain protein to methylate DNA.
MBD: Methyl-CpG Binding Domain protein family; HDAC: Histone DeACetylase; HMT: Histone MethylTransferase; SET: catalytic domain involves in histone methylation; DNMT: DNa MethylTransferase. Adapted from [Zhang 01]

methylation suggest that one single chromatin modification doesn't act on its own, but a combination of modifications may function cooperatively to regulate cellular biological functions.

1.4 Defining cell identity based on histone modifications

1.4.1 Genome-wide mapping of histone modifications

Analysis of histone modifications mainly relies on Chromatin ImmunoPrecipitation technique (ChIP), in which an antibody is used to enrich genomic regions carrying a specific histone modification.

Originally, the presence or absence of a pre-defined regions in the immunoprecipitated DNA was determined by Polymerase Chain Reaction (PCR) but such studies were limited in the number of loci interrogated. The method rapidly evolved by

combining chromatin immunoprecipitation with DNA microarrays to profile chromatin modifications over large genomic regions (ChIP-chip). However this technique suffered from amplification bias and cross-hybridization. The emergence of sequencing technologies (and later Next-Generation Sequencing) contributed to the development of the Chromatin ImmunoPrecipitation followed by sequencing technique (ChIP-seq). ChIP-seq overcomes previous limitations as fewer amplification of immunoprecipitated DNA is required for sequencing and sequencing reads are directly aligned to the genome to create chromatin-state maps [Mikkelsen 07]. ChIP-seq is still considered today as the gold standard method for genome-wide analysis of histone modifications.

Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq)

Briefly, chromatin is fragmented by microccocal nuclease (MNase). This enzyme is particularly adapted as it preferentially cuts linker DNA, generating principally mono-nucleosomes under optimized conditions. The fragmented chromatin is immunoprecipitated with an antibody specific for the histone mark of interest. Unbound chromatin is discarded, whereas immunoprecipitated DNA is purified and amplified before sequencing. The sequenced fragments are aligned to the reference genome and yield genome-wide binding site maps. The number of sequencing reads detected at a genomic region correlates with the modification level of the region (see Fig. 1.4.1).

Alternatively, the enzymatic fragmentation of chromatin may be replaced by mechanical fragmentation (e.g. sonication). In this latter approach, chromatin is first crosslinked with formaldehyde and then sonicated. This second method offers the possibility to profile not only histone modifications but more generally DNA binding proteins and chromatin-modifying proteins. However, the crosslinking step may disrupt the target epitope, thus reducing the immunoprecipitation efficiency.

The resolution of ChIP-seq is directly linked to the size of the chromatin fragments, as well as the sequencing depth. The optimization of chromatin digestion conditions is essential to ensure that most of the sequenced fragments originate from a single nucleosome [Barski 07].

Constraints of chromatin immunoprecipitation methods

Several constraints are inherent to the ChIP method and should be kept in mind before any experiments [Kidder 11].

- Highly specific antibodies are essentials to generate good quality results. Only well-characterized antibodies must be used to ensure high specificity

and good sensitivity to the target epitope.

- Different method of chromatin fragmentation might lead to different results. Both mechanical and enzymatic fragmentation introduce specific bias such as fragment length or selective digestion. In addition, the number of cells, the fixation conditions, the type of sonicator and sonicator settings are also source of variation in ChIP-seq results. Therefore, technical and biological replicates as well as input control (fraction of DNA not immunoprecipitated) are valuable to generate reliable data.

- ChIP-seq was requiring large number of cells (10^6 to 10^8 cells per experiment), thus limiting the use of the technology to rare samples. Considerable efforts have been made recently to reduce the number of starting cells from millions down to tens cells [Ma 18]. However, the chromatin profiles obtained remain an average snapshot of the modification status, which could contain contributions from very heterogeneous modifications states of different cells.

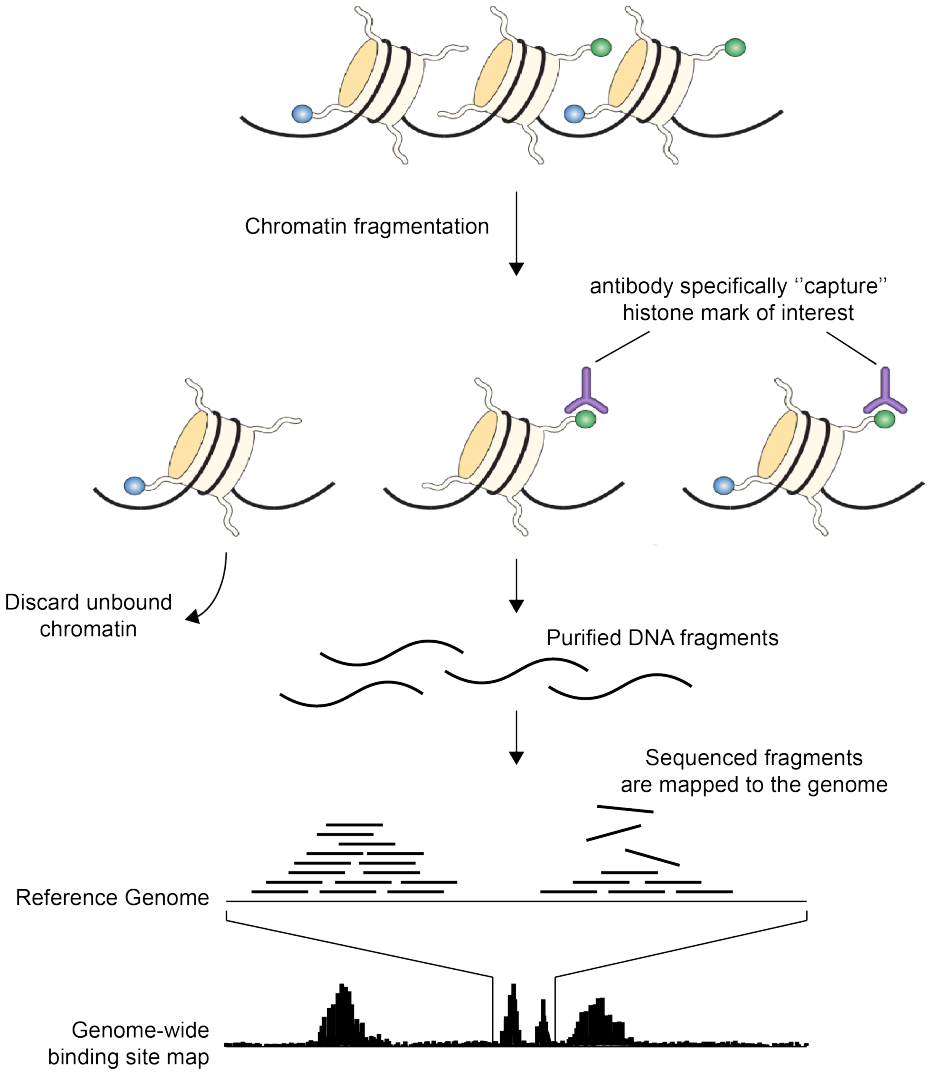


Figure 1.4.1: **Genome-wide mapping of histone modifications by Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq).**
Chromatin is fragmented to generate preferentially mono-nucleosomes. Nucleosomes bearing a histone modification of interest are "captured" by immunoprecipitation using an antibody that is specific to this particular modification. DNA fragments wound around those histones are purified, amplified and sequenced. Sequenced DNA fragments are mapped to the reference genome to generate genome-wide binding site maps in which peaks correlate with regions carrying the target modification. Adapted from [Schones 08]

1.4.2 Epigenomic signatures define cell-type identity

Systematic genome-wide mapping of histone modifications (also referred to as epigenomic profiling) have revealed reproducible patterns in their distribution, allowing predictions to be made about transcriptionally active or repressive chromatin states (see Fig. 1.4.2) [Bernstein 05, Barski 07, Ram 11].

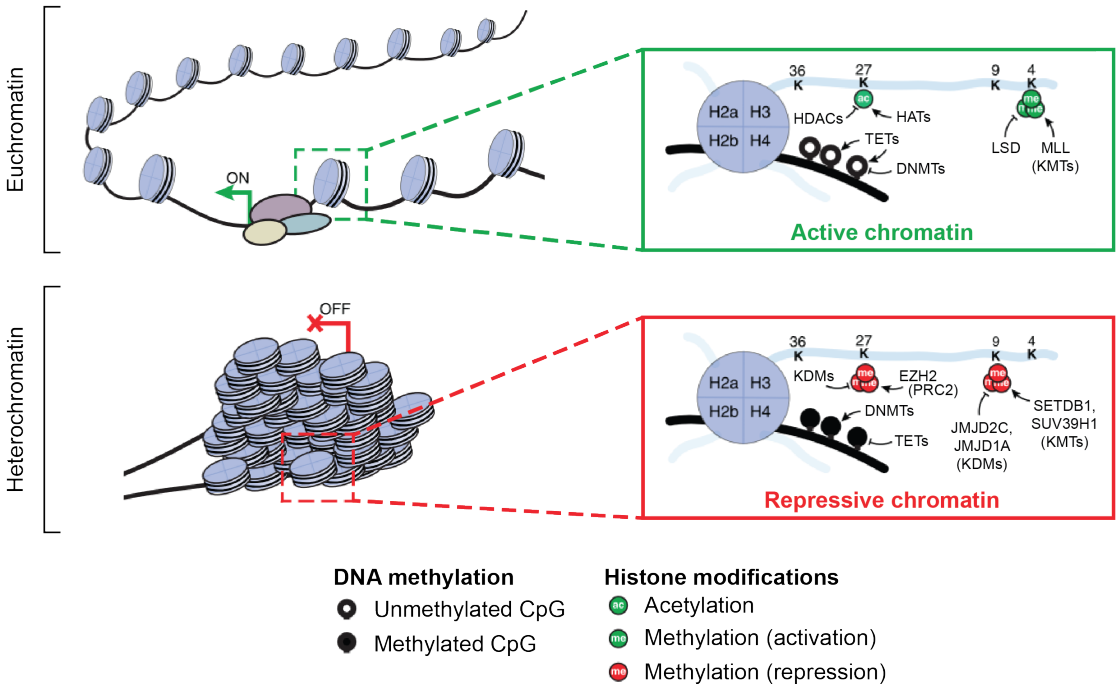


Figure 1.4.2: **Chromatin states define cell identity**
Active chromatin states are associated with high level of H3K4me3 in promoter regions and H3K27ac in enhancer regions. CpG islands are demethylated. Conversely, repressive chromatin states are associated with high level of H3K27me3 and H3K9me3. CpGs are methylated. Histone-modifying enzymes (HATs: Histone AcetylTransferases; HDAC: Histone DeACetylase; KMTs: Lysine MethylTransferase; LSD: Lysine Specific Demethylase; KDMs: Lysine DeMethylase containing Jumonji catalytic domain); DNMTs: DNa Methyl-Transferases; TETs: Ten-Eleven Translocation. Adapted from [Flavahan 17]

A core set of histone modifications has been attributed to specific genomic regulatory elements and chromatin domains. Active chromatin states are associated with high level of H3K4me3 in promoters, H3K4me1 and H3K27ac in enhancers

and H3K36me3 in the body of transcribed genes. Conversely, transcriptionally inactive chromatin state including heterochromatin regions, have been characterized with high levels of H3K27me3 and HK9me3 [Epigenomics 15].

These histone modifications contribute to the definition of epigenomic signatures within distinct chromatin states, which are highly indicative of cell type and tissue identity. The genome-wide profiling of these marks can be leveraged to understand the global landscape of genome regulation and then, for example, distinguish epigenomic differences in the context of normal and disease cell states [Epigenomics 15]. However the current state of the chromatin profiling technologies doesn't allow studying cellular heterogeneity nor detect cell-to-cell variation in chromatin states.

Chapter 2

Droplet-based microfluidics for single-cell epigenomic profiling

Cellular heterogeneity is a universal property of multicellular organisms, which contain diverse cell types originally classified on the basis of their phenotypic characteristics (location in the organism, morphology...). The definition of these cell types has started to evolve with the investigation of molecular characteristics (such as DNA, RNA, proteins, metabolites) but bulk analyses of tissues and cell populations only represent an average snapshot of the cellular components.

Recently, advances in cellular profiling has enabled the characterisation of the molecular heterogeneity at single-cell resolution, revealing a large diversity of cell "states" among similar phenotypes. Single-cell RNA-seq is at the forefront on the development of single-cell methods and provides in-depth analysis of gene expression profiles. However, quantifying gene expression in individual cells is challenging and limited by technical issues (e.g. capture and amplification of low amount of mRNA) as well as the stochasticity inherent in biological processes [Elowitz 02, Li 11a].

In order to reduce the impact of the noise present in single-cell measurements, two strategies have driven the technological development: (i) the increase of the number of variables measured, and (ii) the increase of the number of cells profiled [Prakadan 17]. The first strategy relies on the intuitive idea that the expression of a single gene might not be reliable of a cell state, but the co-variation of a set of genes is less impacted by noise. Similarly, increasing the number of cells contributes to a more effective characterisation of cell subpopulations that compose the sample. An illustration of the dramatic increase over the last decade in the number of cells profiled per single-cell RNA-seq experiment is shown in Fig. 2.0.1 (for review

[Angerer 17, Svensson 18]).

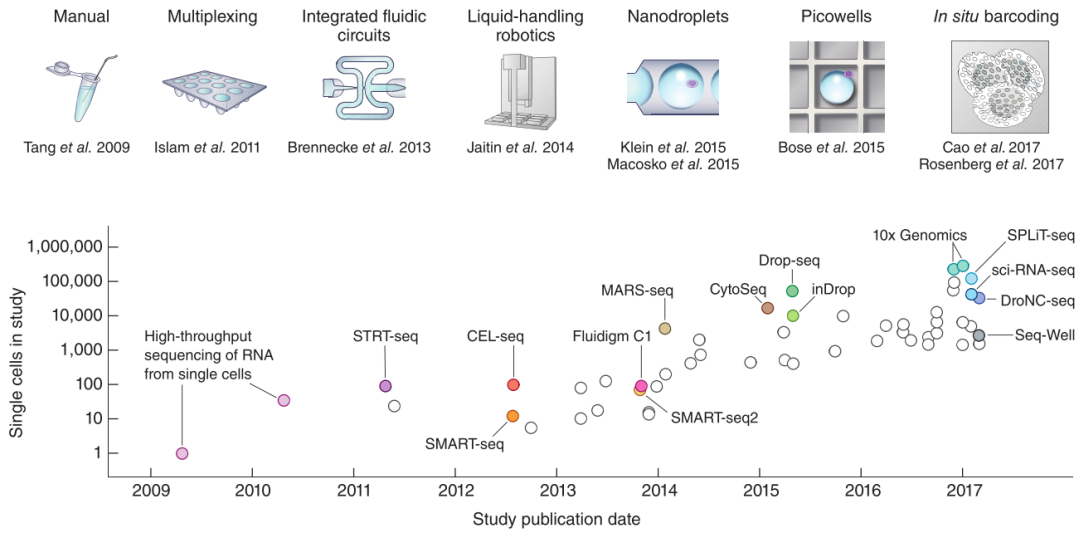


Figure 2.0.1: **Evolution of the number of cells profiled in single-cell RNA-seq experiments.**
Sequencing the whole transcriptome of one single-cell was first described by Tang et al in 2009 [Tang 09]. The number of cells assayed increased to ~100 with sample multiplexing, ~1,000 with the development of robotics and ~10,000 with the use of microfluidic devices. Recently, combinatorial *in-situ* barcoding methods have been developed to simultaneously process ~100,000 single-cells. Reprinted from [Svensson 18]

Tang et al sequenced for the first time in 2009 the whole transcriptome of a single cell [Tang 09]. In the following years, the method has been adapted on plate to allow multiplexing and sequencing of hundreds cells [Islam 11] and thousands cells with the use of robotics [Jaitin 14]. The development of microfluidic devices allowed a jump in throughput from hundreds [Brennecke 13] to thousands [Klein 15] and tens of thousands cells [Macosko 15, Bose 15]. Recently, combinatorial *in-situ* barcoding methods have been used to profiled hundred of thousands cells in parallel, opening doors for large scale studies [Cao 17, Rosenberg 18].

In this introductory chapter the scope of the discussion is limited to microfluidic systems with a focus on the droplets format as a method of choice to profile molecular states at single-cell resolution.

2.1 Scaling by shrinking

Opportunities offered by the microfluidic systems for single-cell analysis

Microfluidics is by definition the study of flows in micrometric systems [Tabeling 05]. The manipulation of liquid at the micrometric scale is of great interest in biology and offers the following advantages:

- Volumes: reduction by 10^6-fold compared to conventional assays in tubes (from milliliters to picoliters). Such volumes are on the same scale as individual mammalian cells.
- Compartmentalization: objects can be confined and isolated from each other in micrometric compartments. In the case of single-cell analysis, cells are captured, lysed and their components retained in the compartment for further processing.
- High-throughput: reduction in volume allows a massive increase in the number of experiments performed in parallel as illustrated in Fig. 2.0.1.
- Cost-efficiency: reduction of reagents, consumables and time [Agresti 10]
- Sensitivity: detection of small amount of analytes is enhanced due to their high concentration in small volume [Najah 12]

Microfluidic confinement strategies for single-cell analysis

Three main types of microfluidic devices have been developed to isolate objects and are used for single-cell analysis purposes: (i) valve-based microfluidic devices, (ii) nano- and picowells and (iii) droplet-based microfluidic devices (see Fig. 2.1.1, for review [Prakadan 17]).

Valve-based microfluidic devices are the first major method developed and probably the most sophisticated. Microfluidic channels are coupled with pressure-controlled valves (see Fig. 2.1.1a; for review on operation of valve-based microfluidics [Unger 00, Thorsen 02, Hong 03]). Opening and closing of valves create walls and isolate objects in nanoliter chambers. The channels and valves can be arrayed and controlled simultaneously on microfluidic devices allowing a broad range of operations such as adding and mixing reagents, incubation... [Fan 11]. The main limitations are the throughput (hundred of cells in parallel) and the complex design/fabrication of the devices.

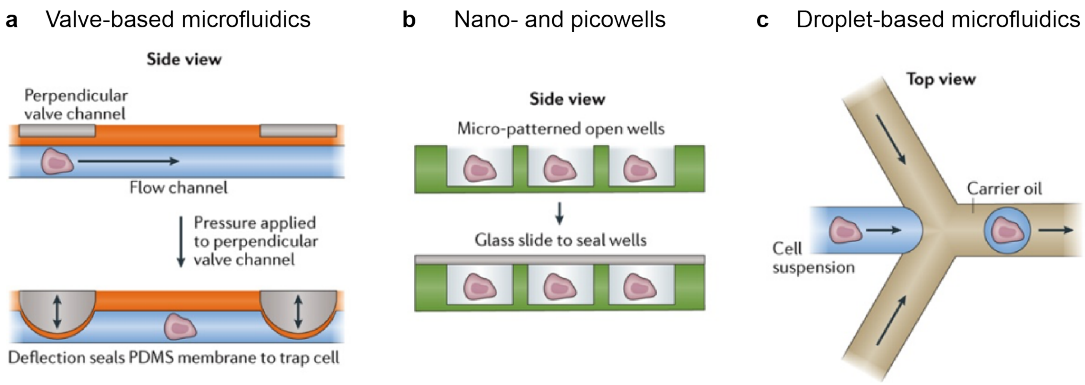


Figure 2.1.1: **Microfluidic confinement strategies for single-cell analysis.** (a) In valve-based microfluidics single-cells can be trapped by applying a pressure on channels perpendicular to the flow channel. The pressure applied deforms the channels and creates reversible "walls" that confine single-cells in microfluidic chambers. (b) Cells can be isolated into nanoliter or picoliter wells by gravity, which can be sealed with a glass slide. (c) Droplet-based microfluidics consist in isolating single-cells in small aqueous droplets dispersed in a continuous oil phase. Adapted from [Prakadan 17]

Arrays of tens of thousands microfabricated wells are a much simpler method to isolate single-cells than valve-based microfluidic systems, whose wells volume can range from nanoliter [Love 06, Gierahn 17] to picoliter [Bose 15] (see Fig. 2.1.1b). Cells can be simply loaded into the wells by gravity, but must be diluted to avoid multiple cells into the same well (see section 2.2.3 for loading occupancy and Poisson statistics). As a result, the number of cells processed in parallel can be scaled up to ten thousands [Bose 15]. In addition to the simplicity, one major advantage is the fixed spatial location of each well allowing multiplexed measurements and kinetic studies.

Alternatively, individual cells can be isolated in small aqueous droplets dispersed in a continuous oil phase (Fig. 2.1.1c). Droplets, whose volume span from nanoliters to femtoliters (10^{-9} to 10^{-15} litres), are stabilized by a surfactant and function as independent biocompatible microreactor. Droplet-based microfluidics have been utilized in this thesis, the opportunities and constraints offered by the droplets and their utility in single-cell analysis are discussed in the following sections.

2.2 Droplet-based microfluidics

Droplet-based microfluidics is a domain of microfluidics, in which aqueous droplets dispersed in a continuous oil phase are used as independent microreactor (sub-section 2.2.1). On-chip microfluidic modules allow droplets generation and manipulation at kHz frequencies and throughput can easily reach millions of drops per hour (sub-section 2.2.2). Droplet microfluidics have been used in a broad range of applications such as the high-throughput phenotypic screening of viruses [Chaipan 17], bacteria [Baret 09], yeast [Agresti 10, Beneyton 17], filamentous fungi [Beneyton 16] and mammalian cells [Clausell-Tormos 08], allowing for example highly efficient directed evolution [Agresti 10, Obexer 17] or the study of immune response dynamics at the single-cell level [Eyer 17]. Recently, droplet-based microfluidic systems combined with DNA barcoding and Next-Generation Sequencing (NGS) have enabled high-throughput single-cell genomics, transcriptomics and epigenomics analysis (discussed in section 2.3 of this introductory Chapter).

2.2.1 Key features of droplet-based microfluidics

Quantitative measurement

An important feature of droplet-based microfluidics is the generation of highly monodisperse droplets [Thorsen 01]. The initial concentration of each reagent is similar between droplets and a slight variation can be directly related to the activity of the encapsulated compounds. For this purpose, fluorophores are available and can be used in microfluidic workflows. For example, fluorogenic substrates are used in enzymatic reactions to measure the catalytic activity of an enzyme as a fluorescence readout. Production and analysis of droplets are generally operated on dedicated microfluidic instruments: each droplet is scanned by a laser beam as they pass in front of a detection point and the fluorescence signals are analysed in real-time [Mazutis 13].

Biocompatibility

Micro-fabrication techniques, such as soft-lithography [Xia 98], contributed to the development of prototype microfluidic device. Those microfluidic chips are easily made of polydimethylsiloxane (PDMS) with micrometric channels imprinted and bound on glass slides. PDMS is a material of choice for producing microfluidic chip as it is inexpensive, optically transparent, gas permeable and importantly, inert to chemical and biological reactions. In addition, oil and water can be injected in the PDMS chips to generate droplets by the shearing of the aqueous phase by the

oil phase [Anna 03].

The continuous oil phase is generally composed of perfluorinated oils which have the advantage to be hydrophobic, lipophobic and inert to biochemical reactions. Those properties are particularly important in droplet-based microfluidics as the solubility of organic molecules is reduced, thus retaining biological compounds inside the droplets. In addition, the high gas solubility in perfluorinated oils enables cellular respiration in droplets [Lowe 98, Mahler 15].

Finally, aqueous droplets are stable after production only if they are stabilized by a surfactant (lowering interfacial tension at the surface of the droplets). Surfactants are organic compounds that are amphiphilic, meaning that they contain both a hydrophilic group (head) and a hydrophobic group (tail). For this purpose, biocompatible surfactants composed of fluorophile perfluoro-tails coupled to hydrophyilic PEG head groups have been developed which enable for example the thermocycling of emulsions while preserving droplets integrity [Hindson 11].

High-throughput

Manipulating liquids at the micro-metric scale reduces the volume of reagents per assay. Droplet-based microfluidics outperform the other categories of microfluidic systems presented in 2.1 with respect to throughput. The frequency of the droplets production and analysis depends on the channels geometry and flow conditions, but a range from 0.1 to 30 kHz has been reported in the literature [Sciambi 15].

2.2.2 Manipulating droplets

Microfluidic devices allow the high-throughput generation of highly monodisperse aqueous droplets. Those droplets can be manipulated using a broad range of microfluidic modules, which can be assembled into fully integrated microfluidic chips (examples of microfluidic modules for manipulating droplets are shown in Fig. 2.2.1).

- Emulsions are made of an aqueous phase dispersed in an oil continuous phase and stabilized by a surfactant. Droplets volume span from nanoliter [Zilionis 17] to femtoliter [Leman 15] and can be produced using different channel designs and geometries [Thorsen 01, Anna 03, Cramer 04, Abate 09a, Li 11b, Li 15].
- Aqueous phases mix rapidly inside droplets by diffusion but microfluidic modules can be added to speed up the mixing of two or more phases [Tice 03, Bo Zheng 04].

- Droplet populations can be merged passively [Tan 04, Niu 08], or actively using a triggered electric field [Mazutis 09b, Niu 09, Zagnoni 09a, Zagnoni 09b].

- Droplets can be incubated on-chip as a single-file in delay lines (seconds), while preserving droplet order [Frenz 09]. Droplets can also be incubated for longer time (minutes) or in stationary chambers but in this case droplets order is not maintained [Courtois 08].

- Droplets fluorescence can be detected and measured in real-time as the droplets passed a laser beam. [Baret 09].

- Droplets of interest can be sorted out based on their physical properties in passive hydrodynamic selection [Chabert 08, Mazutis 09c]. Also, droplets can be actively sorted using external forces such as dielectric forces [Ahn 06, Sciambi 15] or acoustic waves [Franke 09].

- Droplets can be splitted symmetrically or asymmetrically in different ratios depending on the microfluidic channels geometry [Link 04, Abate 11]. Active splitting using a triggered electric field can also be used to divide droplets into smaller ones [Link 06].

- Droplets can be collected and incubated off-chip before being re-injected into a dedicated microfluidic device [Mazutis 09a].

The microfluidic modules illustrated in Fig. 2.2.1 are examples of high-throughput operations that can be performed in droplet-based microfluidics. It is tempting to combine these modules on integrated workflows and, for example, simultaneously generate droplets, measure their fluorescence, and sort the droplets that display the desired properties. However, the hydrodynamic resistance is inversely proportional to the volume of a rectangular microfluidic channel [Fuerstman 07]. Also, multiplying the number of microfluidic modules on the same device increases the hydrodynamic resistance and can make the device difficult to control.

Complex microfluidic workflows are usually performed using different microfluidic devices. However, emulsions are fragile and manipulating droplets in several microfluidic chips implies the collection and re-injection of the droplets, which can potentially induce coalescence. As a result, droplets might not be monodisperse which is problematic to precisely control the droplets inside the channels but also to compare them as identical independent microreactor. Again, the number of microfluidic operations that can be performed sequentially is limited.

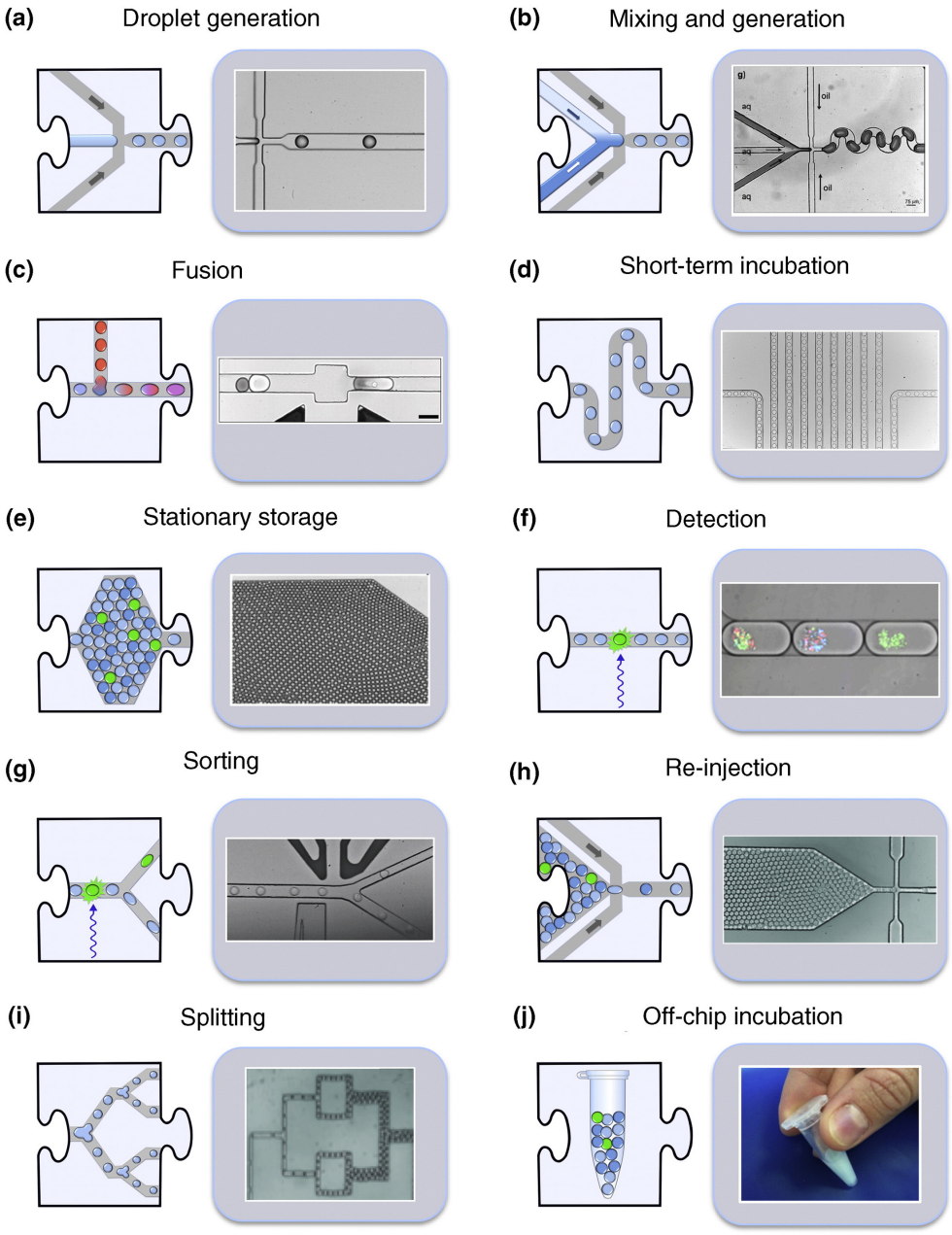


Figure 2.2.1: **Examples of microfluidic modules for manipulating droplets.** (a) droplet generation, (b) mixing two aqueous phases in droplets after generation, (c) active fusion between two droplets by electrocoalescence, (d) incubating droplets (seconds) without loosing droplets order, (e) incubating droplets (minutes to hours) or stationary storage, (f) measurement of droplet fluorescence, (g) fluorescence activated droplet sorting, (h) re-injection of droplets, (i) passive droplet splitting, and (j) off-chip incubation of droplets. Reprinted from [Kintses 10]

2.2.3 Predicting cell compartmentalization

Compartmentalization of non-interacting discrete objects in droplets follows a Poisson distribution [Shapiro 03], which describes the probability of finding a mean number λ of x objects per droplet (Fig. 2.2.2). The random encapsulation of cells requires diluting the samples to ensure at most one cell per droplet, resulting in a large majority of empty droplets [Köster 08, Clausell-Tormos 08, Huebner 08]. With a mean number of cell per droplet in the range of $\lambda = 0.1$ - 0.3, 90% to 74% of the droplets are empty. Some microfluidic devices were also developed to reduce the impact of the Poisson distribution, for example by sorting out empty droplets or ordering the cells in the microfluidic channel before encapsulation [Edd 08, Collins 15].

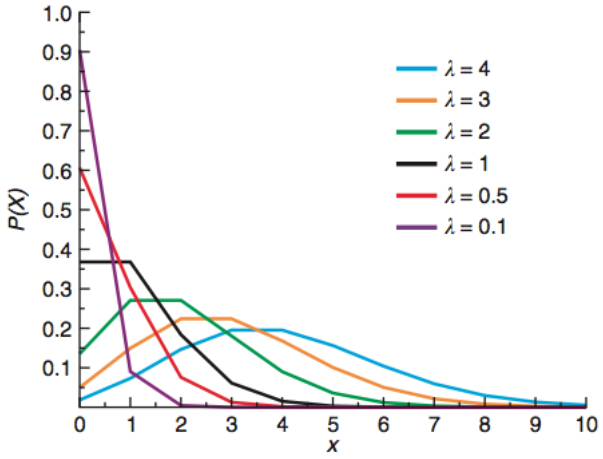


Figure 2.2.2: **Predicting the number of cells per droplets**. Distribution of the number of cells per droplets for different λ values. Reprinted from [Mazutis 13]

2.3 Single-cell -*omics* in droplets

2.3.1 Droplets barcoding

Why barcoding?

Barcoding refers to the need of multiplexing and analyzing multiple cells at once. RNA or DNA are present in minute amount in single cell and such quantities are by

far too low to be sequenced without prior amplification [Tang 09]. Adaptors were used to capture and drive amplification of nucleic acids from individual cell but they rapidly evolved to incorporate unique sequence specific to each cell isolated in well plates [Islam 11, Hashimshony 12, Jaitin 14]. With the emergence of high-throughput techniques such as the random encapsulation of cells in droplets, larger barcode libraries were needed to avoid the barcoding of two distinct cells with the same sequence.

General barcoding constraints

The barcoding strategy should be carefully designed to fullfil requirements and specificity of each application, notably in terms of number of cells multiplexed per assay. Two major methods have been described to generate large barcode libraries for single-cell analysis in droplets: (i) the combination of multiple barcode sequences and (ii) the synthesis of random barcode sequences [Svensson 18].

Combinatorial synthesis method was first described for the solid-phase synthesis of peptides in 1963 and immediately adopted to generate quickly large number of compounds [Merrifield 63]. This approach was adapted to measure gene expression of thousands single hematopoietic cells in a microwell array in which beads functionalized with oligonucleotide primers for mRNA capture were loaded in each well [Fan 15]. The oligonucleotide primers comprised a DNA barcode generated by combinatorial split-pool synthesis method. Briefly, beads were distributed into a 96-well plate with each well containing a different short DNA barcode that is added by ligation to the beads. The beads were then pooled and splitted again in a second 96-well plate. By repeating this split-pool three times, a library of 96^3 combinations is created (884,736 possible barcodes).

Similarly, Klein et al developed the InDrop method [Klein 15] by using hydrogel beads grafted with barcoded primers generated with 2 rounds of split-pool from 384-well plates. The diversity of the barcode library is reduced by a factor 6 resulting in 147,456 combinations. The authors estimated the maximum number of cells that can be sequenced to limit to 1% the probability of having 2 cells with the same barcode. They presented the problem as analogous to the "birthday paradox", with the total number of barcodes assimilated to the days of the year and the number of cells assimilated to the group of persons. In other words, the problem is to sample, without replacement, n elements among N. They defined the number of observed barcodes n_{obs} when sampling n elements by the following relation:

$$n_{obs} = N\,(1 - e^{-n/N})$$

Considering that the barcoding error is the fraction of cells with the same barcode:

$$error \approx 1 - \frac{n_{obs}}{n}$$

In practice, the number of encapsulated cells is much smaller than the total number of barcode combinations (n << N). The equation can be simplified:

$$error \approx \frac{n}{2N} \tag{2.3.1}$$

Applying the equation (2.3.1) to limit to 1% the probability of having 2 cells with the same barcode:

$$n_{cells} = 0.01 \text{ x } 2 \text{ x } 384^2 = 2,949 \text{ cells}$$

Increasing the length of the barcode or the number of split-pool cycles would automatically increase the number of cells than can be multiplexed. Alternatively, a second method consisting of split-pool single-based DNA synthesis on beads has been used to dramatically increase the diversity of the barcode library (12 split-pool rounds generating 4^{12} ~16.7 million combinations). Such method has been reported in Drop-seq [Macosko 15] to explore mouse retinal tissues and identify ~40 transcriptionally distinct cell populations from ~44k cells.

How delivering barcodes in droplets?

In Drop-seq [Macosko 15] and InDrop [Klein 15] methods, barcodes are delivered in droplets by co-encapsulation of cells and barcoded beads. However, in Drop-seq, barcoded beads are small microparticles that follow a distinct Poisson distribution compared to the cells. As a result, the probability to find both one cell and one bead in the same droplet is defined by the product of the probability of finding one cell and the probability of finding one bead. Considering a typical mean number of cell per droplet of $\lambda = 0.3$ (22.2% of single cell), the mean number of bead should also be $\lambda = 0.3$ to ensure at most one bead per droplet (22.2% of single bead), resulting in only ~5% of droplets containing both one cell and one bead. The majority of the cells (~4/5) won't be co-encapsulated with a barcoded bead, leading to important loss of information which might be an issue when processing limited cell numbers.

Klein et al used hydrogel beads instead of solid microparticles in InDrop method. Hydrogel beads have the particularity to be deformable [Kim 07] and can be closely packed in a single-file as they approach the droplet generation junction. By synchronizing the regular flow of hydrogel beads with the periodicity of the droplets

formation, Abate et al demonstrated nearly 100% hydrogel droplet occupancy [Abate 09b]. As droplets must contain one cell and one bead to produce a sequencing library, Klein et al limited cell loss with the use of those hydrogel beads.

In other applications requiring enzymatic DNA digestion such as in single-cell epigenomics studies, barcodes can't be loaded directly with the cells in droplets and as to be delivered *a posteriori*. For this purpose, alternative strategies have been developed. Droplet-based libraries of oligonucleotide barcodes have been first described by Rotem et al for single-cell RNA-seq [Rotem 15b] and later adapted for single-cell ChIP-seq [Rotem 15a]. Oligonucleotides are directly emulsified from 2 x 384-well plates and collected through a single output. The barcode-containing droplets are then fused with cell-containing droplets. The low complexity of the barcode library (1,152 possible barcode combinations) limits the collection of ~100 cells to ensure 95% of the barcodes are unique to a single-cell. This issue can be mitigated by collecting multiple samples and adding a second sample-specific barcode during sequencing library preparation.

To overcome this limitation, single-molecule amplification in droplets [Zhang 12] has been used to generate droplet-based libraries of barcodes [Lan 16, Lan 17]. The barcodes are generated by diluting oligonucleotides in accordance with Poisson statistics so that ~1 in 10 droplets contains a single molecule. Oligonucleotides are encapsulated with PCR reagents in droplets for amplification and generating a clonal population of the single molecule. This method is efficient to generate large libraries of million combinations but leads to a majority of empty droplets due to the Poisson distribution as previously described.

2.3.2 State of the art in single-cell -*omics* using droplet-based microfluidics workflows

Single-cell analysis is a rapidly evolving field of research and development. The flexibility of the droplets format can be leveraged to investigate different molecular layers of individual cells (and hence cell identity and function). Striking examples with respect to genomic, transcriptomic and epigenomic profiling of single cells using droplet-based workflows are discussed in this sub-section.

Interrogating genome complexity at single-cell resolution

The development of single-cell genomics in droplet microfluidics is hampered by technical challenges in isolating, purifying and amplifying genomic DNA from single-cells and by the high read coverage necessary to cover the entire genome.

Consequently, most of the single-cell genomics studies still rely on valve-based microfluidics systems, in which the number of cells is *de-facto* limited.

Fu et al and later Hosokawa et al overcame part of the technical issues regarding Whole Genome Amplification (WGA) quality by adapting mulitple-displacement amplification in droplets [Fu 15, Hosokawa 17]. The method has been used to detect Copy Number Variation (CNV) and call Single Nucleotide Polymorphism (SNP) in ovarian cancer cell lines revealing distinct populations of cells with multiple CNVs and chromosomal abberations [Leung 16].

Quantifying gene expression at single-cell resolution

Study of the gene expression at the single-cell level is already having an important impact in exploring tissue heterogeneity, developmental processes or identifying gene regulatory mechanisms. Two major methods pioneered the development of single-cell RNA-seq using droplet-based microfluidics: (i) InDrop by Klein et al [Klein 15] and (ii) Drop-seq by Macosko et al [Macosko 15]. Both methods rely on the co-compartmentalization of cells in droplets with a bead carrying barcoded polydT primers for mRNAs capture and initiation of the reverse transcription. In Drop-seq, mRNA are captured on beads and the reverse transcription is taking place in bulk, while in InDrop, the reverse transcription occurs in droplet before breaking the emulsion. They showed similar technical performance with >95% cell specificity and 5k to 10k recovered genes per cell. A slight advantage in terms of sensitivity for the Drop-seq method with ~12% RNA capture (~7% for InDrop).

The commercialization of both methods has "democratized" the use of single-cell RNA-seq, making it available to a larger public: 1CellBio Inc. is commercializing the InDrop method, Drop-seq is available via Dolomite Bio, Illumina and BioRad have started a collaboration to develop the ddSEQ Single-Cell Isolator platform and 10x Genomics is providing a popular hybrid method between InDrop and Drop-seq.

This popularity has led to the recent development of alternative methods:

The requirement for tissue dissociation and the preparation of single-cell suspension in single-cell RNA-seq can be difficult to achieve with clinical samples or fragile tissues such as brain samples. To overcome this limitation, Habib et al introduced last year DroNc-seq for massively parallel single-nucleus RNA-seq [Habib 17]. They combined previous low-throughput single-nuclei RNA-seq methods [Grindberg 13, Habib 16, Lake 16, Lacar 16] with Drop-seq to profile ~40k nuclei from mouse and human brains revealing most of the known brain cell types.

Single-cell RNA-seq doesn't provide phenotypic information and measuring simultaneously gene expression and proteins in single-cells remained limited in scale to few genes and proteins in parallel [Ståhlberg 12, Frei 16, Albayrak 16]. Recent studies overcome in part those limitations by pairing transcriptome sequencing with cell surface protein markers of thousands cells using droplet-based microfluidics [Shahi 17, Stoeckius 17, Peterson 17]. For this purpose, cells were pre-labeled before compartmentalization in droplets with antibodies conjugated to DNA tags that can be captured and amplified as mRNAs by the oligo-dT primers on the barcoded beads. Using this method, Peterson et al quantified up to 82 proteins to characterize the activation of naïve $CD8^+$ T-cells isolated from the blood of 3 donors when treated with agonist monoclonal antibodies to CD27. They found 16 differentially expressed proteins across the 3 individuals and showed that the gene expression level and the protein abundance were not always correlated, suggesting a better stability of the proteins compared to mRNAs. Alternatively, antibodies can be replaced by aptamers probes to allow multiplexing and avoid the use of expensive antibody-tag compounds [Delley 18].

More specific applications also emerged such as the high-throughput sequencing of immune repertoires (for review [Georgiou 14, Chattopadhyay 14, Friedensohn 17]). Methods for recovery of V_H/V_L antibody coding sequences have been described [DeKosky 13, McDaniel 16], as well as the pairing of α and β chains of T-cells receptors [Grigaityte 17].

Profiling chromatin states at single-cell resolution

Chromatin accessibility has been investigated at single-cell resolution but not adapted to the droplet format yet (single-cell ATAC-seq). These methods were developed using Fluidigm C1 system with a limited thoughput of hundred cells [Buenrostro 15] or by combinatorial indexing of fragmented DNA [Cusanovich 15]. However, new studies based on droplet microfluidics could take advantage of the announcement made by 10x Genomics during the AGBT 2018 conference (Advances in Genome Biology and Technology). 10x Genomics is about to release in 2018 a new single-cell ATAC-seq product to interrogate chromatin accessibility of thousands cells at single-cell resolution. As proof of concept, researchers from 10x Genomics profiled 1,000 PBMCs and the data analysis revealed all the major cell types present in the human blood with a genomic coverage profile similar to bulk ATAC-seq assay (unpublished, data available on 10x Genomics website).

Single-cell profiling of chromatin landscapes is largely uncharted and only one method has been reported in the literature for the mapping of histone modifications

by single-cell Chromatin ImmunoPrecipitation followed by sequencing (scChIP-seq) [Rotem 15a]. This method is detailled in the following section 2.4.

2.4 Mapping histone modifications at single-cell resolution

Constraints specific to single-cell epigenomics assay

Optimizations of ChIP-seq protocols in terms of immunoprecipitation conditions and amplification procedures have reduced input material from millions to hundreds cells without loosing resolution in the identification of enriched or depleted regions [Adli 10, Brind'Amour 15, van Galen 16, Ma 18]. However, achieving single-cell resolution in epigenomics studies implies additional constraints and requires careful consideration in analysis as the presence or absence of sequencing read is the primary readout of the method:

1. ChIP-seq is an indirect chromatin profiling method in which antibodies are used to enrich target loci (see introductory Chapter 1, section 1.4.1). Background noise arises from non-specific antibody pull-down which tends to increase as the amount of target epitope decreases [Schwartzman 15, Rotem 15a, Clark 16].

2. Most of the sequencing reads are non-specific and there is no direct way to distinguish true reads from false positives at single-cell resolution [Clark 16].

3. Single-cell epigenomics libraries might have low mappability rates and high proportion of duplicates reads (e.g. as reported in single-cell bisulfite sequencing [Smallwood 14]).

4. Estimation and control of technical sources of variation is challenging. Unlike single-cell RNA-seq in which spike-in standards can be used to assess the level of technical noise, such strategy can't be applied in single-cell ChIP-seq assay.

5. Single-cell epigenomics data are sparse as hundreds reads are typically recovered per individual cell.

For those reasons, cell-to-cell variation in histone post-translational modifications remains largely uncharted. Only one method has been reported so far in the literature and is detailled below.

Drop-ChIP

Rotem et al published in 2015 the Drop-ChIP method to perform ChIP-seq at single-cell resolution by combining droplet-based microfluidics, DNA barcoding and Next-Generation Sequencing (see the procedure in Fig. 2.4.1) [Rotem 15a]. The use of droplet microfluidics can mitigate the limitation associated with non-specific noise in low-input experiment by indexing chromatin from thousands cells isolated in droplets and combining the content of all droplets before immunoprecipitation.

The authors developed the Drop-ChIP method with the following characteristics:

- Barcode-containing droplets are generated from 2 x 384-well plates as previously described in sub-section 2.3.1 and reported by Rotem et al [Rotem 15b]. The diversity of the barcode library is low (only 1,152 possible combinations) thus limiting the collection of ~100 cells to ensure that 95% of the barcodes are unique to a single-cell. To increase the throughput, the authors collected and processed multiple fractions of ~100 cells in parallel and added a sample specific barcode before sequencing to allow multiplexing.

- Cells are compartmentalized in droplets, lysed and incubated off-chip for chromatin digestion by microccocal nuclease enzyme (MNase).

- Nucleosomes-containing droplets and barcode-containing droplets are paired and fused in addition with ligation reagents in a three-point merger device (Fig. 2.4.1a). The rationale behind the separate encapsulation of cells and beads is to prevent the barcodes digestion by MNase. To inactivate the enzymatic activity, EGTA is injected with the ligation reagents in order to chelate Ca^{2+} ions which are necessary for MNase activity.

- After fusion, droplets are incubated off-chip to allow ligation of the barcodes to the nucleosomes. Barcodes are ligated to both ends of the nucleosomes allowing the attribution of each sequencing read to its originating cell.

- Barcoded-nucleosomes from ~100 cells are combined and mixed with chromatin carrier from a different specie for chromatin immunoprecipitation. Enriched DNA is then amplified by PCR using primers complementary to a universal sequence on the barcode (Fig. 2.4.1b).

- Sequencing reads are aligned to the reference genome and deconvoluted based on their barcode sequence. Clustering of the single-cell data highlights subpopulations whose chromatin profiles can be reconstructed by aggregating reads from all the cells present in the given subpopulations (Fig. 2.4.1c).

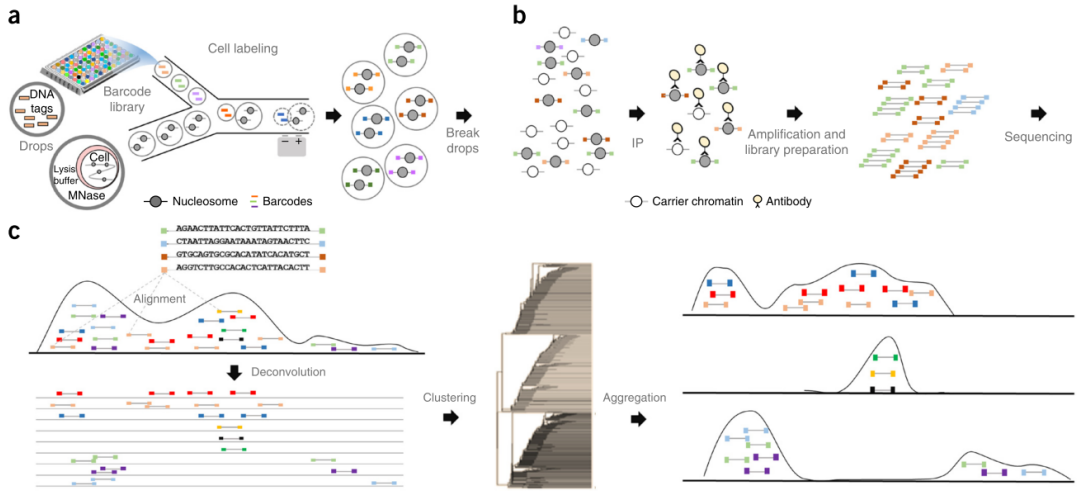


Figure 2.4.1: **Overview of Drop-ChIP procedure developed by Rotem et al.**
(a) Cells are compartmentalized in droplets with the lysis buffer and the micrococcal nuclease enzyme (MNase) for chromatin digestion. Nucleosomes-containing droplets are re-injected in a microfluidic device and merged one-to-one with droplets containing DNA barcodes to allow nucleosomes barcoding at single-cell resolution. (b) Barcoded nucleosomes are combined and mixed with carrier chromatin for immuno-precipitation. Enriched DNA is amplified and sequenced. (c) Sequencing reads are deconvoluted by their barcode sequence to generate single-cell binding site maps. Single-cell profiles are clustered and reads are aggregated to generate chromatin profiles specific to each sub-populations. Reprinted from [Rotem 15a]

The authors demonstrated the performance of Drop-ChIP by profiling H3K4me3 of a mixture composed of mouse embryonic stem cells (ES), mouse embryonic fibroblasts (MEF) and erythroid myeloid lymphoid cells (EML, hematopoietic cell line). Despite the sparsity of the data (only hundreds reads per cell), they showed that 95% of the cells were correctly clustered indicating that single-cell data were sufficient to distinguish cell types based on their epigenetic profiles.

In a second step, they used Drop-ChIP to distinguish variation in H3K4me2 chromatin states in mouse embryonic stem cells cultured in serum completed with Leukemia Inhibitory Factor (LIF). Initial unsupervised clustering didn't reveal subpopulations but they defined a set of 91 "signatures" derived from combination of published ChIP-seq datasets for histone modifications, transcription factor

and chromatin regulators binding sites... and performed a signature-based clustering. This second approach revealed 3 subpopulations with distinct chromatin patterns over Polycomb and pluripotency-related signatures, suggesting different stages of cellular differentiation among ES cells.

Chapter 3

Studying epigenetic intratumoral heterogeneity to better understand the emergence of therapeutic resistance

Genetic alterations such as mutations, gene rearrangement and copy number variation are the traditional molecular hallmark of cancer. Tumors are composed of highly heterogeneous cell subpopulations, which are universally associated with aberrant gene expression, abnormal cell growth and plasticity to internal and external stimuli. However genetic intratumoral heterogeneity can't explain solely the large diversity in morphological and physiological properties of tumor cells. It might not be just a coincidence that most of the genetic mutations are located on genes encoding regulators of the epigenome, indicating that changes in epigenetic states also profoundly affect initiation and cancer progression [Shen 13].

Several studies have shown that the response to therapy of tumor-derived cell lines led to the emergence and survival of rare subpopulations with distinct drug-tolerant states. Interestingly, the phenotype of these "persister" cells is reversible, in other words, the cells are able to restore their original phenotype after removal of the drug and reacquire a drug-tolerant state after re-exposure to treatment. This reversibility suggests that non-mutational mechanisms (i.e epigenetic mechanisms) and cell-to-cell variability may also drive the therapeutic response and give rise to drug resistance [Sharma 10, Knoechel 14].

3.1 Epigenetic abnormalities in tumors

Cancer is historically considered as a genetic disease but cancer cells also exhibit profound alteration of their epigenome. Nearly all epigenetic mechanisms play important role in cancer initiation, as well as in the progression of the disease or in response to therapeutic treatment. This section focuses on functional consequences of the deregulation of covalent chromatin modifications in tumor cells and introduces examples of chromatin-mediated drug resistance.

3.1.1 Chromatin modifications are disrupted in malignant cells

Epigenetic abnormalities can disrupt genome integrity on a global scale (e.g complete loss of heterochromatin structure), but also alter specific gene expression programs notably through aberrant regulation of tumor suppressor genes and overexpression of oncogenes (see Table 3.1).

Epigenetic mark	**Alteration**	**Functional consequences**
DNA methylation	Hypermethylation	Transciption repression
	Hypomethylation	Transcription activation
	Repeats hypomethylation	Transposition, recombinant genomic instability
Histone modifications	Loss of H3 and H4 acetylation	Transciption repression
	Loss of H3K4me3	Transciption repression
	Loss of H4K20me3	Loss of heterochromatic structure
	Gain of H3K9me and H3K27me3	Transciption repression

Table 3.1: **Functional consequences of altered chromatin modifications in cancer.**
Reprinted from [Portela 10]

DNA methylation level is globaly reduced in malignant cells compared to normal cells, with 20% to 60% loss of methylation in CpG islands depending on the cancer type. Global genome-wide hypomethylation has been associated with chromosomal instability, which have the potential to induce tumor initiation as reported in the formation of T-cell lymphomas in mice [Gaudet 03]. In addition, hypomethylation

has been associated with transcription activation of oncogenes and with loss of imprinting in several cancer types [Ito 08].

Conversely, hypermethylation of normally unmethylated CpG islands have been well characterized in several cancer types so that some discrete hypermethylated loci have been proposed as new biomarkers in tumor diagnostic and prognosis (for review [Li 09b, Kelly 10]). Unsurprisingly, the latter hypermethylated CpG islands have been found on promoter regions of tumor suppressor genes, DNA repair genes, cell cycle control genes or apoptosis-associated genes (for review [Esteller 07]).

As discussed in Chapter 1 section 1.3, histone modifications play important role in the maintenance of chromatin structure and in cellular processes including transcription and DNA repair. Consequently, alterations in histone modification patterns have been extensively related to tumor initiation and progression, both at the global scale and at specific loci. Histone modifying and reading enzymes (*writers*, *erasers* and *readers*) are frequently mutated in tumors [Shen 13], as reported with the overexpression of HDACs resulting in a global loss of histone acetylation and transcription repression in different cancer types [Fraga 05]. Similarly, gain and loss of histone methylation are mainly due to aberrant expression of methylases and demethylases. An important example is the overexpression of the methyltransferase EZH2, a subunit of the Polycomb Repressive Complex 2 (PRC2), which is notably responsible for the methylation of H3K27 leading to genes silencing [Zhou 02].

3.1.2 Epigenetic intratumoral heterogeneity contributes to therapeutic resistance

Chemotherapy is a standard method of treatment for cancer but the effectiveness is often reduced by drug resistance. An important challenge in cancer research is to understand the origin of such therapeutic failure. Drug resistance might arise due to the selection and expansion of rare pre-existing sub-clones (*intrinsic resistance*) or by the development of new resistant sub-clones (*acquired resistance*) [Almendro 14]. Unfortunately, the mechanisms of resistance remain poorly understood, in part due to the lack of methods that can resolve cell-to-cell variability and characterize rare subpopulations.

A striking example of therapeutic resistance driven by epigenetic intratumoral heterogeneity was reported in the testing of γ-secretase inhibitor (GSI) in T-cell Acute Lymphoblastic Leukemia (T-ALL). Knoechel et al identified a pre-existing subpopulation of GSI-tolerant "persister" cells that were able to expand during GSI treatment [Knoechel 14]. The reversibility of their phenotype seemed to in-

dicate that the heterogeneity is driven by epigenetic rewiring rather than genetic alterations. This hypothesis was highlighted experimentally by a more condensed chromatin structure associated with elevated level of repressive histone modifications and Heterochromatin Protein 1 (HP1). In addition, the authors demonstrated the dependency of the persister cells to BRD4, a bromodomain protein reader of histone acetylation. Interestingly, the combination of GSI and JQ1 (inhibitor of bromodomain proteins) led to a significant prolonged survival *in vivo* using patient-derived xenograft models of T-ALL.
Similarly, Sharma et al identified a rare drug-tolerant subpopulation in non-small cell lung cancer-derived cell lines. The phenotype of these persister cells was also reversible, suggesting an epigenetic-mediated mechanism of therapeutic resistance. Like Knoechel et al, the authors also combined chromatin modifying agents (inhibitors of histone deacetylase and histone demethylase) with classical anticancer agents to dramatically reduce the survival and the proliferation of the persister cells [Sharma 10].

These findings established an important role for epigenetic intratumoral heterogeneity in the emergence of therapeutic resistance. Combined therapeutics (e.g. GSI + JQ1 in T-ALL) offer exciting new possibilities for cancer treatment via the incorporation of epigenetic modulators with more conventional anticancer agents.

Altogether, these results also highlight the limitations with profiling bulk tumor samples. Such measurements yield averaged "snapshots" of epigenetic landscapes, which are not representative of the sample heterogeneity. To get a systematic understanding of the epigenetic intratumoral heterogeneity and gain insight in tumor evolution / response to treatment, single-cell epigenomic technologies are required.

3.2 Deciphering intratumoral heterogeneity at the single-cell level

Conventional methods applied to tumor samples are facing important limitations due to the extensive heterogeneity of tumor cells. Those methods yield average read-out, principally driven by the dominant tumor cell population. Consequently, signals from rare subpopulations such as drug-tolerant persister cells are masked, making them undetectable. The recent progress in single-cell technologies have provided valuable insight about genomic and transcriptomic heterogeneity in tumors, but little is known about epigenomic landscapes at the single-cell level.

Single-cell sequencing approaches can be leveraged in two ways in cancer studies: (i) the decomposition of intratumoral heterogeneity with the characterization of tumor cell subpopulations as well as their interactions with the tumor microenvironment, and (ii) the analysis of rare cell subpopulations such as drug-tolerant persister cells, Circulating Tumor Cells (CTCs) or Cancer Stem Cells (CSCs).

Originally, single-cell genomic studies were carried out to infer genetic intratumoral heterogeneity through the identification of somatic mutations and copy number variations occuring during tumor evolution [Navin 11, Wang 14]. Transcriptomic studies now complement genomics as part of an indirect readout of intratumoral epigenetic states. Indeed, massively parallel single-cell RNA-seq has proven its utility in identifying and characterizing unknown subpopulations (e.g. haematopoietic lineages [Jaitin 14, Villani 17]). Patel et al used scRNA-seq to investigate intratumoral heterogeneity in glioblastoma, revealing new expression subtypes compared to bulk-derived classification. Importantly, they showed that a higher intratumoral heterogeneity might lead to a decreased patient survival, highlighting again the importance of such studies at single-cell resolution [Patel 14]. Single-cell RNA-seq has also been used to profile malignant cells taking into account their spatial context including the surrounding microenvironment. Depending on their location, tumor cells can be associated with distinct gene expression signatures related to either stress, hypoxia, cell cycle or drug resistance [Tirosh 16, Puram 17].

The technological development of single-cell epigenomics is not as advanced as single-cell transcriptomic, mainly due to technical constraints. Particularly, single-cell epigenomics applied to cancer research is almost nonexistent, except for few studies, nonetheless limited to cancer cell lines.

Several studies, all measuring chromatin accessibility and cell-to-cell variation in transcription factor binding sites within cancer cell lines, have been reported in the literature [Buenrostro 15, Cusanovich 15]. Datasets are sparse but still sufficient to distinguish cell types and evolution of functional regulatory elements during progression of Acute Myeloid Leukemia (AML) [Corces 16]. Interestingly, Corces et al suggested that accessible chromatin states defined by single-cell ATAC-seq could more precisely reflect cell identity and disease evolution trajectory than transcriptomic profiles obtained by single-cell RNA-seq. Single-cell ATAC-seq is becoming a method of choice for mapping chromatin landscapes and efforts have been recently made to make the technology more widely applicable with the use of primary tissues [Preissl 18].

DNA methylation has been extensively studied in bulk tumor samples following

the development of Whole Genome Bisulfite Sequencing method. WGBS provides a high coverage with ~90% of ~28.7 million CpGs in the human genome. Its adaptation to single-cell is not straightforward as it requires few nanograms of DNA and the harsh chemical reactions necessary to transform 5mC also degrade DNA. As a result, the coverage per cell is relatively sparse and the resolution is too low to reliably identify cell-to-cell variation in methylated CpGs [Farlik 15, Smallwood 14]. Like scATAC-seq, single-cell methylome techniques are being optimized, applied to primary tissues [Luo 17] and might become a valuable tool to profile DNA methylation patterns across individual cells.

Single-cell sequencing approaches have the potential to assess the complexity of tumors. Genomic, transcriptomic and epigenomic measurements at single-cell resolution can provide valuable insight about multiple facets of cancer biology including intratumoral heterogeneity, subpopulation characterization, disease progression and outcome prediction, emergence of drug resistance...

3.3 Scope of the thesis

Histone post-translational modifications play important role in structuring the chromatin inside the nucleus, as well as in DNA-based processes including the regulation of genetic expression. Epigenomic profiling has been applied to create genome-wide histone modifications maps and define cell-type specific chromatin states. However, current methods are insensitive to cell-to-cell variability and only yield averaged profiles, limiting its application to heterogeneous samples. Yet, only one system has been reported in the literature to profile histone modifications at single-cell resolution. Rotem et al used Drop-ChIP to reveal distinct chromatin states within a population of embryonic stem cells (see introductory Chapter 2, section 2.4) [Rotem 15a]. However the number of enriched loci detected per cell is limited to few hundreds, which doesn't provide sufficient information to reliably decompose intratumoral heterogeneity nor differentiate rare cell types such as drug-resistant cells from tumor samples.

For this purpose, we imagined and conceived an alternative single-cell ChIP-seq platform combining droplet-based microfluidics with DNA barcoding technique and Next-Generation Sequencing technology. The method relies on the compartmentalization of single cells and chromatin indexing in droplets. After barcoding, the content of all droplets is combined for immunoprecipitation and enriched fragments are amplified prior sequencing. Chapter 4 on page 49 is centred around the development of the microfluidic workflow towards the generation of reliable and high quality single-cell chromatin profiles.

We next sought to determine whether the single-cell chromatin profiles generated enabled the precise identification of distinct cell types. For this purpose, we benchmarked the scChIP-seq platform in a serie of model experiments and confirmed that the single-cell chromatin profiles recapitulate cell type-specific chromatin states with high accuracy. Importantly, the final cell coverage increased by 5 to 10-fold as compared to previously reported Drop-ChIP method, enabling to distinguish patterns of cell-to-cell variation in complex heterogeneous samples. The proof of concept study is discussed in Chapter 5 on page 83.

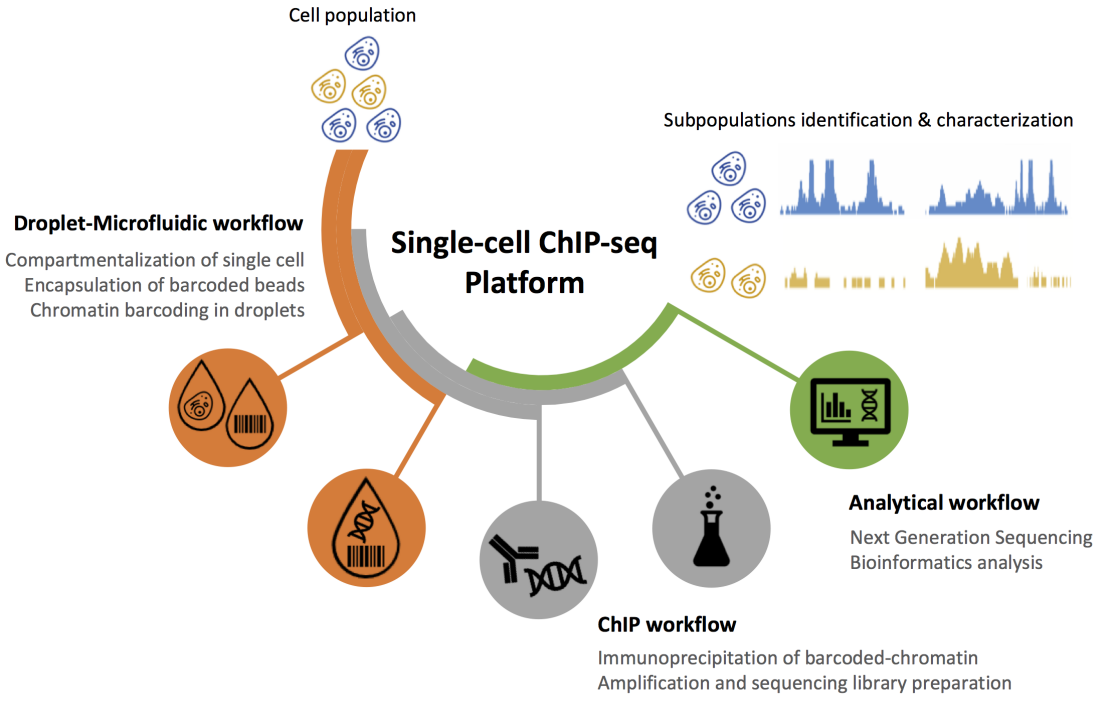
These promising results led us to collaborate with Dr. Céline Vallot's group at Institut Curie. In patient-derived xenograft (PDX) models of breast cancer with acquired resistance, scChIP-seq identified rare populations of cells in untreated drug-sensitive tumor with a chromatin landscape similar to the resistance cells after treatment. The outcome of the study is presented in the final part of Chapter 5 on page 97 in the form of a publication manuscript [Grosselin et al, *in preparation*].

This PhD thesis is part of a collaboration between the Laboratory of BioChemistry at ESPCI Paris and HiFiBiO Therapeutics. It has been supervised by Prof. Andrew Griffiths, director of LBC, and Dr. Annabelle Gérard, director of external partnership at HiFiBiO Therapeutics. The LBC has pionereed the development of droplet-based microfluidics and is applying such technologies in a wide range of applications, from high-thoughput screening to evolutionary biology. HiFiBiO is an antibody drug discovery company which has notably developed innovative single-cell approaches for deep-mining of immune repertoires [Gérard et al, *submitted*].

Results

Chapter 4

Development of a single-cell ChIP-seq platform for the mapping of histone post-translational modifications at the single-cell level



Chromatin states defined in part by histone post-translational modifications are highly indicative of cell type and tissue identity. Genome-wide mapping of histone modifications have revealed patterns in their distributions, allowing predictions to be made about transcriptionally permissive or repressive chromatin states [Bernstein 05, Barski 07]. However, these chromatin maps are insensitive to cell-to-cell variation. Rare cell subpopulations (e.g drug-tolerant tumor cells) with dictinct epigenomic signatures can not be detected and can potentially initiate drug resistance and relapse. Single-cell technologies have the potential to interrogate cellular heterogeneity at different molecular layers. Such technologies can be leveraged in epigenomic studies to generate chromatin maps at single-cell resolution and highlight epigenomic differences in the context of normal and disease cell states [Epigenomics 15].

The diagram accompanying the cover page of Chapter 4 illustrates the single-cell ChIP-seq platform established during this thesis. Starting from a cell suspension, the end result is the identification and the epigenetic characterization of the subpopulations that compose the input sample. For this purpose, the scChIP-seq platform is built around 3 main elements: a droplet-based microfluidic workflow, a Chromatin ImmunoPrecipitation workflow and an analytical workflow. All these three parts are connected to each other and evolved concurrently throughout the thesis. Nonetheless, the development of the single-cell ChIP-seq platform has raised paramount questions inherent to all single-cell technologies: What is limiting the amount of information recovered per cell? How can we optimize the system to retrieve as much information as possible per cell? Is cell-to-cell variation originating from true biological differences rather than technical artefacts?

We reasoned that most of the limiting factors and sources of variation reside in the droplet-microfluidic workflow. This Chapter 4 is centred around the building blocks of the droplet-microfluidic workflow with respect to chromatin barcoding in droplets, namely the fine-tuning of enzymatic activities and the design of DNA barcodes. Also, Chapter 4 aims to provide answers (ideally), or at least review possible lines of action, which would lead to the generation of reliable and high quality single-cell epigenetic profiles.

4.1 Introduction to the droplet-microfluidic workflow

Genome-wide mapping of histone modifications with traditional ChIP-seq method requires a large number of cells to generate high quality binding site profiles (see introductory Chapter 1, section 1.4.1 on page 14 for description of the ChIP-seq technique). Several studies have shown optimized ChIP-seq protocols to reduce input material from millions to hundreds of cells without loosing resolution in the detection of enriched or depleted regions [Adli 10, Brind'Amour 15, Ma 18]. However, these methods only yield an averaged snapshot of the modification status, without providing insight into the epigenetic heterogeneity.

Profiling histone modifications at single-cell resolution remains challenging, in part because the level of noise associated with non-specific binding during the immunoprecipitation tends to increase with low quantity of starting material. Immunoprecipitating chromatin from one single cell is technically feasible, but would lead to highly variable results.

To overcome this issue, chromatin from isolated single-cells can be "indexed" beforehand with a specific and unique DNA sequence (aka DNA barcode), and then combined with indexed chromatin from thousands of cells to perform immunoprecipitation in bulk as in a traditional ChIP-seq protocol. This method circumvents the issue associated with high experimental noise in the immunoprecipitation of low input material, while retaining the single-cell information. Indeed, barcodes being specific of one cell, each read can be attributed to its originating cell after sequencing. However, like other single-cell technologies in which molecular indexing is involved, only indexed nucleosomes (barcoded nucleosomes) have the potential to be amplified and sequenced. The single-cell coverage is directly linked to the number of indexed nucleosomes, making it clear that chromatin indexing is one central problem to be overcome.

Rotem et al employed chromatin indexing method adapted to droplet-based microfluidics to profile histone modifications of thousands of cells at single-cell resolution [Rotem 15a]. The droplet format provides a versatile tool for performing single-cell assays (see introductory Chapter 2). Briefly, cells were compartmentalized in droplets, lysed and their chromatin fragmented by micrococcal nuclease (MNase). The droplets were then merged one-to-one with a second population of droplets containing DNA barcodes, allowing chromatin indexing at the single-cell level. This system is the only one reported in the literature so far (see introductory Chapter 2, section 2.4 on page 36 for complete description of Drop-ChIP).

Limitations of Drop-ChIP [Rotem 15a]

Although Drop-ChIP was used to reveal distinct chromatin states within a population of embryonic stem cells, the single-cell information was limited to as few as hundreds of unique enriched loci detected per cell. As mentioned above, a low cell coverage may be related to a low chromatin indexing efficacy or a poor recovery of indexed nucleosomes from droplets. In addition, this study was carried out on *in-vitro* cultured cell lines, suggesting that the cell coverage and the capability to detect variation in chromatin patterns may be even lower with more complex biological samples (e.g tumor specimens). Notably, we reasoned that Drop-ChIP suffers from two major limitations, which may negatively impact the amount of information recovered per cell:

- Only symmetrically indexed nucleosomes can be amplified and are part of the sequencing library. This requirement dramatically increases the stringency of the system and imposes a strong selection on the nucleosomes (i.e. only those with both ends ligated to a barcode are amplified).
- Amplification of indexed nucleosomes relies on numerous cycles of Polymerase Chain Reaction (PCR), which increases the probablity to introduce amplification bias and errors.

Also, from a practical point of view:

- The complexity of the microfluidic workflow and the lack of real-time monitoring of droplet operations can be a source of loss of material and variability between single-cells and/or across experiments.
- The low complexity of the barcode library (only 1,152 possible barcode combinations) restricts the collection of maximum 100 cells to ensure at least 95% of the barcodes are unique to a single-cell. This issue can be mitigated by collecting multiple independent fractions and by adding a second fraction-specific barcode during the sequencing library preparation. Nevertheless, the immunoprecipitation is only performed on 100 barcoded cells (and eventually chromatin carrier) and multiplied by the number of fractions collected, again increasing the risk of loosing material and introducing bias between fractions.

Constraints on the development of a droplet-based microfluidic workflow for high-throughput chromatin indexing

For these reasons, we sought to develop an alternative droplet-based microfluidic workflow, which should ideally have the following characteristics:

1. Like Rotem et al, we reasoned that indexing chromatin with DNA barcodes in droplets and combining their content for immunoprecipitation is the easiest route to achieve single-cell resolution while limiting experimental noise.

2. Unfortunately, cells and DNA barcodes can't be loaded simultaneously in the same droplets as MNase would digest equally either nuclear DNA or barcodes. Therefore, cells and barcodes have to be encapsulated separately and later merged to inactivate MNase and allow chromatin indexing.

3. Cell emulsion (also referred to as "nucleosomes-containing droplets"): chromatin from single-cells should be efficiently released in droplets to get cleaved by microccocal nuclease. Lysis buffer composition, MNase activity and incubation time should be precisely calibrated.

4. The barcoding strategy is an essential feature of the technology. The barcode design should allow amplification of nucleosomes for which only one end is ligated to a barcode. Conveniently, the number of possible barcode combinations should be large enough to allow multiplexing of thousands of cells in a single assay.

5. Barcode emulsion (also referred to as "barcodes-containing droplets"): for practical reasons, DNA barcodes and all reagents necessary for MNase inactivation and barcode-nucleosome ligation should be present in these droplets. Doing so, the microfluidic workflow would comprise one less droplet operation as compared to Rotem et al, which would make the microfluidic device easier to control (droplet fusion in a merger device). Also, enzymatic activity (MNase inactivation and ligation efficiency) should be calibrated towards the generation of high coverage single-cell profiles.

6. Finally, all microfluidic operations should be monitored in real-time to ensure robustness and reproducibility of the experiments. In particular, the monitoring of the fusion step is important as it determines the number of cells co-encapsulated with a barcoded bead, which would later contribute to the single-cell libraries. Discrepancies in the number of different barcodes identified between the droplet count and the sequencing data would indicate elevated level of noise, or on the other hand, low overall efficacy of the system (high loss of material).

Taking into account the main features listed above and the limitations observed with Drop-ChIP, we conceived an alternative droplet-microfluidic workflow whose principles of operation are illustrated in Fig. 4.1.1 on page 55.

As imposed by ChIP assay (see features #1 & #2 listed above), droplets containing cells and droplets containing barcodes are separately produced before being re-injected and merged one-to-one in a dedicated microfluidic fusion device (Fig. 4.1.1a). One major difference is the use of an alternative barcoding strategy. We replaced soluble barcodes emulsified from microtitre plates containing oligonucleotides by hydrogel beads carrying an average of ~5×10^7 copies of the same and unique DNA sequence (see introductory Chapter 2, section 2.3.1 on page 29). These barcodes are synthesized directly on beads by combinatorial synthesis using a split-pool method from 3×96 well plates, resulting in 96^3 ~8.8×10^5 possible combinations (i.e up to 18k cells can be multiplexed in a single assay while limiting to 1% the probability of finding 2 cells with the same barcode). An important improvement over Drop-ChIP lies in the design of the barcode structure, which allows linear amplification of all barcoded nucleosomes (and not only symmetrically barcoded nucleosomes on both ends, see feature #4 listed above). The barcodes are then released from the beads within fused droplets by photocleavage for nucleosomes barcoding in droplets (Fig. 4.1.1b). Finally barcoded-nucleosomes from all droplets are combined, immunoprecipitated and the enriched fragments linearly amplified. The deconvolution of the barcodes-associated sequencing reads attributes all sequences to their originating cells, thus generating genome-wide single-cell chromatin profiles.

In addition and as suggested by feature #6 listed above, all microfluidic operations (droplets production and fusion) can be precisely monitored in real time by scanning each droplet when crossing a laser beam at a detection point (Fig. 4.1.1a). Then, the number of fused droplets containing both a cell and a barcoded-bead can be accurately counted. This information is valuable as these droplets determine the final number of different barcodes expected in the sequencing data.

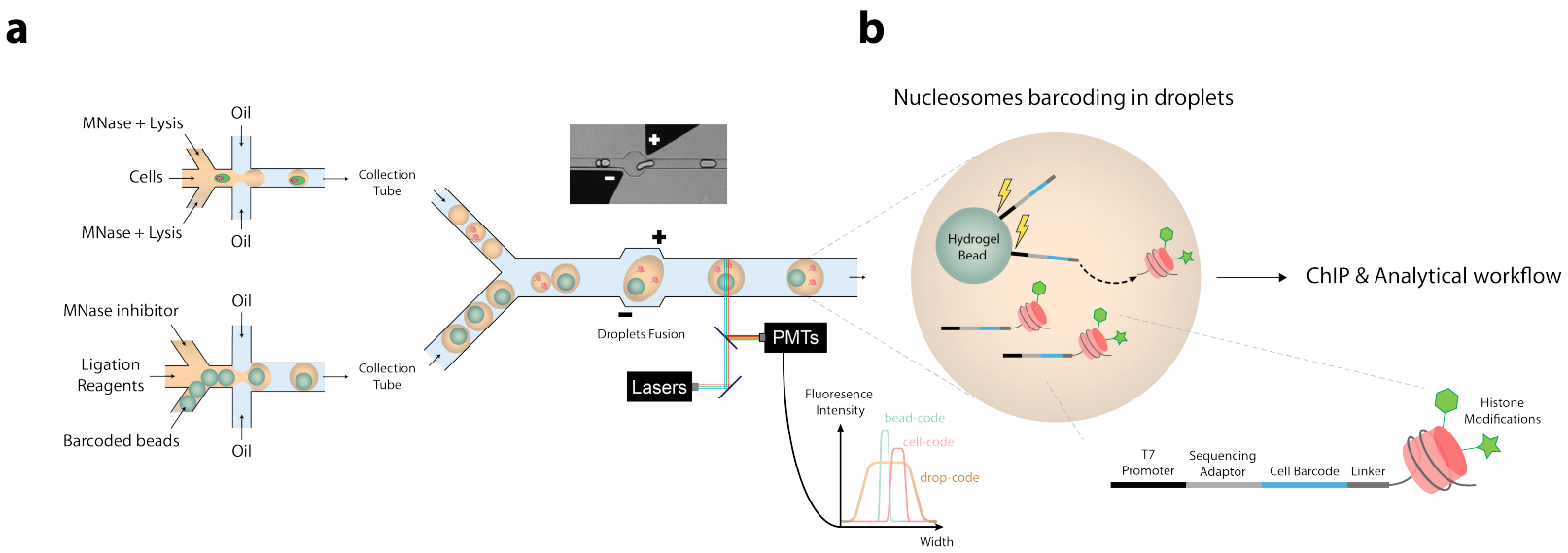


Figure 4.1.1: **Schematic illustrating the microfluidic workflow of the single-cell ChIP-seq procedure.** (a) Cells are compartmentalized in 45 pl droplets with the reagents necessary for cell lysis and chromatin fragmentation. In parallel, hydrogel beads carrying DNA barcodes are encapsulated in 100 pl droplets with ligation reagents and MNase inhibitor. The two emulsions are re-injected in a fusion device, the barcode-containing droplets (100 pl) and the nucleosome-containing droplets (45 pl) are paired asymmetrically and merged by electro-coalescence triggered by an electric field. Fused droplets are scanned one-by-one with a laser beam to analyze the composition of each droplet in real time. (b) Emulsion of fused droplets is collected and incubated off-chip for nucleosomes barcoding in droplets. Barcodes are released from the beads by photo-cleavage and ligated to the nucleosomes. After indexing, the content of droplets is combined, immunoprecipitated and the enriched DNA is sequenced [ChIP workflow]. The deconvolution of the barcode-associated reads attributes all sequences to their originating cell to generate genome-wide single-cell chromatin profiles [Analytical workflow].

4.2 Synchronizing & pausing chromatin fragmentation in droplets

Analysis of histone modifications relies on Chromatin ImmunoPrecipitation technique (ChIP), in which an antibody is used to enrich genomic regions carrying the histone modification of interest. For this purpose, chromatin is first fragmented into nucleosomes, either by enzymatic or mechanical cleavage of linker DNA (e.g. by sonication). In the second approach, the chromatin is crosslinked beforehand with formaldehyde and then sonicated (see introductory Chapter 1, section 1.4.1 on page 14 for the description of the technique).

To achieve single-cell resolution, cells are compartmentalized in picolitre volume droplets. We hypothesized that the mechanical fragmentation approach is not compatible with the droplet format as the sonication might destroy the integrity of the droplets. In addition, the crosslinking step might also disrupt the target epitope, thus reducing the immunoprecipitation efficacy. For these reasons, we opted for the enzymatic digestion of nuclear DNA by microccocal nuclease (MNase).

Cells are compartmentalized in 45 picolitre droplets with a digestion mix comprising the lysis buffer and the MNase (see schematic of the microfluidic workflow, Fig. 4.1.1 on the preceding page). After complete cell lysis, chromatin is released into the droplets and accessible to get cleaved by the MNase. This section presents a typical calibration of MNase activity in droplets in order to yield preferentially fragments of the size of a nucleosome. However, performing enzymatic assays in droplets might be challenging as cells are processed individually with different time scales. Consequently, fine-tuning enzymatic activity is necessary to avoid chromatin digestion discrepancy between droplets and single-cells.

4.2.1 Compartmentalization of cells in droplets

Monitoring cells encapsulation

The number of cells per droplet follows a Poisson distribution, which describes the probability of finding a mean number λ of x cells per droplet (see introductory Chapter 2, section 2.2.3 on page 29). In single-cell ChIP-seq experiment, we adjusted cell density to encapsulate $\lambda = 0.1$ cells in 45 pl droplets, resulting in 90.5% of empty droplets, 9% of droplets containing one single cell, 0.5% containing two cells and 0.015% containing more than two cells.

To monitor in real-time the compartmentalization of cells in droplets, cells were

pre-labeled with Calcein AM, a cell-permeant dye usually used for testing cellular viability. Calcein AM is the non-fluorescent derivative of calcein but the -AM group (AcetoMethoxy group) gets cleaved by intracellular esterases of living cells, releasing a strong green fluorescence (excitation/emission: 495/515nm). The fluorescence is acquired as the droplets crossed the laser beam at the detection point, enabling the counting of the number of cells encapsulated (Fig. 4.2.1).

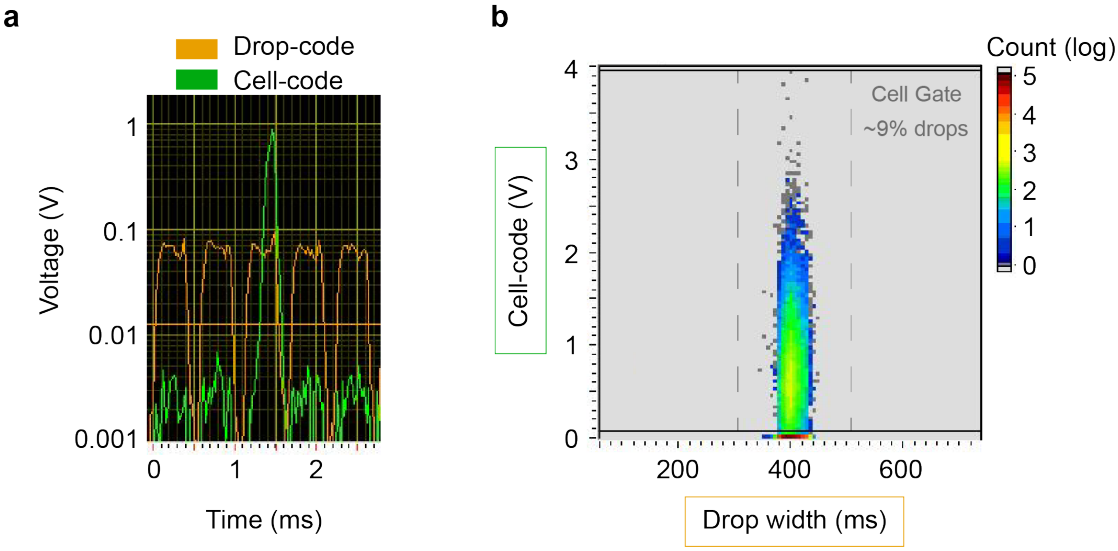


Figure 4.2.1: **Monitoring cell encapsulation in 45 pl droplets.**
(a) Experimental time trace recorded for droplets analyzed at 1.8 kHz. Orange fluorescence is present in all droplets and used to control for the size of the drops (drop-code). Green fluorescence indicates the presence of a cell in a droplet (cell-code).
(b) Plot of cell-code intensity (green) vs drop-code intensity (orange) in each droplet. Droplets containing a cell have a high green cell-code fluorescence allowing the counting of encapsulated cells via the definition of a cell-gate above the noise level. A cell density adjusted to $\lambda = 0.1$ results in ~9% of the droplets that contain one single cell.

Droplets collection

Droplets were collected in a collection tube on ice until the end of the encapsulation, typically 10 min to 20 min depending on the number of starting cells. After encapsulation, droplets were incubated off-chip at 37°C for MNase digestion.

4.2.2 MNase calibration in droplets

At the end of the encapsulation, droplets were incubated off-chip for single-cell chromatin fragmentation. Cells were lysed in droplets, making their nuclear DNA accessible for MNase enzyme. The kinetic of the digestion is particularly important to yield preferentially mono-nucleosomes, which are then retained in the droplet. The ideal incubation time is defined as the time necessary to get 100% of nuclear DNA fragmented into mono-nucleosomes.

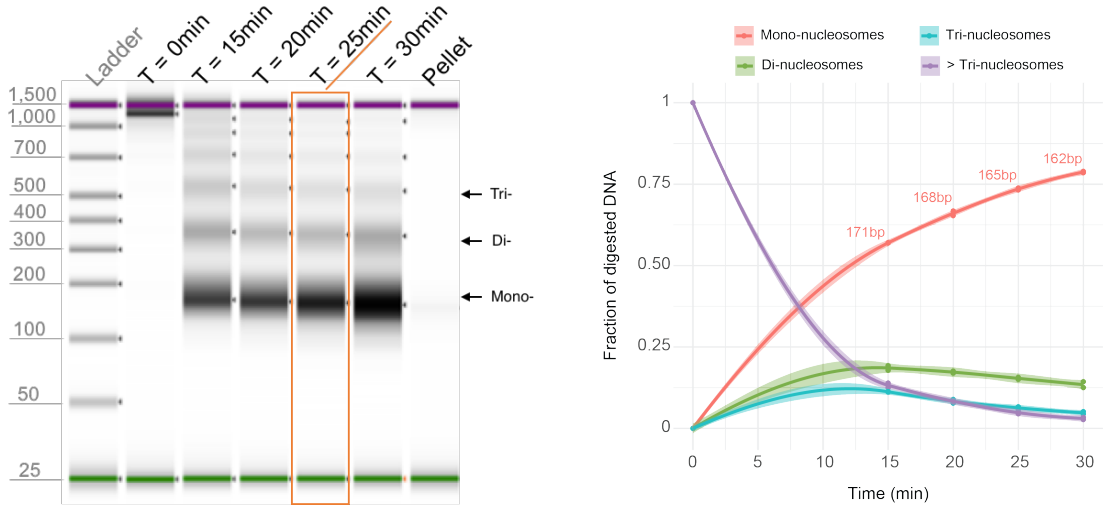
The digestion conditions including lysis buffer composition, MNase concentration and incubation time were precisely calibrated beforehand for each sample by performing a time-course study. The calibration was carried out as follows: 45 pl droplets containing cells, lysis buffer and MNase were produced, collected in a collection tube and placed at 37°C for different incubation time. At each time point, a fraction of droplets were broken and MNase immediatly inactivated by addition of EGTA. DNA fragments were then purified and analyzed by electrophoresis (an example of a time-course MNase fragmentation in droplets for human Jurkat T-lymphocyte cell line is shown in Fig. 4.2.2).

Controlling for proper cell lysis

For effective MNase digestion and to limit loss of information, chromatin has to be efficiently released inside the droplets from the nucleus. After breaking the emulsion, DNA fragments in the soluble fraction were separated by centrifugation from the insoluble fraction composed in part of cellular debris and sometimes insoluble pelleted chromatin with tightly associated proteins. The pellet was recovered, purified and processed in parallel of the soluble fraction. If a non-negligible proportion of DNA remains in the pellet rather than in the soluble fraction, it might be necessary to adjust the cell lysis buffer composition. This is particularly true for heterochromatic histone modification studies such as H3K27me3 or H3K9me3. Here, electrophoresis of DNA fragments reveals minute amount of material remaining in the pellet (~1%), suggesting appropriate lysis buffer composition (see lane "Pellet" on Fig. 4.2.2a).

Nucleosome ladder

The calibration of MNase fragmentation in droplets for Jurkat cells (human T-lymphocyte cell line) is shown in Fig. 4.2.2a. From 15 min to 30 min of incubation, the bands on the electropherogram show a classical MNase digestion profile (sometimes referred as "nucleosome ladder") with expected size of mono-, di-, tri-nucleosomes... Intuitively, the proportion of mono-nucleosomes increases over time as longer DNA fragments get cleaved.



(a) Time-course MNase digestion in droplets

(b) Fraction of nucleosomal fragments at each time point

Figure 4.2.2: **Calibration of MNase activity in droplets.**
(a) Electrophoresis of DNA fragments after MNase fragmentation in droplets shown at different time point of incubation. First lane is the DNA ladder. The proportion of mono-nucleosomes (band at ~160 bp) increases with time as longer DNA fragments get cleaved. The lane "Pellet" shows that cells are properly lysed and no material will be lost at this stage. t = 25min is selected as incubation time for this particular cell type.
(b) Chromatin digestion by MNase generates a mix of mono-, di- and tri-nucleosomes. The fraction of each nucleosomal fragments is calculated for each time point based on electrophoresis profiles as shown in panel (a). Data points are duplicates and trend lines are plotted with a 95% confidence interval. Average size of mono-nucleosomes is also indicated above corresponding points. At t = 25min, ~75% of the nuclear DNA is fragmented into mono-nucleosomes with a mean size of 165 bp. Long DNA fragments (higher than tri-nucleosomes) represent less than 5% of the original DNA and originate mostly from tightly packed chromatin.

The proportion of each nucleosomal fragments is extracted from electropherograms for each time point and plotted in Fig. 4.2.2b. After 25 min of incubation, chromatin had been digested in mainly mono-nucleosomes (~75% with a mean size of 165 bp) and di-, tri-nucleosomes (15% and 6% with a mean size of 352 bp and 534 bp respectively). A small fraction (< 5%) of original DNA remains longer than trinucleosomes and can be associated with tightly packed chromatin.

The choice of the incubation time is a balance between having the highest proportion of mono-nucleosomes but, in the same time, preventing nucleosomal DNA to be overdigested. Indeed, we hypothesized that DNA protruding from the nucleosome should be long enough to enable an efficient ligation of the barcodes in the subsequent steps of the procedure (hypothesis not confirmed experimentally).

4.2.3 Controlling MNase activity in droplets

Performing enzymatic assays on individual cells in droplets is challenging as cells are processed sequentially with different time scales. For example, the duration of encapsulation is about 20 min, in other words, in the same order of magnitude as MNase incubation time (see MNase calibration subsection 4.2.2 on page 58). Intuitively, cells encapsulated at the beginning will be longer in contact with MNase than the cells encapsulated at the end of the droplets production. Similar observation can be made regarding the re-injection of the droplets in the fusion device (see schematic of the microfluidic workflow, Fig. 4.1.1 on page 55). Indeed, the fusion of the two emulsions can last between 1h to 4h depending on the design of the experiment, meaning that some droplets containing fragmented DNA "wait" for hours before being fused and the MNase inactivated by EGTA. Consequently, synchronizing and pausing enzymatic activity are critical in order to avoid the introduction of chromatin digestion variation between individual cells.

We hypothesized that collecting the droplets on ice upon cells encapsulation may prevent MNase activity. To confirm this hypothesis, we performed a similar digestion control experiment in droplets as detailled in the MNase calibration subsection on page 58.

We validated our hypothesis that collecting the droplets on ice actually prevented MNase activity. The time point "t = 0 min" on Fig. 4.2.3, which corresponds to a fraction of droplets taken at the end of the droplets production but just before incubation, indicates that nuclear DNA were not digested yet by MNase. Therefore, we confirmed that chromatin digestion was not occuring upon droplets production but started immediately with incubation at 37°C (see time point "t = 5 min" on Fig. 4.2.3).

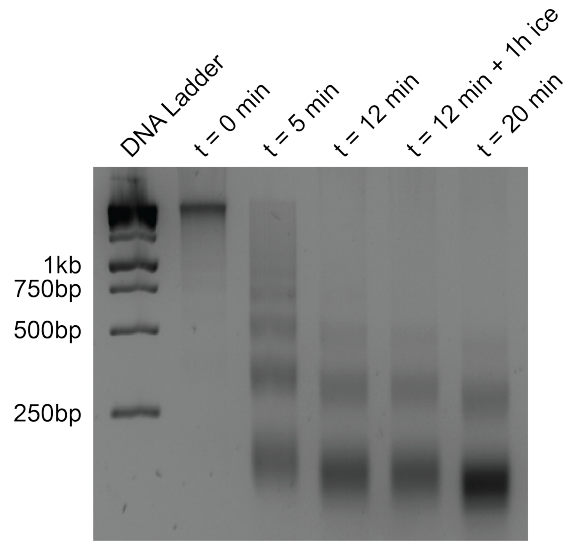


Figure 4.2.3: **Synchronizing & pausing MNase activity between droplets.** Gel electrophoresis of DNA fragments at different time point of MNase incubation. At t = 0 min, DNA is not fragmented yet, confirming that MNase activity is synchronized upon droplets collection on ice. Time point t = 12 min + 1h ice shows similar digestion profile as after 12 min of incubation, confirming that MNase activity is paused when droplets are stored on ice.

In conventional bulk ChIP-seq assays, the inactivation of MNase occurs immediately after MNase incubation with the addition of EGTA. Conversely, in single-cell ChIP-seq assays, EGTA can't be added immediately within the droplets and MNase is only inactivated after fusion with the barcodes-containing droplets (also containing EGTA, see schematic of the microfluidic workflow, Fig. 4.1.1 on page 55). Also, we hypothesized that placing the droplets on ice after incubation and upon re-injection in the fusion device may "pause" MNase activity and limit cell-to-cell variation in chromatin digestion.

For this purpose, two droplet fractions were taken after 12 min of MNase incubation: one fraction was immediately processed to control for digestion, while the second fraction was stored beforehand for 1h on ice, then processed similarly. Both time points show similar fragmentation profiles, confirming that the MNase is no longer active in the fraction stored on ice (as compared with "t = 20 min" time point, see Fig. 4.2.3). Consequently, storing droplets on ice "pauses" MNase activity, thus preventing again variation in chromatin digestion between droplets[1].

[1]Control also confirmed with a 4h incubation on ice (data not shown).

4.3 DNA barcoding strategy for efficient chromatin indexing at single-cell resolution

Barcoding refers to the need of analyzing multiple cells at once. Single-cell genomic, transcriptomic or epigenomic technologies require a way to index molecules of interest in order to first drives amplification of the target molecules and to be later used as a "barcode" sequence specific to the cell of origin. One possible solution is to use multiple copies of a unique nucleic acid sequence to index DNA or RNA molecules from individual cells with the same sequence. Bioinformatic analysis of the sequencing data and the deconvolution of the barcode-associated reads attributes each sequence to one barcode (i.e to one cell). Recent improvements in single-cell RNA-seq technology have stimulated the development of such barcoding methods in droplets allowing analysis of thousands of cells in parallel [Macosko 15, Klein 15, Zilionis 17]. Such methods can be adapted to single-cell epigenomic studies to improve the throughput and efficiency of the methods (see introductory Chapter 2, section 2.3.1 on page 29 for complete description of droplets barcoding).

In our scChIP-seq platform, chromatin from single-cells are indexed in droplets before being combined for the immunoprecipitation and subsequent sequencing library preparation steps. Section 4.1 on page 51 of this Chapter already introduced the importance of the chromatin indexing on the overall efficacy of the technology, in which DNA barcodes are a core component. The following parts of this section outline the barcoding strategy utilized as well as the optimization made on the barcode structure to ensure efficient and reliable chromatin indexing in droplets.

4.3.1 Introduction to the DNA barcoding strategy

In Drop-ChIP, Rotem et al took advantage of the droplet-based libraries of oligonucleotide barcodes previously reported by Rotem himself for single-cell RNA-seq purposes [Rotem 15b]. Briefly, oligonucleotides were directly emulsified from microtitre well plates and collected through a single output. One advantage of the method was the presence of soluble barcodes at high concentration in droplets (up to 10^9 molecules per droplet). However, the complexity of the microfluidic workflow used to generate the droplets and the low diversity in the total number of possible barcode combinations hamper the use of this method for high-throughput single-cell chromatin barcoding (see Limitations of Drop-ChIP, section 4.1 on page 51 of this Chapter).

To overcome Drop-ChIP limitations, we centred our chromatin indexing strat-

egy on DNA barcodes bound to hydrogel beads as later reported by Klein et al for single-cell RNA-seq [Klein 15, Zilionis 17]. Highly monodisperse hydrogel beads composed of a polymerized network of Streptavidin and PolyEthylene Glycol Di-Acrylate (PEG-DA) are produced in a microfluidic device. Hydrogel beads are porous and deformable beads of 18 pl in volume (33 µm in diameter), which are used as solid supports for DNA barcodes synthesis by combinatorial synthesis using a split-pool method.

DNA barcodes are grafted to the hydrogel beads via a streptavidin-biotin linkage and a photo-cleavable moiety enabling release from the beads after exposure to UV light. The synthesis of the barcodes consists in distributing the beads in a microtitre well plate containing ligation reagents and 96 combinations of a 20 bp oligonucleotide (later referred to as Index 1). Index 1 are ligated to the beads and the latter are pooled before being distributed again in a second microtitre plate containing 96 new combinations of a 20 bp oligo (later referred to as Index 2). By repeating 3 times this split-pool cycle, a library of 96^3 possible barcode combinations is easily generated (i.e 884,736 combinations). The production of the hydrogel beads and the synthesis of the barcodes on beads are further detailled in Appendix A on page 141.

Quality controls of barcoded hydrogel beads

Barcoded beads are one of the core reagents of the scChIP-seq technology, their quality has been systematically controlled to ensure that cell-to-cell variations originate from true biological differences in their histone modification patterns rather than technical artefacts.

We first controlled that the majority of the barcodes on the beads were complete after synthesis. Electrophoresis profile of DNA barcodes released from the beads revealed that >75% were full length (larger peak at 146 bp), as well as the presence of intermediates that failed to be completed (Fig. 4.3.1a). In average, the number of full-length barcodes was estimated to $\sim 5 \times 10^7$ copies per barcoded hydrogel bead.

As the recovery of barcoded-nucleosomes is directly linked to the final cell coverage, the barcodes must be efficiently released from the hydrogel beads. In addition, Klein et al showed that the efficiency of droplet-based scRNA-seq procedure was enhanced when barcoded primers were released within the droplets before the capture of mRNAs and the reverse transcription reaction [Klein 15]. Similarly, we hypothesized that chromatin indexing would also benefit from early release of

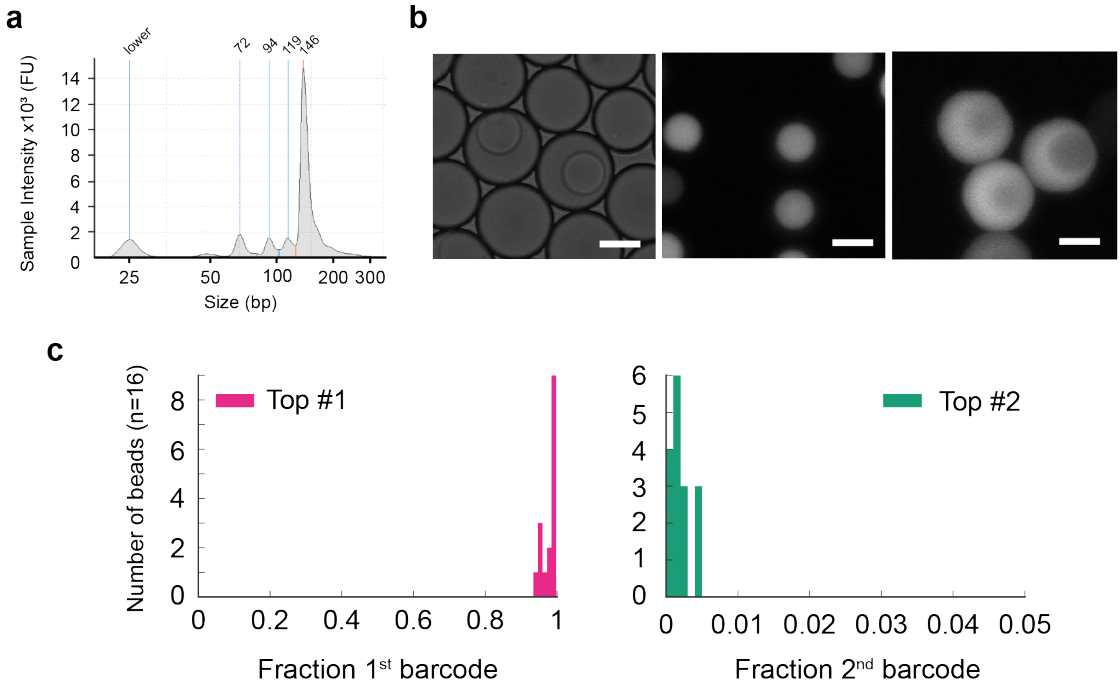


Figure 4.3.1: **Quality controls of barcoded hydrogel beads.**
(a) Tapestation profile of DNA barcodes after photo-cleavage from hydrogel beads showing the presence of full-length barcodes (larger peak at 146 bp), as well as intermediates that failed to be completed (peaks at 72 bp, 94 bp and 119 bp).
(b) Imaging of hydrogel beads in droplets using epifluorescence microscopy. From left to right: (i) bright field image; (ii) imaging after hybridization of a DNA probe complementary to the Illumina sequencing adaptor onto the barcodes; (iii) same as (ii) after release of the barcodes in droplets by photo-cleavage. Scale bars are 35 µm.
(c) Single-bead deep sequencing results showing the fraction of the first-two most abundant barcodes of 16 beads. In average, 97.7% of the barcodes present on a bead match to the same sequence while the second most abundant barcode represents only 0.17% of all sequencing reads.

DNA barcodes before the incubation for ligation reaction rather than keeping the barcodes bound on the beads. To validate the release of the barcodes from the hydrogel beads, complementary DNA probes were hybridized to the barcodes onto the beads. The latter were then encapsulated in 100 pl droplets and collected off-chip as in a scChIP-seq experiment. We re-injected part of the droplets as a single file into a micrometric chamber as reported by Eyer et al [Eyer 17] and imaged the beads by epifluorescence microscopy. As expected, the fluorescence was localized on the beads (see second image on Fig. 4.3.1b). A second fraction of beads-containing droplets were exposed to UV light to initiate barcodes release and loaded into the micrometric chamber as described above. Epifluorescence microscopy of bead-containing droplets after photo-cleavage revealed a uniform distribution of the fluorescence in the droplets, suggesting complete barcode release (see third image on Fig. 4.3.1b). We didn't quantify the kinetic of release into the droplets but we visually confirmed the speed of diffusion as only few minutes elapsed between UV exposure and imaging.

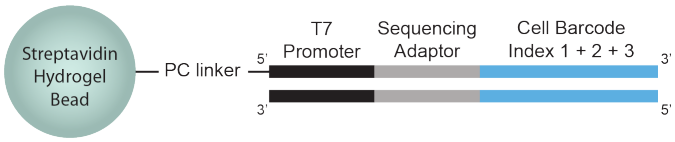
Lastly, we performed single-bead sequencing on every new batch of barcoded-beads produced. Single-beads were isolated by limiting dilution in a 384-well plate. Only wells containing one bead were selected by imaging for amplification and sequencing of the barcodes. Hundred thousands of different barcodes were identified per bead but, in average, the most abundant barcode represented 97.7% of the sequencing reads. The second most abundant barcode took on average as few as 0.17% of the reads, suggesting that all the other barcodes were negligeable (see Fig. 4.3.1c).

4.3.2 Precise DNA barcode design improved chromatin indexing in droplets

In Drop-ChIP, the structure of the barcodes was probably one of the main limitation in the final single-cell coverage (hundreds of unique loci detected per cell). Indeed, their structure imposed a strong selection as only symmetrically indexed nucleosomes on both ends had the potential to be amplified by PCR (see Limitations of Drop-ChIP, section 4.1 on page 51 of this Chapter). We reasoned that indexing nucleosomes from only one end would increase the single-cell coverage and ultimately the capacity of the system to distinguish more subtle variations between single-cell chromatin profiles.

As previously mentioned, barcodes are bound to the beads via streptavidin-biotin linkages, which are in turn separated from the 5'-end of the oligonucleotide by photo-cleavable entities. The latters are composed of the photo-cleavable group

as well as an alkyl spacer that minimize steric interactions (the all entity is referred to as PC-linker). The first biotinylated and PC-linker oligonucleotide is common to all barcodes and comprises a T7 promoter sequence and an Illumina sequencing adaptor. The T7 promoter sequence serves as a recognition site for the T7 RNA polymerase to initiate the linear amplification of enriched barcoded-nucleosomes post-immunoprecipitation in an *in-vitro* transcription reaction (IVT). This amplification strategy is widely adopted for unbiased, sensitive and reproducible amplification of cDNAs post-reverse transcription in single-cell RNA-seq protocols [Hashimshony 12]. In a second step, the Illumina sequencing adaptor serves as PCR handle to complete the preparation of the sequencing library. Also, this adaptor is necessary for the Next Generation Sequencing of the samples as the primer that initiates the reading of the barcode sequence (see Appendix B, section B.2 on page 150). The beads grafted with this first common oligonucleotide are then used for barcode synthesis with successive ligation of the 3 index as previously described and reported in Appendix A on page 141. A schematic representation of the structure is illustrated in Fig. 4.3.2a.



(a) Scheme of barcode structure v1

	Barcode structure v1
Correct barcodes	**37.71%**
Only index 1	18.08%
Only index 1-2	24.31%
Barcode in Read #1	8.7%
Other	11.2%

(b) Proportion of correct and rejected barcode reads

Figure 4.3.2: **The original barcode design yields low proportion of reads with a correct structure.**
(a) The first version of the barcode comprises the T7 promoter, the Illumina Read #2 sequencing primer and the 3 index barcode.
(b) Only 37.7% of the reads were identified with a correct barcode structure. Analysis of rejected reads revealed that 42.4% of the raw reads were associated with barcodes missing 1 or 2 index.

Unfortunately, the analysis of one of our first single-cell ChIP-seq dataset revealed that only few reads (~38%) were having a complete and correct barcode structure (see Fig. 4.3.2b). We investigated for common patterns among rejected reads and found that:

- 42.4% of raw reads (68% of rejected reads) had a non-complete barcode structure missing the 2nd or the 3rd index. Among those, 18.08% had only the first index and 24.31% had the first two index. Barcodes missing all three index can't be rigorously identified but they may be part of the "Other" rejection flag on Fig. 4.3.2b.

- 8.7% of raw reads (14% of rejected reads) had a barcode with only the first index identified in Read #1 instead of Read #2, indicating the presence of barcode concatemers.

The high proportion of reads within each rejection flags suggested that we didn't anticipate 2 major issues in the first version of the barcode structure:

1. The barcodes are bound to the beads via a streptavidin-biotin linkage and the biotin is separated from the 5'-end of the oligo by a photo-cleavable linker. Under exposure to UV light, this PC-linker is cleaved, releasing a 5'-phosphorylated oligonucleotide:
   a) The retainment of the 5'-phosphate makes the barcode suitable for self-ligation and formation of concatemers.
   b) As our barcode design is not symmetric, nucleosomes ligated to the released 5' side of the barcodes won't be amplified.

2. As revealed by the Tapestation profile of Fig. 4.3.1a, all the barcodes on the beads are not complete (75% are full-length barcodes). Obviously, reads associated with non-full length barcodes can't be attributed to their originating cells, introducing experimental noise and higher proportion of unusable reads after sequencing.

Taking into account previous results, we proposed an optimized structure allowing the digestion of barcode concatemers as well as reducing the ligation of non-full length barcodes. A schematic representation of this optimized structure is illustrated in Fig. 4.3.3a.

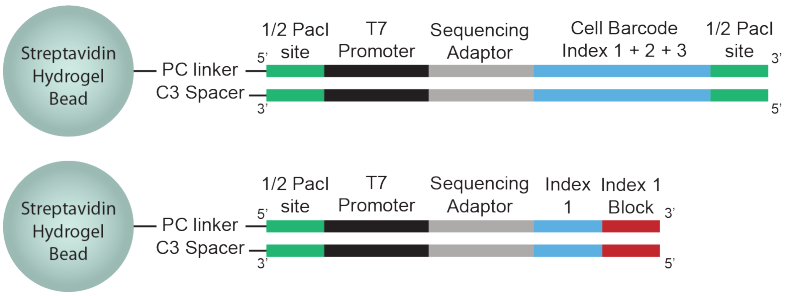
1. Barcodes are framed with one half of Pac1 restriction site, which is only reconstructed in case of formation of concatemers. Those are digested after the immunoprecipitation but before the linear amplification to clean-up the library.

2. Barcodes photo-cleaved side is modified with the incorporation of a 3' C3-spacer. This modification introduces a spacer arm at the 3' hydroxyl group of the 3' base and blocks the ligation. With the addition of the spacer group, the direction of the ligation is enforced from the 3' end of the barcodes to the nucleosomes.

3. Non full-length barcodes are completed with a "block" oligonucleotide sequence comprising a 3' C3-spacer and a 5' inverted dideoxy-T base. Again, both modifications aim at limiting unwanted ligation events.

These optimizations improved the proportion of reads with a correct barcode structure (see Fig. 4.3.3b). A quarter of the reads are still associated with non-full length barcodes suggesting that some of them are not blocked. It is not surprising as the "block" sequence is added by ligation, which is also limited by the efficiency of the ligation. As an alternative, we have also considered digesting the non-full length barcodes rather than blocking them. This other possibility relied on the modification of the last adaptor with the introduction of two phosphorothioate bonds in Pac1 half restriction site. Phosphorothioate bonds subsitute sulfur atoms for oxygen in the phosphate backbone of an oligo and are generally introduced to inhibit exonuclease degradation. In our scenario, only full-length barcodes would be protected from exonuclease treatment. We didn't continue with this solution as the first experimental tests were not convincing and we were concerned that the efficiency of the digestion of barcode concatemers by Pac1 would have been impacted by the presence of phosphorothioate bonds in the backbone of Pac1 restriction site[2].

[2]Including phosphorothioate bonds in the oligo reduces activity of nucleases [Putney 81]



(a) Scheme of barcode structure v2

	Barcode structure v1	Barcode structure v2
Correct barcodes	**37.71%**	**73.57%**
Only index 1	18.08%	7.49%
Only index 1-2	24.31%	10.08%
Barcode in Read #1	8.7%	5.4%
Other	11.2%	3.46%

(b) Proportion of correct and rejected barcode reads

Figure 4.3.3: **Optimizing the barcode structure greatly enhanced the quality of the sequencing data.**
(a) The new structure enables digestion of barcode concatemers by addition of one half of Pac1 restriction site on both ends of the barcodes. A C3-spacer at the 3' end of the bottom strand is also added to enforce the direction of the ligation. Non-full length barcodes are completed with a "block" adaptor containing modified bases modifications that prevent unwanted ligation.
(b) The v2 structure greatly improved the quality of the sequencing data. The proportion of correct barcodes is doubled between the first version and the second version of the barcodes. Similarly, the proportion of non-full length barcodes is reduced by a factor 2.5.

4.4 Nucleosomes barcoding in droplets

In all single-cell technologies, the ability to identify and characterize subpopulations relies on the final number of sequencing reads obtained per cell (also referred to as "cell coverage"). In Drop-ChIP, Rotem et al obtained as few as hundreds of reads per single-cell but they demonstrated that this coverage was still sufficient to identify distinct chromatin profiles from *in-vitro* cultured cell lines using a supervised clustering approach (see introductory Chapter 2, section 2.4 on page 36 for Drop-ChIP description) [Rotem 15a]. However, such coverage is a hurdle to reliably identify subpopulations from tumor samples. As discussed in the introductory section 4.1 on page 51 of this Chapter, chromatin indexing in droplets is one of the limiting step of the scChIP-seq procedure. Indeed, the efficacy of the barcoding is directly linked to the number of reads per cell obtained after sequencing as only barcoded-nucleosomes are amplified in the preparation of the sequencing library.

As imposed by ChIP assays, cells can't be encapsulated with DNA barcodes at the first place due to the risk of barcodes digestion by MNase. Chromatin indexing is only taking place after the delivery of DNA barcodes into nucleosomes-containing droplets. This section presents a typical droplets fusion monitoring in order to precisely count the number of cells co-encapsulated with a barcoded bead. This is an excellent proxy to estimate the overall performances of the platform by comparison with the number of detected barcodes in the sequencing data. Again, important questions related to enzymatic activity in droplets arise and are discussed in the following parts of this section: is MNase fully inactivated upon droplets fusion to prevent barcodes digestion? Can we estimate the efficacy of the ligation of the barcodes to the nucleosomes as a probable cause of the final number of loci detected per cell?

4.4.1 Delivering DNA barcodes into nucleosomes-containing droplets

Encapsulation of barcoded beads in droplets

Loading discrete objects such as hydrogel beads into droplets can be estimated by a Poisson distribution. As discussed in the introductory Chapter 2 section 2.2.3 on page 29, the random loading of beads would yield a majority of empty droplets that would be useless if later fused with a nucleosomes-containing droplet.

The best-case scenario would be to "beat the Poisson statistics" by loading hydrogel beads one-by-one into the droplets. This can be achieved by taking ad-

vantage of the physical properties of the beads. Indeed, they are highly deformable and they can be squeezed in a close-packed organization without clogging the microfluidic device. The synchronization of the periodic flow of hydrogel beads with the frequency of the droplet production results in a regular loading of the beads into the droplets (see How delivering barcodes in droplets, introductory Chapter 2, section 2.3.1 on page 31) [Abate 09b].

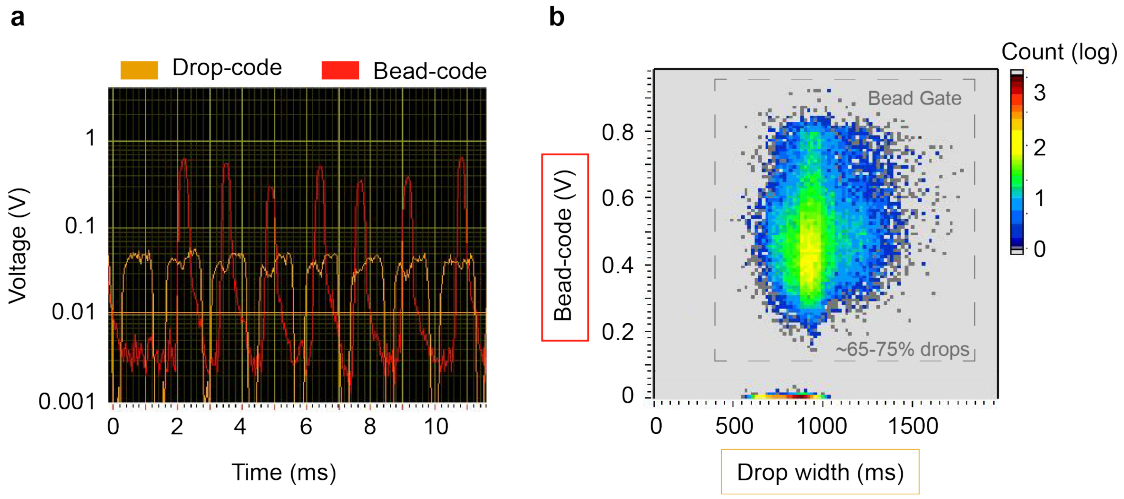


Figure 4.4.1: **Monitoring hydrogel beads loading in 100 pl droplets.** (a) Experimental time trace recorded for 100 pl droplets analyzed at 650 Hz. Orange fluorescence is present in all droplets and used as a drop-code to control for the size of the droplets. Red fluorescence indicates the presence of a hydrogel bead in a droplet (bead-code). (b) Plot of bead-code intensity (red) versus drop-code intensity (orange) in each droplet. By using a close-packed ordering of the beads, 65% to 75% of the droplets contain one bead.

In the same way as the encapsulation of the cells, the loading of the beads is monitored in real time as the droplets crossed the laser beam at the detection point. The beads are pre-labeled with biotin coupled to a Cy5 dye (excitation/emission: 650/670 nm) that bind free streptavidin sites available on the beads. The fluorescence intensity allows counting the number of droplets containing at least one bead (see Fig. 4.4.1). In single-cell ChIP-seq experiment, we typically achieved between 65% to 75% of droplets that contained a barcoded hydrogel bead.

Merging nucleosomes-containing droplets with barcodes-containing droplets

Cells and DNA barcodes are separately encapsulated to prevent barcodes digestion by MNase. To index chromatin at the single-cell level, DNA barcodes have to be delivered in a second microfluidic step into nucleosomes-containing droplets. This is achieved by active fusion of the two droplets populations in a dedicated microfluidic fusion device using a triggered electric field.

Droplets from the "cell emulsion" and droplets from the "barcode emulsion" are re-injected as a single-file in the microfluidic fusion device. Achieving proper electro-coalescence requires one-to-one pairing of the droplets from the two emulsions. Hydrodynamic forces enable the faster smaller 45 pl droplets ("cell emulsion") to catch up and come in contact with the 100 pl droplets ("barcode emulsion"), as contact is necessary for the two droplets to fuse [Mazutis 09b]. Similarly to droplet production, fluorescence intensity of fused droplets is acquired as they crossed the laser beam at the detection point (see schematic of the microfluidic workflow, Fig. 4.1.1 on page 55 of this Chapter).

Fig. 4.4.2a depicts an experimental time trace of possible fusion events. For clarity, droplets are numbered from the left to the right:

- droplet #1 contained a cell as revealed by the high "cell-code" green fluorescence. Unfortunately, no bead was present in this droplet, in other word, the chromatin profile of this particular cell is lost.
- droplets #2 and #5 both contained a bead (high "bead-code" red fluorescence) but no cell (low "cell-code" green fluorescence).
- droplet #3 was an empty droplet which did not contain any bead nor cell.
- droplet #4 contained both a bead and a cell as revealed by the high "bead-code" red fluorescence and the high "cell-code" green fluorescence. Chromatin from this cell was indexed and part of the sequencing library.

In addition to the common "drop-code" (orange) and the "cell-code" (green), droplets from the "cell emulsion" also contain blue fluorescence which is specific for this emulsion (referred to as "drop-code cell" in Fig. 4.4.2). By plotting the "drop-code cell" signal intensity (blue) versus "drop-code" signal intensity (orange) in each droplet, four main droplets populations are highlighted, allowing to determine the overall efficacy of the fusion (populations labeled (A), (B), (C) and (D) on Fig. 4.4.2b):

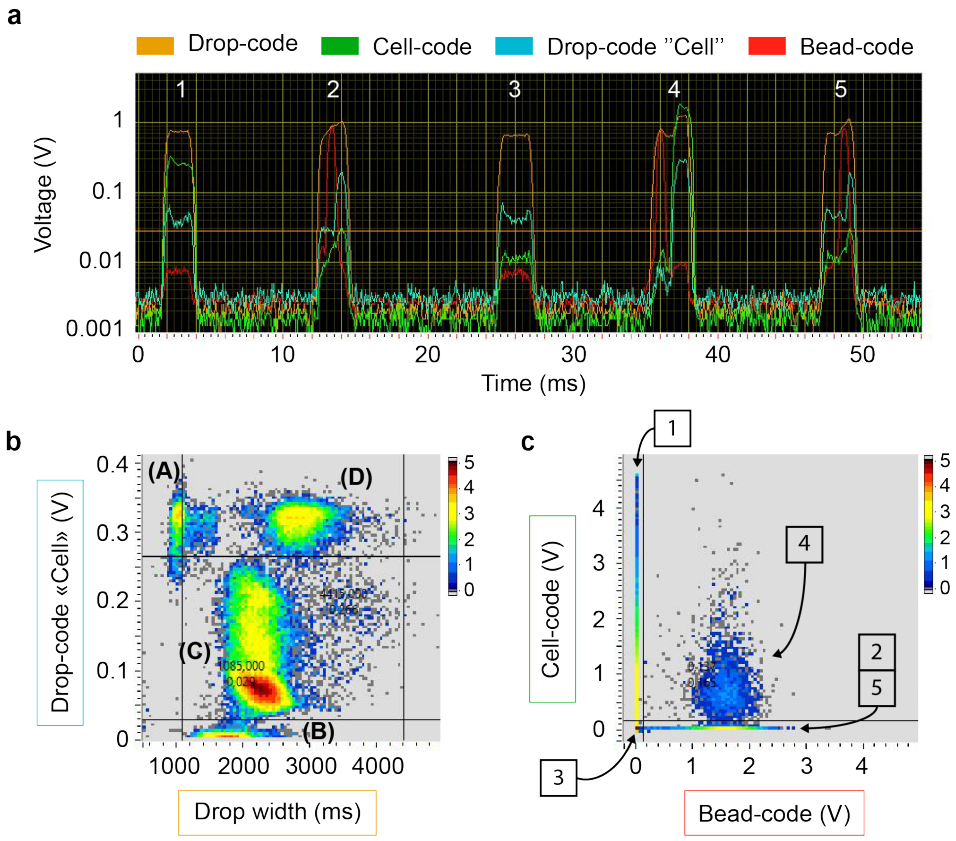


Figure 4.4.2: **Monitoring droplets fusion.**
(a) Experimental time trace recorded for droplets after fusion at 150 Hz. Orange fluorescence is present in all droplets and used as a drop-code to control for the size of the fused drops. Green fluorescence indicates the presence of a cell and the red fluorescence indicates the presence of a hydrogel bead. Blue fluorescence is a drop-code specific of the cell emulsion.
(b) Plot of drop-code "cell" intensity versus drop-code intensity in each droplet defining 4 main populations of droplets after fusion. The main population in the middle represents droplets correctly paired and fused (70% to 80%). Unpaired droplets from the bead-emulsion are in the bottom population and unpaired droplets from the cell-emulsion with a high blue-fluorescence intensity are in the top-left population. The last population (top right) is associated with incorrectly-paired droplets containing 2 cell-droplets fused with one bead-droplet.
(c) Plotting cell-code intensity versus bead-code intensity in each droplet enables precise counting of useable drops (those containing one cell and one bead). Droplets from the time trace in panel (a) are indicated as example of the different populations.

(A) Unpaired droplets from the "cell emulsion". Smaller size compared to other population and higher blue signal intensity ("drop-code cell"). Typically, this population represented 5% to 10% of the droplets after fusion.

(B) Unpaired droplets from the "barcode emulsion". These droplets have a background level of blue intensity ("drop-code cell"). Typically, this population represented 5% to 10% of the droplets after fusion.

(C) Correctly paired droplets. Typically, this population represented 70% to 80% of the fused droplets, which defined the efficacy of the fusion.

(D) Incorrectly paired and fused droplets. The high droplet width and blue fluorescence intensity suggest that one droplet from the "barcode emulsion" was fused with 2 droplets from the "cell emulsion". Typically, this population represented 5% to 10% of the fused droplets and was more likely to increase when the pairing rate was dropping.

Scanning each droplet enables to precisely count the number of cells that are co-encapsulated with a barcoded bead. This information is valuable as only these cells are indexed and amplified in the subsequent sequencing library preparation procedure. Differences in the number of barcodes identified in the sequencing data and the number of droplets counted on the microfluidic station would indicate elevated level of noise or low overall efficacy of the system (high loss of material).

Secondly, by plotting the "cell-code" signal intensity (green) versus the "bead-code" signal intensity (red) for each droplet, we can identify 4 main populations (Fig. 4.4.2c). The droplets numbered from 1 to 5 in Fig. 4.4.2a are indicated as example of each subpopulation of droplets.

- Droplets alongside the x-axis only contained barcoded hydrogel beads (e.g droplet #2 & droplet #5).
- Droplets alongside the y-axis contained a cell but no barcoded hydrogel bead (e.g droplet #1). These cells were "missed", meaning that their nucleosomes were noy indexed nor amplified in the sequencing libary preparation steps.
- Droplets without fluorescence in green and red were empty droplets (e.g droplet #3).
- Droplets with positive signal in both green and red contained a cell and a bead (e.g droplet #4). The number of droplets counted in this area corresponds to the expected number of barcodes that we should obtain after sequencing.

In terms of performance, 40% ± 5% of input cells are co-encapsulated with a barcoded hydrogel beads and contribute to the final sequencing library. The remaining

proportion of input cells constitutes "missed cells" that failed to be co-encapsulated with a barcoded hydrogel beads. Two factors can explain the high proportion of "missed cells": the loading of the barcoded beads in 100 pl droplets (70% ± 5% of the droplets contain a barcoded bead) and the fusion efficacy (75% ± 5% of the droplets are correctly fused). By further optimizing these two steps of the droplet-microfluidic workflow, we assume that the performance of the system in terms of proportion of cells interrogated would be enhanced.

4.4.2 Inactivating MNase in droplets

Micrococcal nuclease is an endo-exonuclease that can digest single-stranded, double-stranded, circular and linear nucleic acids. As MNase activity is strictly dependent on Ca^{2+}, the enzyme is easily inactivated by addition of tri-Ethylene Glycol diamine Tetraacetic Acid (EGTA) that quench the calcium ions.

In conventional bulk ChIP-seq assays, MNase inactivation occurs immediately after incubation time is over by addition of EGTA. Conversely, in the single-cell ChIP-seq assays, MNase is only inactivated after fusion with the "barcode emulsion", in which EGTA is present. Section 4.2 on page 56 of this Chapter already demonstrated that MNase activity can be synchronized and paused, limiting variations in chromatin digestion between droplets. Here, we sought to determine whether the MNase is completely inactivated after droplets fusion and over long incubation time (i.e ligation of barcodes to the nucleosomes). Indeed, even a residual enzymatic activity would lead to the digestion of all nucleic acids present in the droplets, including DNA barcodes.

The MNase inactivation in droplets including the final concentration of EGTA was calibrated by performing a simple model experiment. The latter was carried out as follows: 145 pl droplets (i.e volume of droplets after fusion) were produced, collected in a collection tube and incubated overnight. Each droplet contained all the reagents of a single-cell ChIP-seq assay excepting cells and barcoded hydrogel beads, replaced by 5×10^7 copies of a 70 bp oligonucleotide mimicking the quantity of nucleic acids normally present in droplets after fusion. To assess the impact of the concentration of EGTA on MNase activity, three emulsions were produced with varying final concentration of EGTA: 0 mM (digestion positive control), 13 mM and 26 mM. A digestion negative control consisting of droplets that did not contain MNase was used as a reference for estimating the proportion of non-digested oligonucleotide. After incubation, droplets were broken and oligonucleotides were then purified and analyzed by electrophoresis on a TapeStation instrument.

The calibration of the concentration of EGTA necessary to fully inactivate

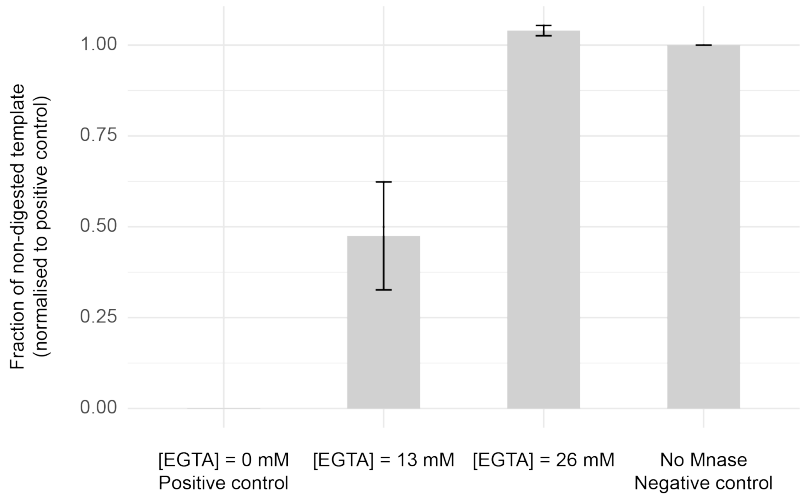


Figure 4.4.3: **MNase inactivation in droplets.**
Droplets containing 5×10^7 copies of a 70 bp oligonucleotide were incubated overnight with various concentration of EGTA (0 mM, 13 mM and 26 mM final concentration) and all reagents used in a single-cell ChIP-seq experiment. The fraction of oligonucleotide remaining after incubation was measured by TapeStation and normalized to the digestion negative control (i.e droplets without MNase). 26 mM final concentration of EGTA fully inactivated MNase in droplets. Barplot shows the mean fraction of non-digested oligonucleotides for duplicates and the error bars correspond to the standard deviation.

MNase activity in droplets is shown in Fig. 4.4.3 on the preceding page. The quantity of remaining oligonucleotides in each emulsion was measured by electrophoresis on a TapeStation instrument and normalized by the negative control. As expected and without addition of EGTA, oligonucleotides were completely digested in the positive control. A final concentration of 26 mM EGTA per droplet fully inactivated MNase as the proportion of remaining oligonucleotides was similar to the negative control and to the initial amount of 5×10^7 copies per droplet. Conversely, only one half of the original quantity of oligonucleotides was remaining with a final concentration of 13 mM EGTA per droplet.

In this manuscript, if not otherwise stated, all single-cell ChIP-seq experiments were performed with 30 mM EGTA final concentration per droplet.

4.4.3 Assessing barcode-nucleosome ligation efficacy in droplets

Nucleosome barcoding in droplets is critical to achieve high coverage single-cell chromatin profiles. Only nucleosomes ligated to a barcode are amplified during the sequencing library preparation while information carried by all other nucleosomes is lost. In order to estimate the performance of the system and the potential oss of information, it is important to estimate and optimize the efficiency of the ligation in droplets.

Estimating the ligation efficiency directly from a single-cell ChIP-seq experiment is challenging due to the complexity in terms of molecular diversity obtained after breaking droplets. The quantification of barcoded-nucleosomes from a mixture containing free barcodes ; barcode concatemers ; barcoded and non-barcoded mono-, di-, trinucleosomes... is almost impossible. To overcome this situation, we sought to design a simplified model experiment allowing an estimation of ligation efficiency in droplets. We also used this model in a second time to define the ligation incubation time that would maximize the proportion of barcoded-nucleosomes.

We reasoned that all variables that would make the interpretation of the results complicated should be eliminated. Therefore, we imagined a model experiment which principles are illustrated in Fig. 4.4.4. The experiment was carried out as follows: 145 pl droplets (i.e volume of droplets after fusion) were produced, collected in a collection tube and incubated off-chip for different incubation time. Each droplet contained all the reagents of a single-cell ChIP-seq assay excepting cells and barcoded hydrogel beads, replaced by 2 synthetic oligonucleotides:

- 5×10^7 copies of a 70 bp oligonucleotide mimicking the barcodes (equivalent

to the mean number of barcodes delivered per droplet). This oligo is framed with half of the restriction site of Pac1, allowing digestion and removal of barcodes concatemers which would artifically increase the ligation efficacy.

- 10^7 copies of a 138 bp oligonucleotide mimicking the nucleosomes (based on an estimation of the number of nucleosomes in a human cell.).

After incubation for ligation, emulsion was broken and treated with Pac1 for barcode concatemers digestion that would artificially increase the proportion of ligated product. The use of synthetic oligonucleotides had the major advantage that we could design qPCR primers to quantify the number of ligated product. We used two couples of primers: primers #1 (in blue) amplified only ligated products while the primers #2 (in red) amplify both ligated and non-ligated "nucleosome-like oligo". The ligation efficiency is then defined as the proportion of the ligated product over the total amount of "nucleosome-like oligo". We also used the electrophoresis profiles obtained by Tapestation as a second measurement method.

Longer incubation time (from 2h to overnight) leads to a 10% increase in the proportion of ligated product, as seen in Fig. 4.4.5. From the two measurement methods we can estimate that one half of nucleosome-like oligonucleotides are ligated with a barcode-like oligo.

Obviously, the results shown in Fig. 4.4.5 are not representative of the reality. The simplicity of the model experiment doesn't take into account important factors that might negatively impact the efficacy of the ligation of the barcodes to the nucleosomes. For example: this model is based on simple and homogeneous oligonucleotides rather than varying size of DNA/proteins complexes (i.e nucleosomes) or DNA barcodes released from hydrogel beads. To overcome in part these limitations, the "nucleosome-like oligo" were replaced by nucleosomes from human T-lymphocytes cell line and processed as previously mentioned. The fraction of ligated nucleosomes was only estimated from electrophoresis profiles on a Tapestation instrument but was similar to the ~50% obtained with the synthetic oligonucleotide (see "T-cell nucleosomes" on Fig. 4.4.5).

In this manuscript, if not otherwise stated, all single-cell ChIP-seq experiments were performed with an overnight incubation for nucleosome barcoding in droplets.

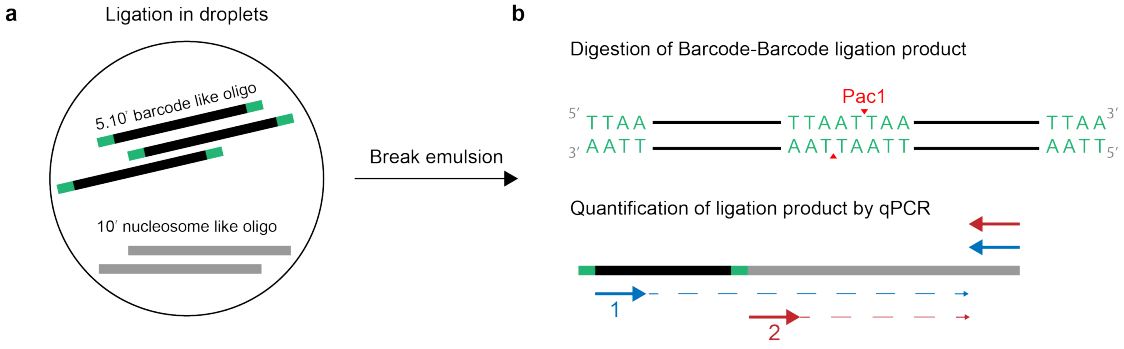


Figure 4.4.4: **Schematic drawing illustrating the model experiment used to assess ligation efficacy in droplets.**
(a) 145 pl droplets containing 5×10^7 copies of a "barcode-like oligo" and 10^7 copies of a "nucleosome-like oligo" were incubated overnight with all the reagents used in a single-cell ChIP-seq experiment (excepting cells and barcoded hydrogel beads).
(b) After incubation, droplets were broken and the aqueous phase was treated with Pac1 restriction enzyme to digest self-ligated product of the "barcode-like oligo". The ligation efficiency was then calculated either from TapeStation profiles or by qPCR. Primers #1 (in blue) amplified only ligated products while primers #2 (in red) amplified both ligated and non-ligated "nucleosome-like oligo". Ligation efficacy is defined as the proportion of the ligated product over the total amount of "nucleosome-like oligo".

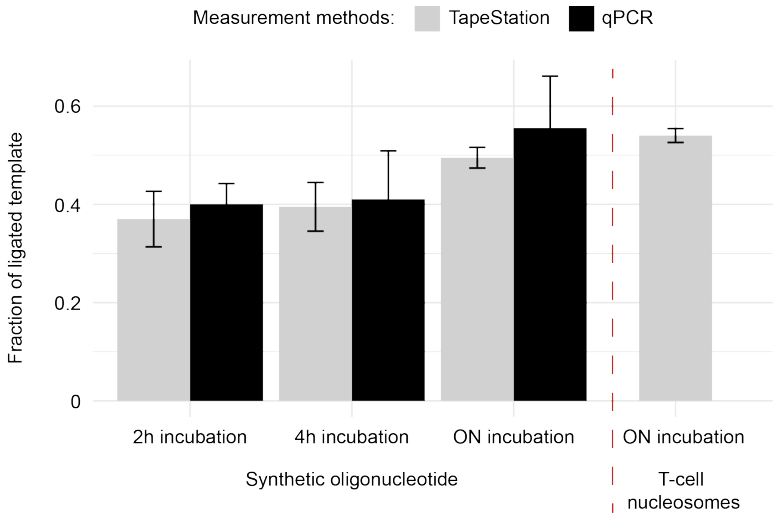


Figure 4.4.5: **Estimation of the ligation efficiency in droplets.**
The fraction of ligated product was measured either by analysis of TapeStation profiles or by qPCR as described in Fig. 4.4.4. The ligation was performed for 2h, 4h and overnight incubation. Both measurement methods gave similar estimation of the fraction of ligated product and a significant increase in efficiency was observed after an overnight incubation (~10%). Nucleosomes from human T-lymphocytes cell line was also used instead of the "nucleosome-like oligo" and confirmed similar results compared to the synthetic oligonucleotides (~50% ligation). Barplot shows the mean fraction of ligated oligonucleotides for duplicates (experiments performed on two different days) and the error bars correspond to the standard deviation.

4.5 Conclusion & perspectives on the droplet-microfluidic workflow

This chapter was centred around the building blocks of the droplet-microfluidic workflow with respect to chromatin barcoding in droplets.

First, we investigated potential sources of technical variability among single-cells and across single-cell ChIP-seq experiments. This led us to anticipate variability in MNase activity between droplets, notably upon single-cell encapsulation. We found that collecting and storing droplets on ice actually synchronized and paused MNase activity in droplets, thus preventing cell-to-cell differences in chromatin digestion.

Secondly, DNA barcodes are core components of the technology and their structure considerably impacted the efficacy of chromatin indexing in droplets. We developed an optimized barcode design allowing linear amplification of all barcoded-nucleosomes without distinction (e.g not only symmetrically barcoded nucleosomes on both ends as in Drop-ChIP). By modeling the ligation in droplets, we evaluated to one half the number of nucleosomes that may be ligated in our experimental conditions.

Finally, our droplet-microfluidic workflow enables precise monitoring of each operation, which provides valuable information about the overall performances of the system. For example, differences between the number of cells co-encapsulated with barcoded hydrogel beads and the number of barcodes identified in the sequencing data would indicate elevated level of noise, or on the other hand, low efficacy of the system (high loss of material).

A patent application related to this droplet-microfluidic workflow has been filed.

Chapter 5

Single-cell ChIP-seq identifies rare sensitive tumor cells with chromatin state similar to resistant tumor cells after treatment

Chromatin states are highly indicative of cell type and tissue identity. The power to identify cell type-specific chromatin states relies on measuring coherent variations between single-cell profiles. A high single-cell coverage is a must-have in order to reveal rich biological information and distinguish underlying patterns of variability in complex heterogeneous samples (e.g tumor specimens).

Previous Chapter 4 introduced the single-cell ChIP-seq platform developed during this thesis, with a main focus on the droplet-microfluidic workflow. It already discussed several lines of action with respect to chromatin barcoding in droplets towards the generation of reproducible and high quality single-cell chromatin profiles. For examples, the fine-tuning of enzymatic activities and the design of the DNA barcodes were determined to play important role in the final single-cell coverage.

On the other hand, a high single-cell coverage doesn't necessarily guarantee the identification of coherent variations across single-cell chromatin profiles, nor the validity of the classification of single-cells into subpopulations. In particular, the single-cell profiles can be highly sensitive to technical attributes and poorly specific to their initial chromatin states. The latter scenario has raised important questions with regard to the validation of the technological approach: Is single-cell information retained throughout the procedure? How specific are the single-cell profiles? How accurate is the classification of distinct cell types on the basis of their single-cell chromatin profiles?

This Chapter 5 outlines the framework established to unambiguously distinguish cell type-specific chromatin states from single-cell ChIP-seq profiles with nearly 100% accuracy. In light with the promising results obtained in this proof-of-concept, we initiated a collaboration with Dr. Céline Vallot's group at Institut Curie to investigate patterns of variability in the context of drug-sensitive and drug-resistant cell states in patient-derived xenograft models of breast cancer.

The outcome of this study is part of a paper manuscript included in section 5.2 at the end of this Chapter.

Grosselin K., Durand A., Poitou A., Marangoni E., Nemati F., Dahmani A., Reyal F., Frenoy O., Pousse Y., Reichen M., Woolfe A., Brenan C., Griffiths A. D.*, Vallot C.*, Gérard A.* *Single-cell chromatin profiling reveals heterogeneity of chromatin states in breast cancer.* In preparation.

5.1 Reconstructing cell type-specific chromatin states from single-cell ChIP-seq profiles

A central question coming with the conception of the scChIP-seq procedure is to determine whether the single-cell chromatin profiles can reveal distinct subpopulations and characterize cell type-specific chromatin states with high accuracy. Following this direction, we conceived a proof of concept study relying on the prior knowledge of the initial composition of the input sample and on the expected outcome. Doing so, we were able to support the robustness and validity of the technological approach.

This section presents the three main stages of the proof of concept study, each stage has been built with an increasing level of complexity compared to the previous one. First, we sought to measure the level of cross-contamination and confirm that the single-cell information was retained during the entire procedure. Then, we asked whether the classification of the single-cells into subpopulations was related to coherent variations in their chromatin profiles or to technical attributes. Doing so, we were able to evaluate the accuracy of the clustering. Finally, we demonstrated the capability of the system to resolve distinct cell types from a heterogeneous cell suspension.

The results of the proof of concept study are also part of the paper manuscript presented in section 5.2 of this Chapter (included in Fig. 1 & fig S2-6 of the paper manuscript). For clarity and to avoid repetition, only the reasoning behind the construction of the study as well as the striking results are detailed in this section.

5.1.1 Maintaining single-cell resolution throughout the scChIP-seq procedure

Identification of distinct cell types based on their patterns of histone modifications is only possible by achieving a single-cell resolution. For this purpose, nucleosomes from single-cells are indexed in droplets with DNA barcodes unique to a cell, before being combined for the immunoprecipitation and sequencing library preparation. The deconvolution of the barcode sequencing reads attributes each sequence to its cell of origin, thus generating genome-wide single-cell chromatin maps of the position of the modified histones. The single-cell resolution is achievable providing that barcodes are not mixing during the microfluidic workflow (e.g droplet coalescence after fusion or during incubation for nucleosome-barcode ligation) nor after combining droplets content for immunoprecipitation (at that time of the process, barcoded-nucleosomes are not purified yet).

We sought to determine if the system was actually maintaining the single-cell resolution by measuring the level of inter-specie cross-contamination. For this purpose, we carried out a specie-mixing experiment in which the scChIP-seq procedure was applied to barcode a mixed cell suspension of mouse and human cells (initial proportion 1/3rd mouse and 2/3rd human cells). We collected 3,000 droplets containing a cell co-encapsulated with a barcoded bead as calculated from fluorescent monitoring of droplet composition, performed chromatin immunoprecipitation targeting H3K27me3 and sequenced the library.

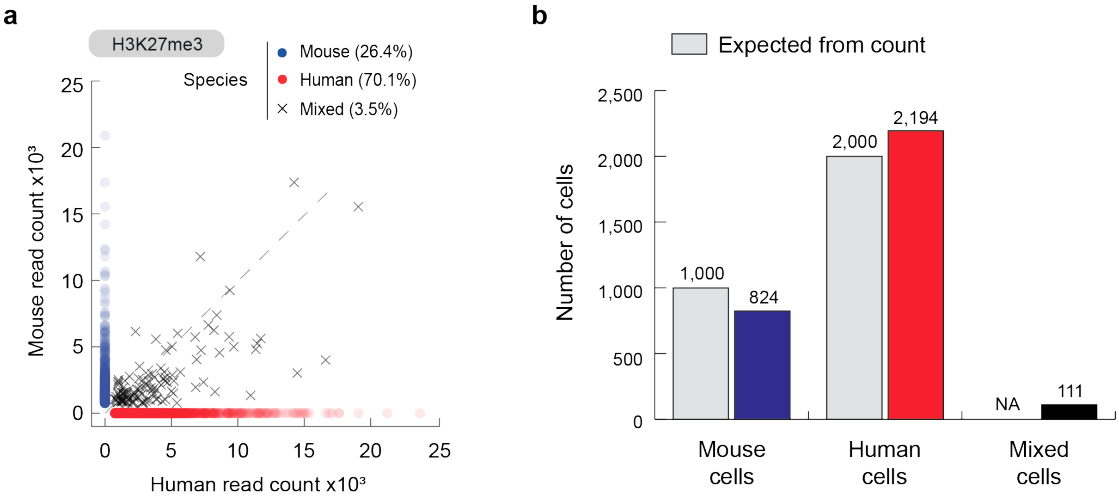


Figure 5.1.1: **Sequencing a mix of human and mouse cells reveals specie-specific mapping and single-cell resolution.**
(a) Scatter plot depicting the number of reads per barcode aligning on the mouse versus human reference genome. Barcodes are attributed to one specie if more than 95% of their reads are specific to this specie. Points are colored by specie and mixed barcodes are represented by black crosses.
(b) Barplot showing the number of barcodes identified for each specie after sequencing in comparison with the number of cells co-encapsulated with a barcoded bead calculated from fluorescent monitoring on the microfluidic station (light grey bars).

Sequencing reads associated with barcodes (i.e read #2) were processed as described in the Material and Methods of the paper manuscript (Appendix A) and in Appendix B. In parallel, sequencing reads associated with the nucleosomal sequence (i.e read #1) were aligned successively on both mouse and human genome.

For each identified barcode, the number of reads aligning the mouse genome was plotted versus the number of reads aligning the human genome. We observed that 96.5% of the barcodes had at least 95% of their reads mapping to either mouse or human genome (Fig. 5.1.1a). Importantly, among these barcodes, the proportion assigned to mouse (27.3%) and human (72.7%) species as well as the number of barcodes (824 and 2,194 respectively) closely matched with the initial proportion and the number of droplets containing a cell and a barcoded bead as calculated from fluorescent monitoring of the droplets on the microfluidic station (Fig. 5.1.1b).

The conclusions of this specie-mixing experiment are particularly important as they validate part of the technological development. First, barcodes are unique to a single-cell confirming that droplets remained independent throughout the entire microfluidic workflow. In addition, combining content from thousands of droplets for the immunoprecipitation step doesn't lead to non-specific nucleosome barcoding which would have been misleading in the reconstruction of cell type-specific chromatin profiles. Second, the robustness of the process is also supported by the striking correlation between the number of different barcodes identified by the droplet count and the sequencing data. The live monitoring of droplet-microfluidic operations is valuable to assess the level of noise and ensure reproducibility across experiments.

5.1.2 Accurate clustering of cell type-specific chromatin states

A high specificity and accuracy of the scChIP-seq procedure are essential to distinguish coherent patterns of variability in complex heterogeneous samples and to classify single-cells according to their distinctive chromatin landscapes. We conceived a second proof of concept experiment enabling to simultaneously measure the specificity and the accuracy of the procedure.

Experimental model applied to distinguish single-cells from the same specie

The design of the proof of concept is the following: human T lymphocytes and human B lymphocytes were separately encapsulated and successively indexed using two distinct barcodes Sets. These Sets of barcodes were slightly adapted to comprise the 3 index part of the single-cell barcodes, as well as an additional 15 bp sequence specific for each cell type (i.e common to all the single-cells of one cell type). To demonstrate the versatility of the scChIP-seq platform in profiling distinct chromatin states, we performed chromatin immunoprecipitation targeting histone modifications associated with either active transcription (H3K4me3) or

repressed gene expression (H3K27me3) (see Fig. 5.1.2a).

Each sequencing read associated with the DNA barcode would carry a double information: (i) the single-cell barcode sequence, which is used to assign the read to its cell of origin and (ii) the cell type-specific sequence, which is used to assign the read to one of the two cell type (B or T-cells). Ideally, the clustering of the single-cell profiles would separate the barcodes into two groups, which identity would be then confirmed by the cell type-specific sequence (see Fig. 5.1.2b).

From these two groups, the accuracy of the scChIP-seq procedure can be calculated by counting the number of mis-classified cells within each cluster (identified from the cell type-specific sequence). In turn, the specificity can be determined by comparing cumulative single-cell profiles with bulk ChIP-seq profiles and by computing genome-wide correlation between the two datasets.

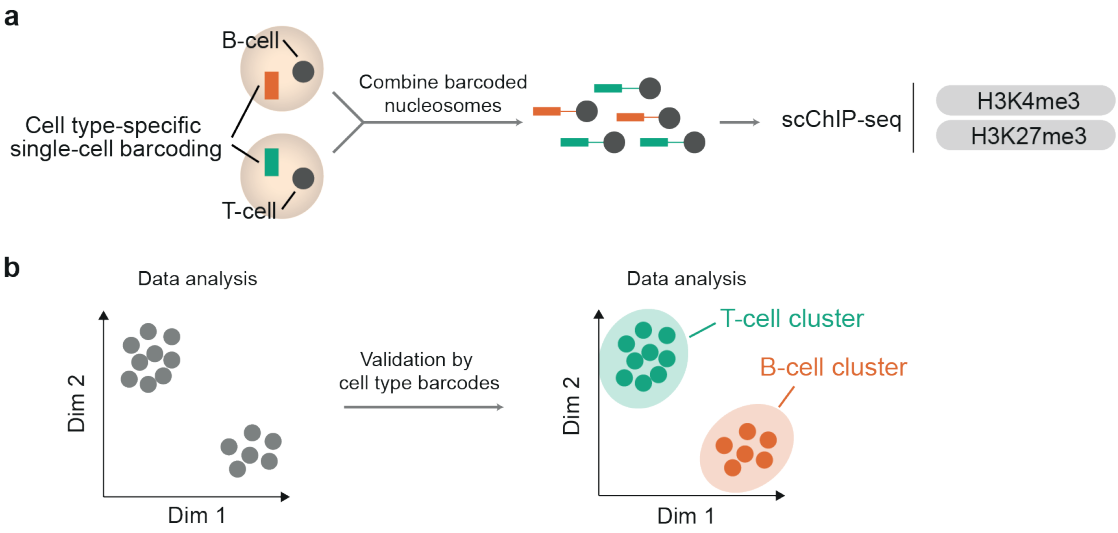


Figure 5.1.2: **Proof of concept study and expected outcome.**
(a) Human B and T lymphocytes were separately encapsulated and specifically indexed in droplets. Indexed chromatin from the two emulsions were then combined for scChIP-seq targeting histone modifications associated with distinct chromatin states (H3K4me3: active transcription; H3K27me3: repressed gene expression).
(b) Expected outcome of the study. Data analysis would group cells into two clusters using an unsupervised clustering approach. The correct clustering of B and T-cells based on their single-cell profiles would then be confirmed by the cell type-specific barcode sequence.

Appropriate precautions were considered to ensure the validity of the separate barcoding of B-cells and T-cells in droplets:

1. Specie-mixing experiment confirmed the independence of each droplet and the maintenance of single-cell resolution throughout the entire procedure.

2. After nucleosomes barcoding in droplets, indexed chromatin from B-cells and T-cells were combined for the immunoprecipitation and sequencing library preparation. Doing so, no bias between the two cell types were introduced at these stages ("batch effect").

3. The droplet microfluidic workflow of the scChIP-seq procedure is reproducible as showed by the high correlation across technical replicates suggesting that separate barcoding didn't introduce technical variability between the two cell types (Pearson correlation scores: 0.96, 0.95 and 0.97 with p-value $< 10^{-15}$ across 3 technical replicates, see supplementary fig. S7D of the paper manuscript)

The scChIP-seq dataset generated in this proof of concept experiment has also contributed significantly to the development of the bioinformatic pipeline. If not otherwise stated, all the results presented in this section were obtained from H3K27me3 scChIP-seq sample (results for H3K4me3 scChIP-seq sample are available in the paper manuscript).

Deconvolution of single-cell barcodes and cell type-specific sequences

Sequencing reads #2 were processed as described in the Material and Methods of the paper manuscript (Appendix A) and in Appendix Bwith a modification in the barcode extraction step to also take into account the cell-type specific sequences. The distribution of the number of reads per barcode provided a first insight about the quality of the dataset and was used as a proxy for eliminating background noise (Fig. 5.1.3a). The distribution can be fitted as a sum of two normal distributions: the first one containing barcodes with few reads (less than few hundreds) was associated with background; the second one grouping barcodes having more than 500 reads was assumed to originate from single-cells. A threshold was set between the two distributions and the barcodes with few reads were discarded (dotted line at 500 reads per barcode on Fig. 5.1.3a). We hypothesized that the split-pool synthesis and the "purity" of the barcodes bound on beads underlied the presence of low read count barcodes in the sequencing data. Indeed, the quality controls of the barcoded beads and notably the single bead sequencing revealed that in average 97.7% of the barcodes on a bead were matching to the same sequence whereas the second most abundant sequence was only representing 0.17% of all

sequencing reads (see Chapter 4, section 4.3 & Fig. 4.3.1 on page 64). As a result, an average of 2.8% of the barcodes bound on a bead were completely different which might lead to low read count barcodes after sequencing.

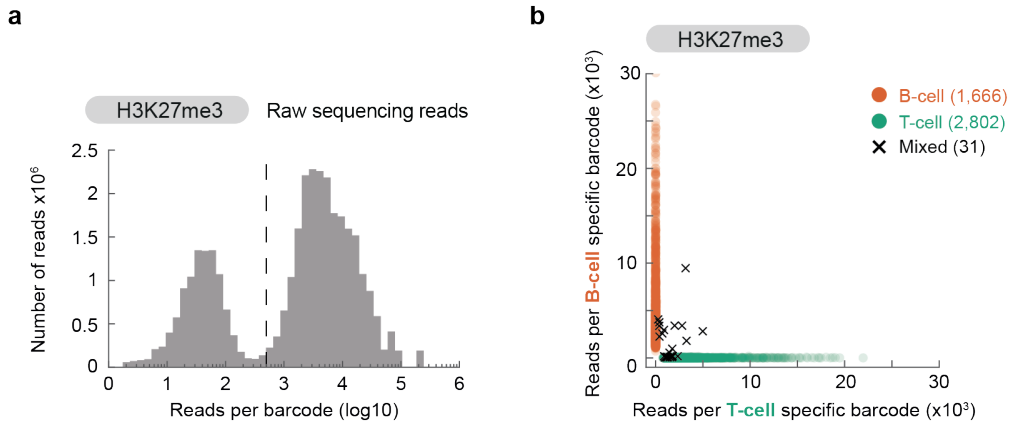


Figure 5.1.3: **Deconvolution of single-cell barcodes associated with cell type-specific sequences.**
(a) Histogram showing the distribution of the number of reads per barcode. Barcodes with few reads (< 500 reads) were associated with background, whereas barcodes with more than 500 reads were assumed to originate from single-cells. A threshold set at 500 reads (dotted black line) was used as a cutoff to eliminate noise.
(b) Scatter plot depicting for each barcode the number of raw reads aligned on the B-cell specific sequence versus aligned on the T-cell specific sequence. Points are colored by cell type-specific sequence and mixed barcodes are represented by black crosses. Barcodes were highly specific to one cell type.

In this particular proof of concept experiment, human T-cells and B-cells were separately encapsulated and indexed in droplets using two distinct barcode Sets. Both Sets comprised the single-cell barcodes, as well as a sequence specific of each cell type. For each barcode surviving the threshold as defined in Fig. 5.1.3a, the number of reads bearing the T-cell specific sequence was plotted versus the number of reads bearing the B-cell specific sequence. As expected, the vast majority of the barcodes (99.3%) were highly specific as having at least 95% of their reads matching one of the two Set. Similarly to the specie-mixing experiment, this result was suggesting again that combining indexed chromatin from thousands of cells for immunoprecipitation and sequencing library preparation didn't lead to non-specific nucleosome-barcoding (Fig. 5.1.3b). On the other hand, few barcodes (0.7%) were sharing reads aligned on both Sets. We hypothesized that

these barcodes were originating from barcoding errors with two cells sharing the same barcode sequence (see introductory Chapter 2, section 2.3.1 or [Klein 15] for original description of the calculation).

The distribution presented in Fig. 5.1.3a is related to raw sequencing reads per barcode. After selection of barcodes above the noise threshold, duplicate reads and reads possibly originating from the same nucleosome were discarded for further analysis. For the two histone marks, we achieved an average coverage of 1,630 uniquely mapped reads per cell (see Fig. 1B of the paper manuscript).

Mark	**# of cells**	**# of raw reads**	**Average # of raw reads per cell**	**Reference**
H3K4me4	~6,000	65 million	~11,000	[Grosselin et al, in preparation]
	~2,300	322 million	~140,000	[Rotem 15a]
	~4,300	250 million	~58,000	[Rotem 15a]
H3K27me3	~4,500	72 million	~14,400	[Grosselin et al, in preparation]

Table 5.1: **Sequencing performance (1/2).**
The table compares the number of expected cells per sequencing library, the number of raw sequencing reads as well as the average number of raw reads per cell in Drop-ChIP and in our scChIP-seq system. #: number

Mark	**# of cells post sequencing**	**# of cells used in analysis**	**Average # of reads per cell after QC**	**Reference**
H3K4me4	6,134	3,112	1,630	[Grosselin et al, in preparation]
	1,716	1,020	381	[Rotem 15a]
	1,279	376	544	[Rotem 15a]
H3K27me3	4,470	2,232	1,637	[Grosselin et al, in preparation]

Table 5.2: **Sequencing performance (2/2).**
The table compares the number of cells identified after sequencing, the final number of cells used in analysis after QC and the average number of useable reads per cell after QC in Drop-ChIP and in our scChIP-seq system. #: number ; QC: Quality Control.

In terms of performance, the final coverage obtained with our system is already 3 to 5-fold higher than the coverage obtained with previously reported Drop-ChIP method [Rotem 15a]. Importantly, our sequencing libraries were not sequenced at saturation and we anticipate an increase of the coverage with higher sequencing depth (see comparison of performances between the 2 systems in Table 5.1 and Table 5.2).

Unsupervised clustering of single-cell chromatin profiles

Unlike single-cell RNA-seq data analysis, in which sequencing reads are associated to genes used as comparison units, reads from single-cell ChIP-seq data are distributed along the entire genome. One possible way to overcome this situation is to align the reads on a set of regions known to be enriched or conversely depleted for a particular histone modification. However, this type of supervised analysis is based on prior knowledge of the composition of the sample to be analyzed. A second possibility is to perform an unsupervised analysis to identify cell-to-cell variation without prior information about their chromatin profiles.

For this purpose, reads from each single-cell were binned in non-overlapping regions spanning the genome, representing each cell as a vector. The latter was aggregated in a m x n coverage matrix, with m rows corresponding to the genomic bins and n columns corresponding to the single-cells. Matrix value c_{ij} corresponded to the number of reads for barcode j aligned in region i. The bioinformatic tools used to generate the coverage matrix from raw sequencing reads have been developed during this thesis and are further described in the Material and Methods of the paper manuscript (Appendix A) and in Appendix B. The resulting coverage matrix was then used as a starting point for downstream analysis specific to each experiment.

Unsupervised analysis and clustering of single-cell H3K4me3 and H3K27me3 profiles for this particular proof of concept experiment are both presented on Fig. 1C of the paper manuscript. For the two histone modification marks, barcodes were grouped by consensus clustering into two well separated clusters, which identity was confirmed by the cell type-specific sequence (H3K4me3: 99.3% accuracy; H3K27me3: 99% accuracy).

H3K4me3 is known to accumulate around promoters of transcriptionally active genes whereas H3K27me3 accumulates in broad domain of facultative heterochromatin. Aggregation of single-cell profiles within each cluster enabled to reconstruct T-cell specific and B-cell specific chromatin states with high accuracy as compared to bulk ChIP-seq profiles. Snapshots of differentially enriched loci with cumulative

single-cell profiles for each identified cluster and bulk profiles are illustrated in Fig. 1D of the paper manuscript. For example, single-cells profiles within the T-cell cluster are enriched in H3K4me3 around CD3 genes (a T-cell specific gene) and depleted around CD22 gene (a B-cell specific gene). Conversely, H3K27me3 is a chromatin mark associated with inactive transcription, meaning that enrichment is inversely correlated with H3K4me3 profiles. As expected, single-cells profiles within the T-cell cluster reveal H3K27me3 enrichment around CD22 gene (a B-cell specific gene) and H3K27me3 depletion around CD3 genes (a T-cell specific gene).

H3K4me3 and H3K27me3 profiles around both CD3 and CD22 loci highlight the local specificity of the aggregated profiles obtained with our scChIP-seq procedure. By computing genome-wide correlation between aggregated single-cell profiles and bulk profiles, we confirmed the specificity of the data at a global scale (Pearson correlation scores: 0.93 [H3K4me3] and 0.97 [H3K27me3] with p-value $< 10^{-15}$. See Fig. 1E of the paper manuscript).

5.1.3 *In silico* simulation of detection limit

The ideal dataset generated in this proof of concept was used to measure accuracy and specificity of single-cell chromatin profiles, but it can be adapted to simulate *in-silico* the sensitivity to detect rare cell populations. For this purpose, 500 T-cell profiles were randomly subsampled from the dataset and a decreasing proportion of randomly selected B-cell profiles were spiked-in with the selected T-cell profiles. To control the impact of the cell coverage, the random selection process was repeated with a minimum of 500, 1,000 and 2,000 uniquely mapped reads per cell (reads corrected for duplicates). The artificially created datasets were then processed using an unsupervised clustering approach as previously described.

The visualization of the clustering results on t-SNE plots confirms the intuitive idea that increasing the cell coverage tends to increase the sensitivity (Fig. 5.1.4). With a coverage of 500 uniquely mapped reads per cell, all the B-cells are mixed within the T-cells population highlighting again the importance of high cell coverage. Conversely, with a coverage of 2,000 uniquely mapped reads per cell, a subpopulation representing 5% of the dataset can still be unambiguously identified from the dominant cell type. The sensitivity would be further enhanced with higher cell coverage.

Other parameters might improve the sensitivity of the system in the detection of rare cell populations. For example, increasing the total number of cells also contributes to lower the detection limit (data not shown in the manuscript).

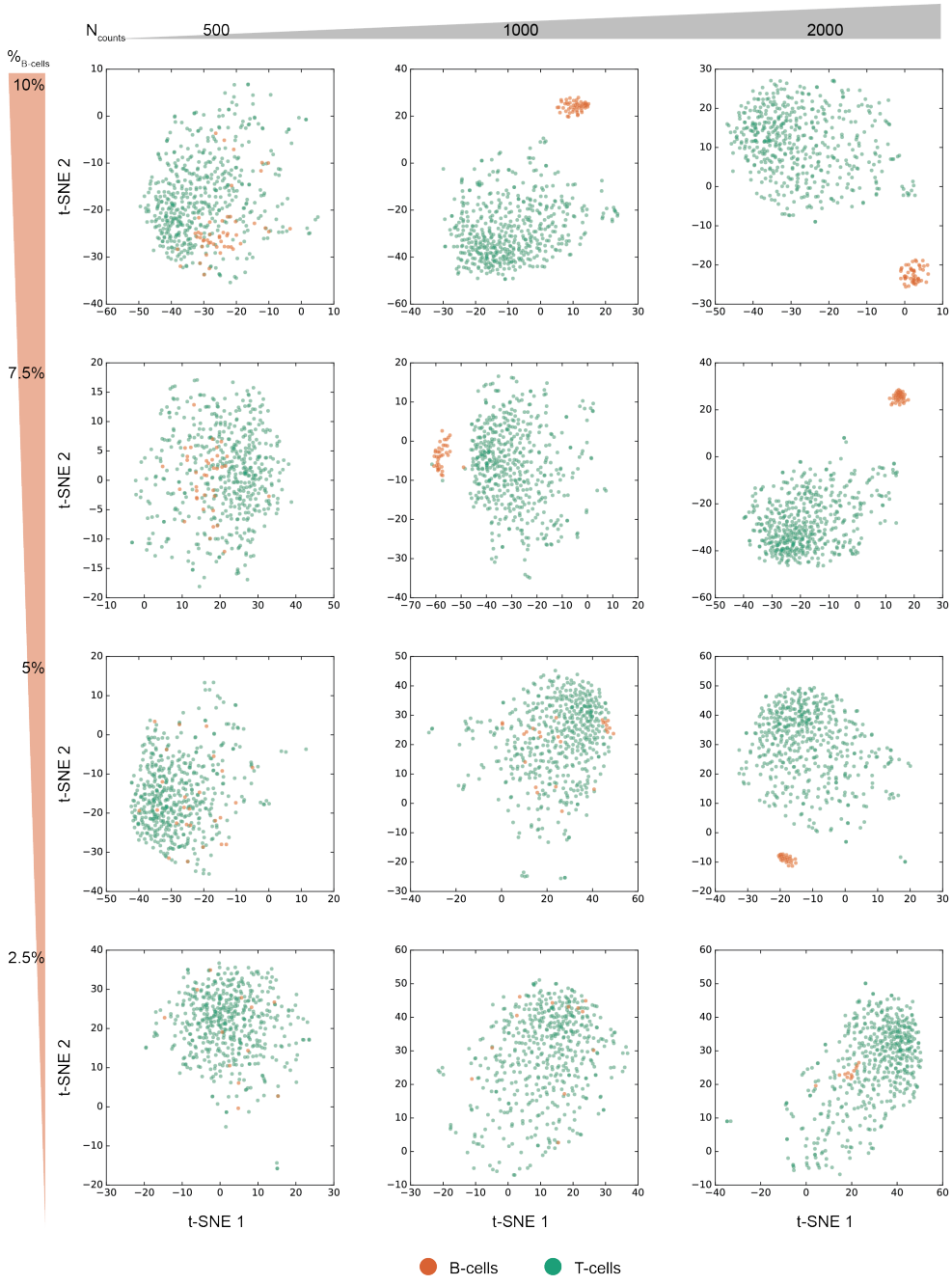


Figure 5.1.4: **Increasing cell coverage lowers detection limit of rare cell populations.**
t-SNE plots representing H3K27me3 scChIP-seq dataset in *in-silico* simulation of detection limit with varying proportion of spiked-in B-cells in T-cells profiles (from top to bottom) and varying thresholds of uniquely mapped reads per barcode (from left to right). Points are colored according to the cell type-specific sequence.

This is a growing trend in single-cell RNA-seq study in which sequencing large number of cells at low depth is prefered to low cell number with high coverage in order to characterize transcriptomic heterogeneity at single-cell resolution [Cao 17, Rosenberg 18].

5.1.4 Distinguishing subpopulations from heterogeneous cell suspension

Single-cell chromatin profiles generated by our scChIP-seq platform allowed a precise identification of cell type-specific chromatin states. However, these results were obtained from ideal datasets, with different cell types being separately processed in the microfluidic workflow and indexed with distinct cell type-specific barcodes. Those ideal scenarios don't reflect the complexity of the samples we wish to analyse with our system (in terms of sample heterogeneity, purity or subpopulation scarcity).

Going back to the mouse/human specie-mixing experiment (see subsection 5.1.1 on page 85 of this Chapter), a simple alignement on both reference genomes revealed that 96.5% of the barcodes were unambiguously attributed to one of the two species. In addition, the initial proportion of starting cells was also preserved (Fig. 5.1.5a; 3,000 cells co-encapsulated with a barcoded bead; initial proportion 33% mouse and 66% human cells). Human cells were actually composed of a 1:1 mixture of B and T lymphocytes cells. To demonstrate the capability of the system in identifying cell types from heterogeneous sample, only the human barcodes were selected and processed in an unsupervised clustering approach as previously described.

Unsupervised clustering of selected human barcodes revealed two well separated clusters independent from cell coverage and containing 1,063 (58%) and 760 (42%) single-cells respectively (Fig. 5.1.5b). In order to characterize cluster identity, single-cell data were aggregated to generate cluster-specific chromatin profiles. Doing so, pattern of H3K27me3 enrichment around CD3 genes (a T-cell specific gene) was comparable between C1 and Bulk B-cells ChIP-seq, suggesting repressed expression of CD3 genes in C1. Conversely, pattern of H3K27me3 around CD22 gene was comparable between C2 and Bulk T-cells ChIP-seq, suggesting in turn repressed expression in C2 (Fig. 5.1.5c).

The number of single-cells comprising each cluster was closely related to the initial number cells from each cell type, echoing previous statement about the robustness of the system. From the initial cellular suspension containing equiva-

lent numbers of mouse cells, human B-cells and human T-cells, 3,000 cells were co-encapsulated with a barcoded bead. After sequencing data analysis, 2,647 barcodes (88%) were unambiguously assigned to one of the three cell type. Among them, 824 barcodes were assigned as mouse cells (31%), 1,063 barcodes were assigned as human B-cells (40%) and 760 barcodes were assigned as human T-cells (29%). We assumed that the sample was not sequenced at saturation and increasing the number of reads would bring the number of identified barcodes closer to the expectation.

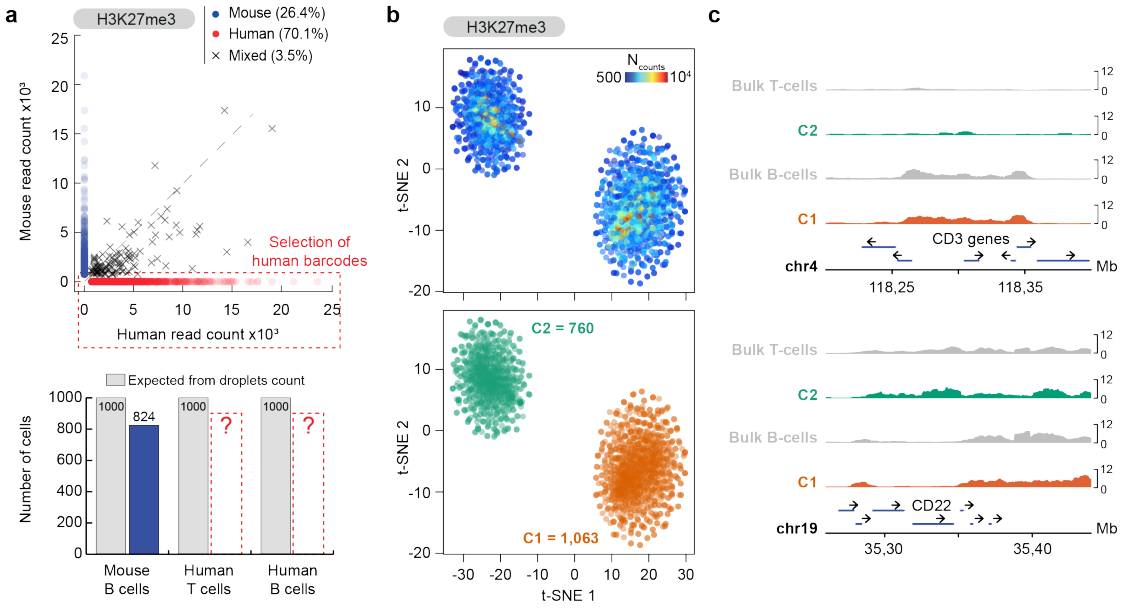


Figure 5.1.5: **Deciphering chromatin states from heterogeneous samples.** (a) Top: scatter plot as shown in Fig. 5.1.1a from specie-mixing scChIP-seq experiment (1/3rd mouse cells; 2/3rd human cells). Human cells were composed of 1:1 mixture of B and T lymphocytes. Bottom: barplot showing the expected number of barcodes in comparison with droplets count.
(b) Clustering of human barcodes reveals two well separated populations independent from cell coverage as shown on t-SNE plots. The number of cells per cluster closely matched with expected number of each human cell type from the original mixture.
(c) Single-cell profiles within each cluster were aggregated to generate population chromatin profiles. Examples shown around a T-cells specific gene (CD3 genes) and a B-cells specific gene (CD22 gene) confirmed the identity of C1 (B-cells) and C2 (T-cells) clusters.

5.1.5 Conclusion & perspectives

This proof of concept study validates our scChIP-seq procedure as a robust method to profile chromatin landscape at single-cell resolution and distinguish patterns of variability in heterogeneous sample. Single-cells are classified with a high accuracy and their profiles specifically recapitulate cell type-specific chromatin states with high fidelity. These results also highlight the importance of our technological development towards the generation of high coverage single-cell profiles. Indeed, the sensitivity of the system to detect rare cell populations is directly related to the single-cell coverage.

Altogether, the single-cell ChIP-seq method developed during the thesis outperforms previously reported Drop-ChIP method and opens new avenues for epigenomic studies at single-cell resolution. In light with these promising results, we initiated a collaboration with Dr. Céline Vallot's group at Institut Curie to investigate the heterogeneity of chromatin modifications in the context of drug-sensitive and drug-resistant cell states within patient-derived xenograft models of breast cancer. This study is part of a paper manuscript included in the following section.

5.2 Grosselin et al, *in preparation*

Single-cell chromatin profiling reveals heterogeneity of chromatin states in breast cancer

Kevin Grosselin[1,2], Adeline Durand[3], Adeline Poitou[1], Elisabetta Marangoni[4], Farida Nemati[4], Ahmed Dahmani[4], Fabien Reyal[4,5,6], Olivia Frenoy[1], Yannick Pousse[1], Marcel Reichen[1], Adam Woolfe[1], Colin Brenan[1,7], Andrew D. Griffiths[2]*, Céline Vallot[3,4]*, Annabelle Gérard[1]*

Affiliations:

[1] HiFiBiO SAS, 29 rue du Faubourg Saint Jacques, 75014 Paris, France

[2] Laboratoire de Biochimie, ESPCI Paris, PSL Research University, CNRS UMR8231 Chimie Biologie Innovation, F-75005 Paris, France

[3] CNRS UMR3244, Institut Curie, PSL Research University, F-75005 Paris, France

[4] Translational Research Department, Institut Curie, PSL Research University, F-75005 Paris, France

[5] INSERM U932, Institut Curie, PSL Research University, F-75005 Paris, France

[6] Surgical Department, Institut Curie, France

[7] HiFiBiO Inc, 700 Main Street, Cambridge, MA, 02139, USA

*Correspondence to: andrew.griffiths@espci.fr; celine.vallot@curie.fr; a.gerard@hifibio.com

Abstract:

The dynamic nature of chromatin and transcriptional features play a critical role in normal differentiation and are expected to participate to tumor evolution. Studying the heterogeneity of chromatin alterations with single-cell resolution is mandatory to understand the contribution of epigenetic plasticity to tumor evolution. Here, we describe a droplet microfluidics system that enables the profiling of chromatin marks in several thousand single cells, with a coverage of up to 10,000 loci per cell. In patient-derived xenograft (PDX) models of breast cancer with acquired drug resistance, scChIP-seq revealed that untreated, drug-sensitive, tumors contain a rare population of cells with chromatin traits similar to that of all resistant cells. Our results highlight the potential of chromatin traits as biomarkers of response and resistance to cancer therapy.

One Sentence Summary:

scChIP-seq discloses rare populations of cancer cells with distinct chromatin traits, characteristic of resistant tumor cells.

Genetic, transcriptional and epigenetic changes contribute to tumor evolution. Understanding both the intra-tumor heterogeneity and the selection dynamics of these molecular alterations is critical to determine their potential as therapeutic targets and design effective treatment rationales. The extensive modeling of clonal dynamics of genetic alterations, thanks to deep sequencing and single-cell approaches, have revealed the key contribution of genetic intra-tumor heterogeneity to the emergence of populations of cancer cells with drug resistance capacities (*1*, *2*). Resistance can arise from a sub-population of cells bearing a key mutation, decisive for the resistance process (*3*). However, while a number of genetic mechanisms driving metastatic or resistance processes has been discovered, it also appears that in many cases genetic mechanisms driving these processes cannot be found. Increasing efforts concentrate on non-genetic aberrations, particularly transcriptional and chromatin alterations, that may account for the adaptability of the cancer cells to cancer therapies (*4–7*). Development of single-cell RNA-seq has opened avenue for deep understanding of transcriptional heterogeneity within tumor samples and have revealed the emergence of transcriptional sub-clones in tumors upon treatment (*8–10*). So far, only few studies have tackled the clonal evolution of epigenetic alterations mostly looking at DNA methylation (*11–13*). The heterogeneity and evolution of chromatin-based alterations remain largely unknown, partly due to the lack of single-cell techniques interrogating the distribution of chromatin marks at the single-cell level within complex biological samples. Recently, our understanding of epigenetic in individual cells has been advanced by the development of single-cell epigenomics to analyze, for example, DNA methylation (*14–16*), chromatin accessibility (*17–19*), chromosome conformation (*20*) and DNA-proteins interactions (*21*). Yet, these methods yield low coverage per single-cell, preventing the analysis of complex samples, such as tumor specimens.

Here we describe a high throughput single-cell ChIP-seq (scChIP-seq) approach combining droplet microfluidics to the high depth of single-cell DNA barcoding technologies (fig. S1), to profile histone post-translational modifications at single-cell resolution with an unprecedented coverage of 10^3 to 10^4 unique loci per cell. Our scChIP-seq approach relies on a microfluidic system with live monitoring of droplets, which enables in three steps the controlled production of 145 ± 10 pl droplets containing both nucleosomes from individual cells and a hydrogel bead with barcodes (Fig. 1A, fig. S2 and movies S1 & S2). Each hydrogel beads carries ~5×10^7 copies of unique DNA barcode generated by split-and-pool synthesis (fig. S3). We monitor at each step the incorporation of cells and hydrogel beads

within droplets, through the addition of fluorophores and measurement in real time of the fluorescence within droplets (Fig. 1A and fig. S4). Off the microfluidic system, droplets are then incubated to allow the ligation of the nucleosomes to the single-cell barcodes released from the hydrogel beads by photo-cleavage. The content of merged droplets is used for immunoprecipitation of nucleosomes with post-translational histone modification of interest and the barcoded immuno-precipitated DNA is amplified and sequenced (fig. S5).

We first validated the efficiency and accuracy of our scChIP-seq procedure using cell lines. To evaluate the accuracy of our microfluidic workflow to produce chromatin profiles at the single-cell resolution, we profiled the distribution of H3K27me3 within a mixed population of mouse and human cells, by equally mixing one mouse and two human lymphoid cell lines (mouse M300.19 cells, human Ramos B cells and human Jurkat T cells). Post-sequencing, we confirmed that 96.5% of the identified barcodes were unambiguously assigned to a single species, indicating that the microfluidic system produced droplets with a single cell (fig. S6A). The total number of cells identified post-sequencing (3,018) fitted the number of cells co-encapsulated with barcoded hydrogel beads calculated from fluorescent monitoring of microfluidic operations (~3,000), validating the robustness of our procedure (fig. S6B). In addition, the proportion of cells for each species (824 mouse cells [27.3%], 2,194 human cells [72.7%]) was in close agreement with the initial proportion of cells from each species (1/3 mouse vs 2/3 human). To validate the specificity of our scChIP-seq procedure to classify single cells according to their distinctive chromatin landscape, we profiled in two cell lines, human Ramos B and Jurkat T cells, the distribution of two histone modifications with different binding profiles: H3K4me3, known to accumulate in narrow peaks around promoters and enhancers respectively, and H3K27me3, shown to accumulate in broad domain of facultative heterochromatin. To control for accuracy of our classification, we separately barcoded human Ramos B cell line and human Jurkat T cell line using two independent sets of barcodes. For the two histone marks, we achieved an average coverage of 1,630 uniquely mapped reads per cell (Fig. 1B) and a high correlation across technical and biological replicates (fig. S7D-E, $r = 0.96$ and 0.98 with $p < 10^{-15}$ respectively). For both single-cell chromatin profiling, we identified by consensus clustering two stable clusters corresponding to each cell line (Fig. 1C and fig. S7F), with a specificity over 99.3% and 99.0% respectively for H3K4me3 and H3K27me3 profiles as attested by the cell-type specific barcodes. Aggregated single-cell profiles recapitulated the bulk ChIP-seq profiles with high accuracy (Fig. 1D-E, $r = 0.93$ and 0.97 with $p < 10^{-15}$ for H3K4me3 and H3K27me3 respectively). Through differential analysis of H3K4me3 and H3K27me3

chromatin traits between Ramos and Jurkat cells, we identified concordant lineage-specific sets of genes as being enriched for H3K4me3 or H3K27me3 (fig. S7G-H). Altogether these results confirm our scChIP-seq procedure as a robust method to detect chromatin landscapes at the single-cell level, to classify single cells with a high accuracy according to their chromatin state and to identify discriminating chromatin traits between cell populations.

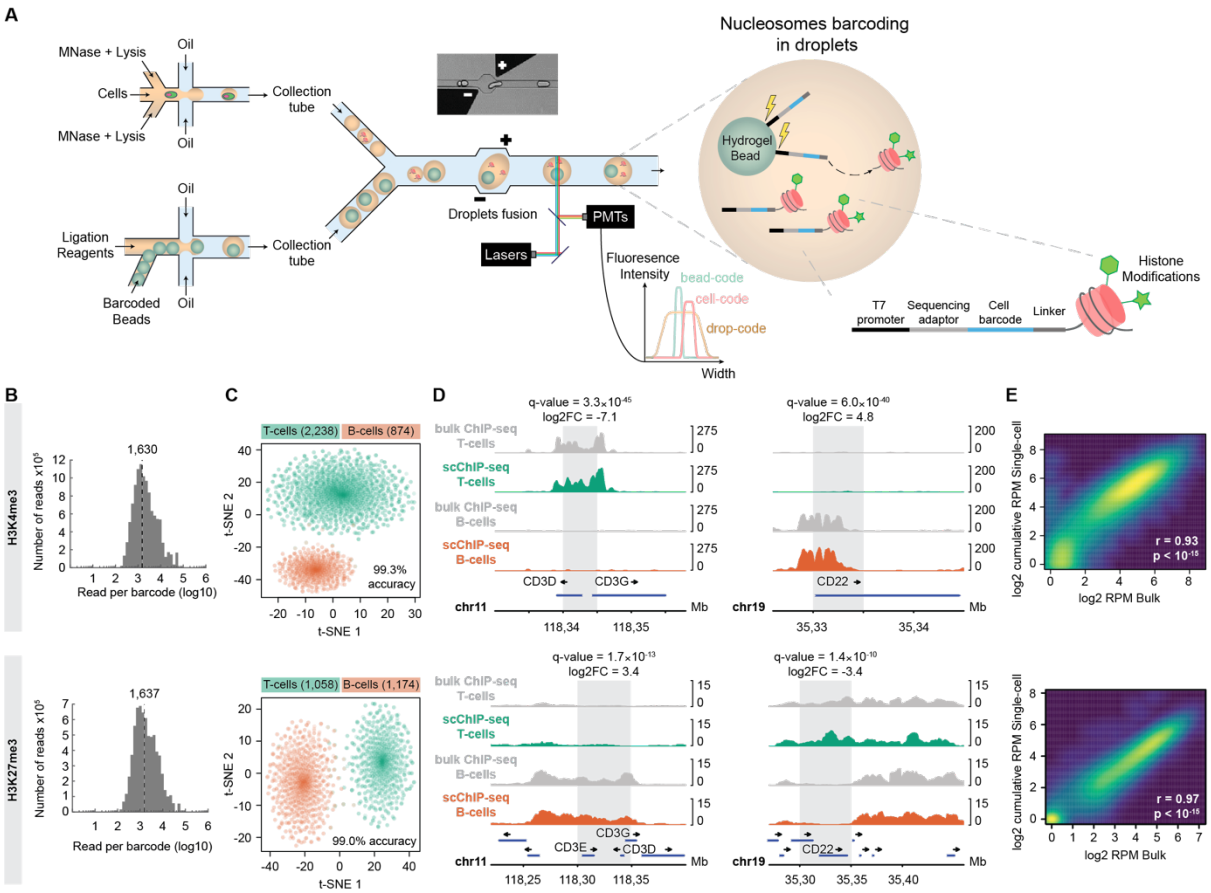


Fig. 1. Reconstructing cell type specific chromatin states from single-cell ChIP-seq profiles. (**A**) Overview of the microfluidic single-cell ChIP-seq workflow. Cells are compartmentalized in 45 pl droplets, lysed and their chromatin is fragmented by MNase. Hydrogel beads carrying single-cell barcodes are loaded in 100 pl droplets and electro coalesced one-to-one with droplets containing the digested chromatin. Droplet content is scanned after fusion to precisely measure the number of captured cells. Barcoded DNA adaptors are released by photo-cleavage and ligated to the nucleosomes in droplets. (**B - E**) Unsupervised analysis of H3K4me3 and H3K27me3 scChIP-seq data from human B and T lymphocytes separately indexed in droplets using single-cell cell-type specific barcodes. (**B**) Histograms of the distribution of uniquely mapped reads per cell, average coverage is indicated with a dotted line. (**C**) t-SNE plots representing H3K4me3 and H3K27me3 scChIP-seq data sets, points are colored according to cell type specific barcode sequence. (**D**) Snapshots of differentially enriched loci (fig. S7) with cumulative single-cell profiles for each

cell type and bulk profiles. Differentially bound regions identified by Wilcoxon signed-rank test are indicated in grey with the corresponding adjusted p-value and log2 fold change. (**E**) Scatter plots displaying log2 enrichments in cumulative single-cell versus bulk ChIP-seq data, calculated based on the count per million of mapped reads in 5 kb genomic bins for H3K4me3 and 50 kb for H3K27me3. Pearson correlation scores and p-values are computed genome-wide.

We then applied our scChIP-seq procedure to interrogate the heterogeneity of chromatin traits within breast tumor models of acquired resistance to cancer therapy. We profiled the H3K27me3 landscape at single-cell resolution of two pairs of patient-derived xenografts (PDX). In parallel, we also performed scRNA-seq to evaluate transcriptional heterogeneity within the same cell populations. We first studied a pair of triple-negative PDX samples (Fig. 2A): HBCx-95, responsive to Capecitabine treatment (*22*), and HBCx-95-CapaR, a tumor derivative with *ex vivo* acquired resistance to Capecitabine (fig. S8A).

We started by studying the diversity of chromatin traits within stromal cells originating from HBCx-95 and HBCx-95-CapaR, by isolating mouse sequences from both datasets (n = 1,766 cells with >1000 unique chromatin traits, average coverage of 3,535 reads per cell). Cumulative single-cell chromatin profiles matched bulk ChIP-seq profiles (fig. S8C, $r = 0.89$ with $p < 10^{-15}$), with a smaller amplitude of signal (fig. S8D, fold-change of 1.89). By unsupervised analysis, we could group cells independently of coverage starting 1,600 unique reads per cell (fig. S9A-E), threshold which was kept for all subsequent analysis (see Material & Methods). Consensus clustering approaches (fig. S9F-G) showed that stromal cells stably grouped in three '*chromatin-based*' populations according to H3K27me3 chromatin profiling, *Chrom_c1*, *c2* and *c3*, irrespectively of the PDX sample of origin (Fig. 2B). By comparing chromatin traits between groups of cells (fig. S10A), we identified loci with specific H3K27me3 enrichment and absence for *Chrom_c2* and *c3* populations (n = 2,933 and n = 2,550 respectively with q-value < 0.01 and |log2FC| > 1), and to a lesser extent for cluster *Chrom_c1* (n = 232).

In parallel scRNA-seq analysis revealed four populations of stromal cells (Fig. 2C and fig. S11): two groups of cells of fibroblast origin (with specific markers *Col12a1* and *Efemp1*), endothelial cells (*Pecam-1*) and macrophage cells (*Ms4a7*). To further compare the identity of populations inferred from both approaches, we focused on genes with a transcription start site located within 1 kb of chromatin traits specific of *Chrom_c1*, *c2* and *c3*. Relying on genes with lineage-specific expression pattern, e.g. *Ptk2* and *Il10ra*, scChIP-seq and scRNA-

seq data together indirectly suggested that *Chrom_c2* might correspond to cells of fibroblast lineage and *Chrom_c3* to cells of immune origin (Fig. 2D-E and fig. S10B). In addition, we performed pathway analysis (fig. S10A), which showed that loci devoid of H3K27me3, specifically in *Chrom_c2*, were located in the vicinity of genes involved in the epithelial to mesenchymal transition (q-value = 1.5×10^{-03}) or apical junction (q-value = 3.0×10^{-02}), confirming the potential fibroblast origin of cells within this cluster. We could not identify relevant genes or related pathways associated to the few chromatin traits characteristic of *Chrom_c1*, suggesting that this cluster of cells shared chromatin traits with both *Chrom_c2* and *c3*. It indeed appears for example that half of cells from this cluster share with immune-like cells an H3K27me3 enrichment for *Ptk2* (Fig. 2D).

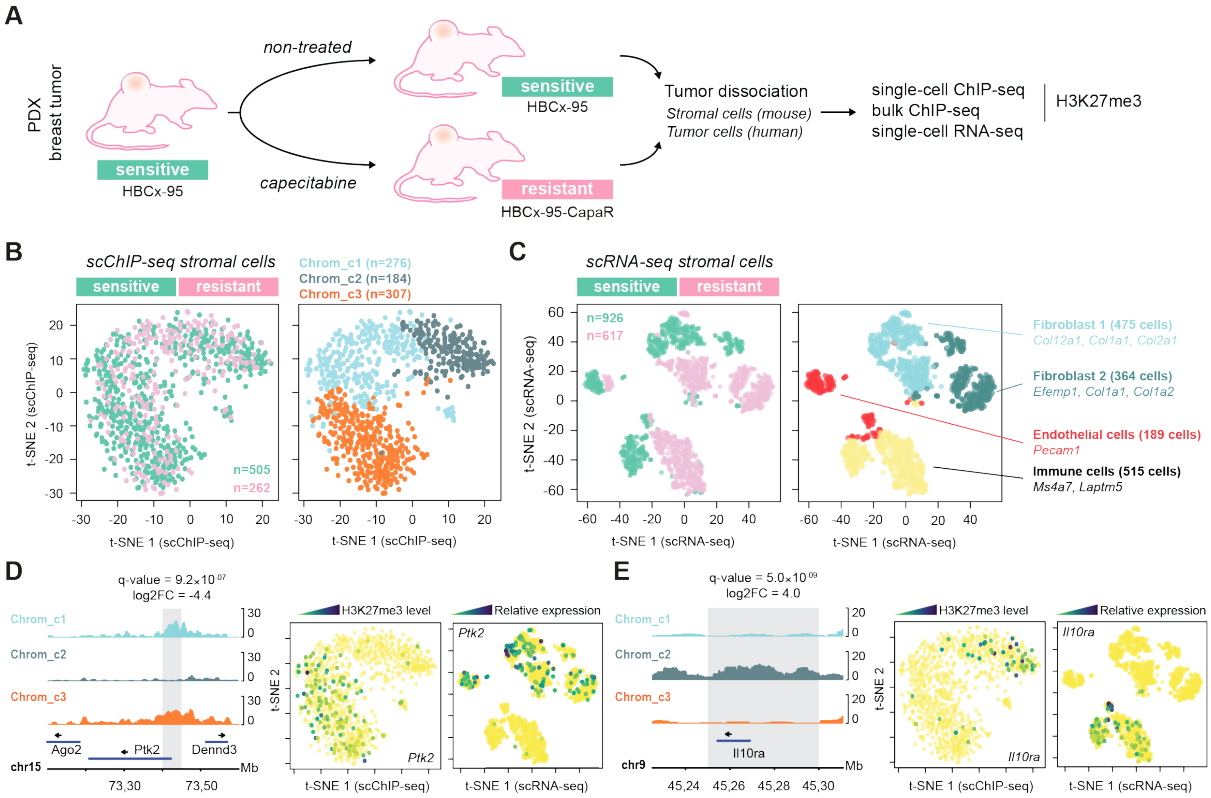


Fig. 2. Single-cell ChIP-seq profiling of mouse stromal cells reveals cell identities from H3K27me3 chromatin traits. (**A**) Cells from sensitive and Capecitabine-resistant breast cancer patient-derived xenografts were processed in parallel to profile H3K27me3 chromatin and transcriptional features at single-cell resolution. (**B**) t-SNE representation of scChIP-seq datasets, colored according to tumor of origin (sensitive or resistant, left panel) or consensus clustering results (fig. S9H). (**C**) t-SNE representation of scRNA-seq datasets, colored according to tumor of origin (sensitive or resistant, left panel) or consensus clustering results (fig. S10B). Marker genes identified by differential expression analysis are indicated for each subpopulation (fig. S10C). (**D-E**) Left panels: snapshots of differentially enriched loci for

Chrom_c3 (*Ptk2*) and Chrom_c2 (*Il10ra*) versus others, with cumulative single-cell profiles for each cell type and cluster. Differentially bound regions identified by Wilcoxon signed-rank test are indicated in grey with the corresponding adjusted p-value and log2 fold change, calculated in respect to Chrom_c2 for both loci. Middle panel: t-SNE representation of scChIP-seq datasets, points are colored according to H3K27me3 enrichment signals in each cell for *Ptk2* and *Il10ra* loci. Right panels: t-SNE representation of scRNA-seq datasets, points are colored according to expression signal for *Ptk2* and *Il10ra* in each cell.

Next, we studied the heterogeneity of chromatin traits among tumor cells from the pair of triple-negative PDX samples. We removed from our analysis loci affected by copy-number variations, as identified from bulk DNA profiles, to focus on potential chromatin-related variations between cells (fig. S12A). Regarding their chromatin and transcriptomic profiles, cells grouped according to their tumor of origin, sensitive or resistant counterpart (Fig. 3A-C and fig. S12B-C), suggesting that the epigenome and transcriptome of resistant cells might have been reprogrammed following exposure to Capecitabine. We could identify distinct chromatin sub-clones within resistant cells (Fig. 3A), suggesting that heterogeneous populations of resistant cells, with distinct chromatin features, might have emerged. Interestingly, consensus clustering (Fig. 3D) shows that 13 cells from the untreated tumor robustly classify with resistant cells of *Chrom_c2*, suggesting they might share chromatin traits with these cells. When comparing chromatin traits between *Chrom_c2* and *Chrom_c1*, we identified 2,352 depleted loci, among which *IGF2BP3* (q-value = 1.1×10^{-20}), a gene known to promote resistance to chemotherapy (*23*) (Fig. 3E-G). A fraction of loci with a loss of H3K27me3 enrichment by HBCx-95-CapaR were not associated to a detection of RNA within the locus (*COL4A1*, *HOXD* cluster, fig. S12D), either due to the absence of transcription activation or the sensitivity of the scRNA-seq procedure.

If no cells from untreated tumor classified with resistant cells according to global transcriptome features, a subgroup of cells from cluster *RNA_c1* (Fig. 3A) show a distinct correlative pattern with resistant cells. Altogether, cells within the untreated tumor display non-genetic traits similar to these of all resistant cells.

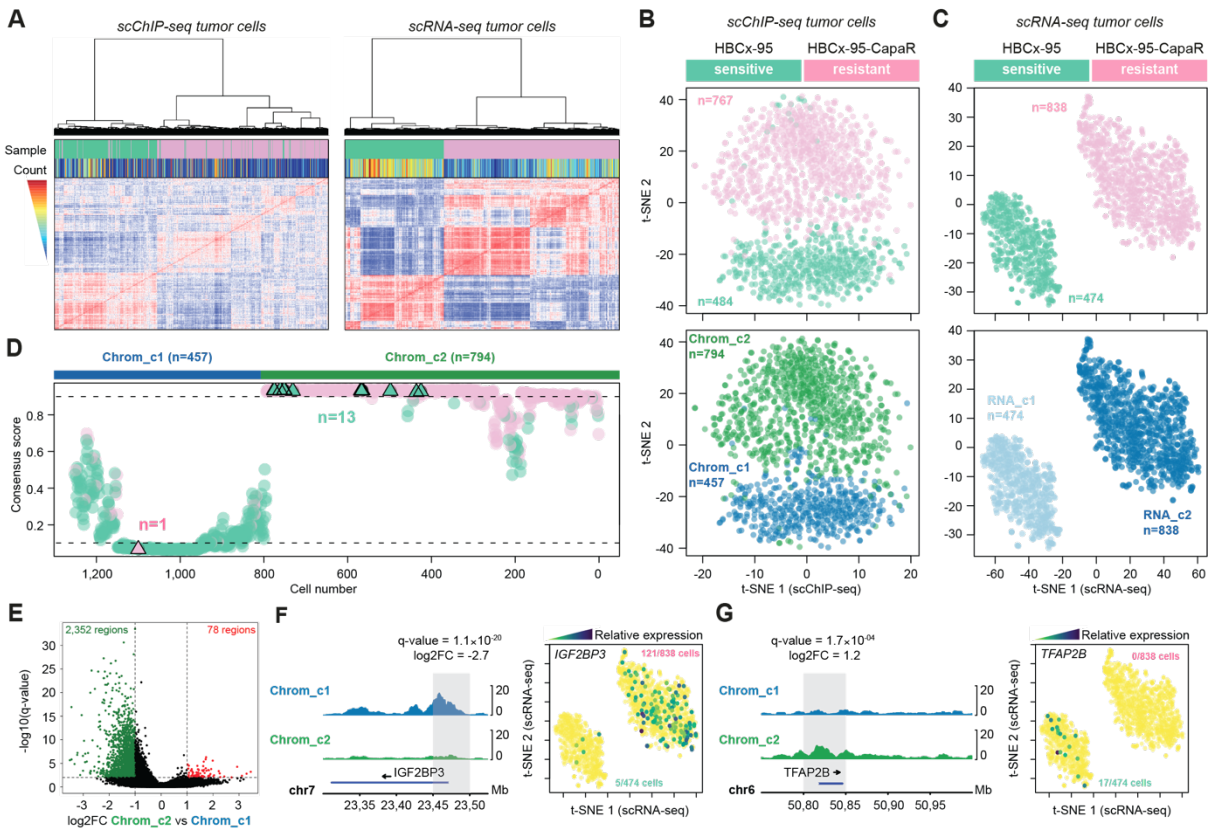


Fig. 3. Single-cell ChIP-seq distinguishes between sensitive and drug-resistance specific H3K27me3 chromatin states. (**A**) Hierarchical clustering and corresponding heatmap of cell to cell Pearson correlation scores. Sample of origin is indicated in green for HBCx-95 and pink for HBCx-95-CapaR, unique read count is indicated above heatmap. (**B-C**) t-SNE representation of scChIP-seq and scRNA-seq datasets, cells colored according to sample of origin and identified cluster (fig. S12C). (**D**) Consensus score of membership to identified clusters, in respect to Chrom_c2. A score of 1 corresponds to a cell as highly representative of Chrom_c2 cluster. (**E**) Volcano plot representing adjusted p-values (Wilcoxon rank's test) versus fold-changes for differential analysis comparing chromatin traits between Chrom_c2 and c1 (thresholds of 0.01 for q-value and 1 for |log2FC|). (**F - G**) Left panels: aggregated H3K27me3 chromatin profiles for each cluster are shown for two top significant loci depleted and enriched in Chrom_c2 (*IGF2BP3* [q-value = 1.1×10^{-20}] and *TFAP2B* [q-value = 1.7×10^{-04}]). Right panels: t-SNE plots representing scRNA-seq datasets, points are colored according to cell expression signal for *IGF2BP3* and *TFAP2B*.

In addition, we profiled a pair of luminal ER+ breast PDX: HBCx-22, responsive to Tamoxifen and HBCx-22-TamR, a tumor derivative with acquired resistance to Tamoxifen (*24*) (n = 822 tumor cells with >1000 unique chromatin traits, average coverage of 10,228 reads per cell). Tumor cells display two major chromatin profiles related to their tumor of

origin, yet 16% (n = 41 out of 255) of cells within sensitive tumor share chromatin traits of all resistant cells (Fig. 4A-D and fig. S13). Differential analysis revealed a massive reshuffling of chromatin states in resistant-like (*Chrom_c2*) cells versus sensitive-like (*Chrom_c1*) cells (Fig.4E, n = 408 loci with q-value < 0.01 and |log2FC| > 1), in agreement with previous reports of chromatin plasticity induced by endocrine therapies (*5*, *25*). One of the top significant chromatin trait lost by resistant-like cells overlaps *EGFR* (q-value = 6.6×10^{-05}), a gene known to be implicated in resistance to Tamoxifen (*26*). We observed that *EGFR* is expressed in resistant cells versus sensitive ones, but also within in a transcriptional sub-clone of HBCx-22 (*RNA_c6*, q-value = 1.1×10^{-05}, Fig. 4C,G-H), suggesting that chromatin and transcriptional traits common to all resistant cells are already found in the sensitive tumor, and could have been selected for by Tamoxifen treatment.

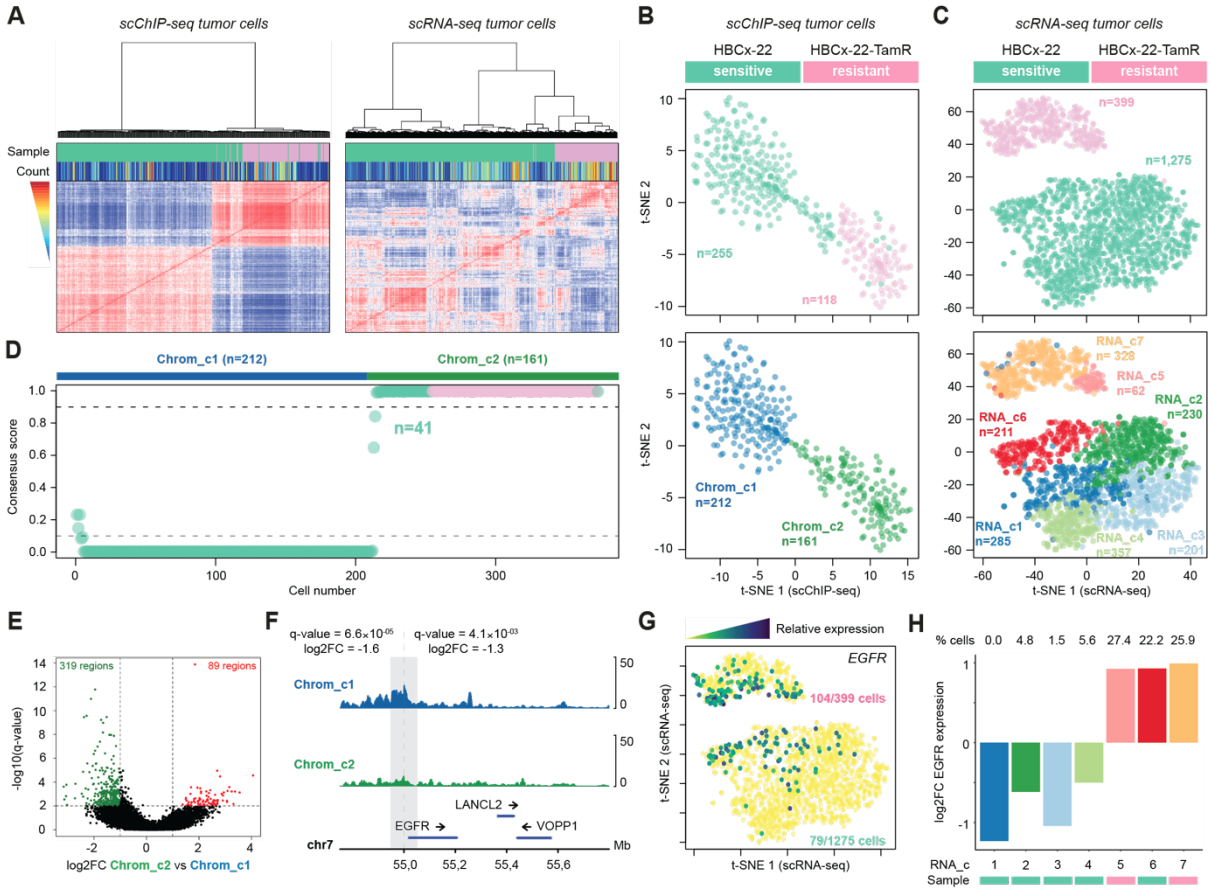


Fig. 4. Single-cell ChIP-seq reveals, within sensitive tumor, a rare population sharing H3K27me3 chromatin traits with resistant cells. (**A**) Hierarchical clustering and corresponding heatmap of cell to cell Pearson correlation scores. Sample of origin is indicated in green for HBCx-22 and pink for HBCx-22-TamR, unique read count is indicated above heatmap. (**B - C**) t-SNE representation of scChIP-seq and scRNA-seq datasets, cells

colored according to sample of origin and identified cluster (fig. S13D). (**D**) Consensus score of membership to identified clusters, in respect to Chrom_c2. A score of 1 corresponds to a cell as highly representative of Chrom_c2 cluster. (**E**) Volcano plot representing adjusted p-values (Wilcoxon rank's test) versus fold-changes for differential analysis comparing chromatin traits between Chrom_c2 and c1 (thresholds of 0.01 for q-value and 1 for |log2FC|). (**F**) Snapshots for *EGFR* locus of aggregated H3K27me3 chromatin profiles for each cluster. For each window, log2 fold-change and adjusted p-value are indicated. (**G**) t-SNE plots representing scRNA-seq datasets, points are colored according to cell expression signal for *EGFR*. (**H**) Barplot displaying the average log2 fold-change in *EGFR* expression level for cells in each cluster versus all remaining cells. The percentage of cells, within each cluster, with detectable *EGFR* expression is indicated above barplot. Sample of origin (green for HBCx-22 and pink for HBCx-22-TamR) is indicated below.

Profiling a chromatin mark at the single-cell level with high coverage was instrumental to reveal the presence of 'epigenetic' clones within complex tumor samples. scChIP-seq represents a unique opportunity to grasp the selection process of chromatin traits and potential epigenetic plasticity (*27*) during tumor evolution. Our parallel study of H3K27me3 layout and transcription points out that 'epigenetic' sub-clones, defined by common H3K27me3 genomic distribution, dot not necessarily fully match 'transcriptional' sub-clones. Loss of such repressive chromatin traits is changing the chromatin to a permissive state, where transcription can happen, and could in this line correspond to a priming event. In the future, combining RNA to chromatin profiling will be decisive to dismantle chromatin from transcriptional plasticity not only in tumors but also during normal differentiation.

References and Notes:

1. S. Nik-Zainal *et al.*, Mutational processes molding the genomes of 21 breast cancers. *Cell*. **149**, 979–993 (2012).
2. P. Eirew *et al.*, Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*. **518**, 422–426 (2015).
3. M. W. Schmitt, L. A. Loeb, J. J. Salk, The influence of subclonal resistance mutations on targeted cancer therapy. *Nat. Rev. Clin. Oncol.* **13** (2016), pp. 335–347.
4. P. Rathert *et al.*, Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. *Nature*. **525**, 543–547 (2015).
5. L. Magnani *et al.*, Genome-wide reprogramming of the chromatin landscape underlies endocrine therapy resistance in breast cancer. *Proc. Natl. Acad. Sci.* **110**, E1490–E1499 (2013).
6. W. Hugo *et al.*, Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*. **165**, 35–44 (2016).
7. E. E. Gardner *et al.*, Chemosensitive Relapse in Small Cell Lung Cancer Proceeds through an EZH2-SLFN11 Axis. *Cancer Cell*. **31**, 286–299 (2017).
8. S. V Puram *et al.*, Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*. **171**, 1611–1624.e24 (2017).
9. W. Chung *et al.*, Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 (2017).
10. C. Kim *et al.*, Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell*. **173**, 879–893 (2018).
11. T. Mazor *et al.*, DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. *Cancer Cell*. **28**, 307–317 (2015).
12. Y. Hao, Y. Cui, X. Gu, Genome-wide DNA methylation profiles changes associated with constant heat stress in pigs as measured by bisulfite sequencing. *Sci. Rep.* **6**, 27507 (2016).
13. M. J. Aryee *et al.*, DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases. *Sci. Transl. Med.* **5** (2013), doi:10.1126/scitranslmed.3005211.
14. S. A. Smallwood *et al.*, Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*. **11**, 817–820 (2014).
15. C. Angermueller *et al.*, Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*. **13** (2016), doi:10.1038/nmeth.3728.
16. Y. Hou *et al.*, Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304–319 (2016).

17. D. A. Cusanovich *et al.*, Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-.).* **348**, 910–914 (2015).

18. J. D. Buenrostro *et al.*, Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. **523**, 486–490 (2015).

19. M. R. Corces *et al.*, Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).

20. T. Nagano *et al.*, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. **502**, 59–64 (2013).

21. A. Rotem *et al.*, Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).

22. E. Marangoni *et al.*, *Clin. Cancer Res.*, in press, doi:10.1158/1078-0432.CCR-17-3490.

23. M. Lederer, N. Bley, C. Schleifer, S. Hüttelmaier, The role of the oncofetal IGF2 mRNA-binding protein 3 (IGF2BP3) in cancer. *Semin. Cancer Biol.* **29** (2014), pp. 3–12.

24. P. Cottu *et al.*, Acquired resistance to endocrine treatments is associated with tumor-specific molecular changes in patient-derived luminal breast cancer xenografts. *Clin. Cancer Res.* **20**, 4314–4325 (2014).

25. V. T. M. Nguyen *et al.*, Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. *Nat. Commun.* **6**, 10044 (2015).

26. S. Massarweh *et al.*, Tamoxifen resistance in breast tumors is driven by growth factor receptor signaling with repression of classic estrogen receptor genomic function. *Cancer Res.* **68**, 826–833 (2008).

27. W. A. Flavahan, E. Gaskell, B. E. Bernstein, Epigenetic plasticity and the hallmarks of cancer. *Science (80-.).* **357** (2017), doi:10.1126/science.aal2380.

Acknowledgments: We would like to thank Bradley Bernstein (Massachusetts General Hospital, Broad Institute, Cambridge MA, USA) for advice on the technology development; Robert Nicol (Whitehead Institute, MIT Center for Genome Research, Cambridge MA, USA) for advice on barcoded sequencing; the Institut Pierre-Gilles de Gennes (IPGG) for use of the clean room facilities. We would also like to thank patients for their approval for the use and sequencing of tumor samples.

Funding:

A.D.G was supported by the by the French "Investissements d'Avenir" program via grant agreements ANR-10-NANO-02, ANR-10-IDEX-0001-02 PSL, ANR-10-LABX-31 and ANR-10-EQPX-34, by the French Agence Nationale de la Recherche (ANR), project CellectChIP. CV was supported by the ATIP Avenir program, Plan Cancer and by the SiRIC-Curie program (#INCa-DGOS-Inserm_12554). NGS was performed by the ICGex platform of the Institut Curie (Paris) and the iGenSeq platform of the Institut du Cerveau et de la Moelle épinière (Paris). High-throughput sequencing has been performed by the ICGex. NGS platform of the Institut Curie is supported by the grants ANR-10-EQPX-03 (Equipex) and ANR-10-INBS-09-08 (France Génomique Consortium) from the Agence Nationale de la Recherche ("Investissements d'Avenir" program), by the Cancéropôle Ile-de-France and by the SiRIC-Curie program -SiRIC Grant INCa-DGOS- 4654.

Author contributions: K.G., C.B., A.D.G., C.V. and A.G. conceived the method and designed the experiments. K.G., A.D., F.N., A.D., A.P. and O.F. conducted experiments. PDX experiment was designed by F.R. and E.M.; Y.P. performed barcoded beads preparation and quality controls, M.R. contributed to microfluidic chip design. K.G., C.V. and A.W. performed data analysis. K.G., A.D.G., C.V. and A.G. wrote the manuscript with input from all authors.

Competing interests: Authors declare no competing interests.

Data and materials availability: All relevant sequencing files were deposited to GEO

Supplementary Figures & Tables for

Single-cell chromatin profiling reveals heterogeneity of chromatin states in breast cancer

Kevin Grosselin[1,2], Adeline Durand[3], Adeline Poitou[1], Elisabetta Marangoni[4], Farida Nemati[4], Ahmed Dahmani[4], Fabien Reyal[4,5,6], Olivia Frenoy[1], Yannick Pousse[1], Marcel Reichen[1], Adam Woolfe[1], Colin Brenan[1,7], Andrew D. Griffiths[2]*, Céline Vallot[3,4]*, Annabelle Gérard[1]*

Correspondence to: andrew.griffiths@espci.fr; celine.vallot@curie.fr; a.gerard@hifibio.com

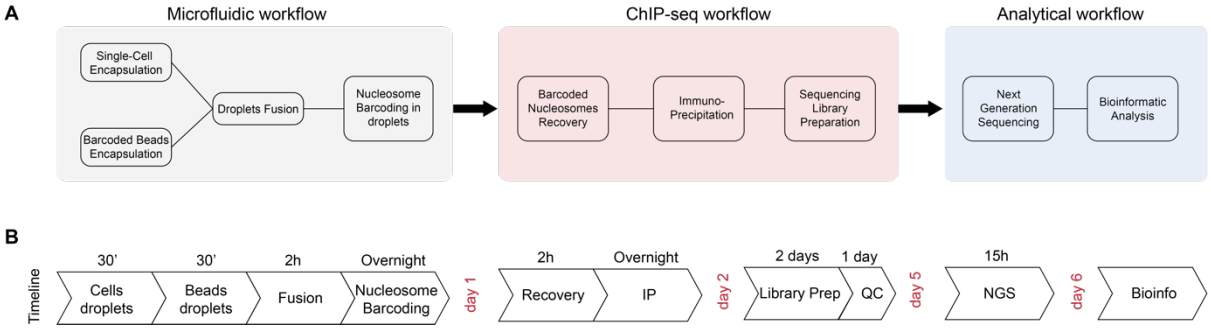


Fig. S1. Overview of the single-cell ChIP-seq procedure. (**A**) The single-cell ChIP-seq procedure is divided into 3 workflows including a droplet-microfluidic workflow, a ChIP-seq workflow and an analytical workflow. Key steps of each workflow are indicated in the black boxes. (**B**) Single-cell ChIP-seq timeline. The entire process from the cell harvesting to the NGS takes 5 days.

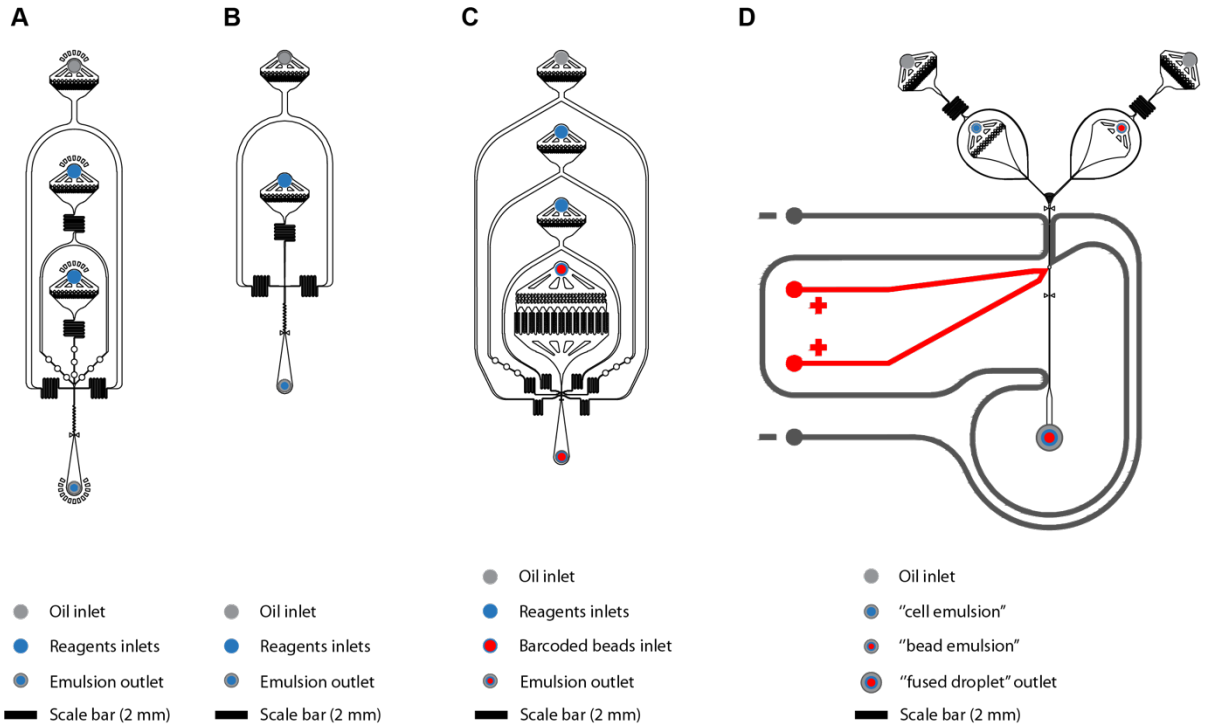


Fig. S2. Design of microfluidic devices. (**A**) Device used to compartmentalize cells in 45 pl droplets. (**B**) Device used to produce 9 pl droplets to make hydrogel beads. (**C**) Device used for compartmentalization of hydrogel beads in 100 pl droplets. (**D**) Device used to merge nucleosome-containing droplets with hydrogel barcoded bead-containing droplets. Potential and ground electrodes are indicated with the '+' and the '-' marks, respectively. Scale bars are 2 mm.

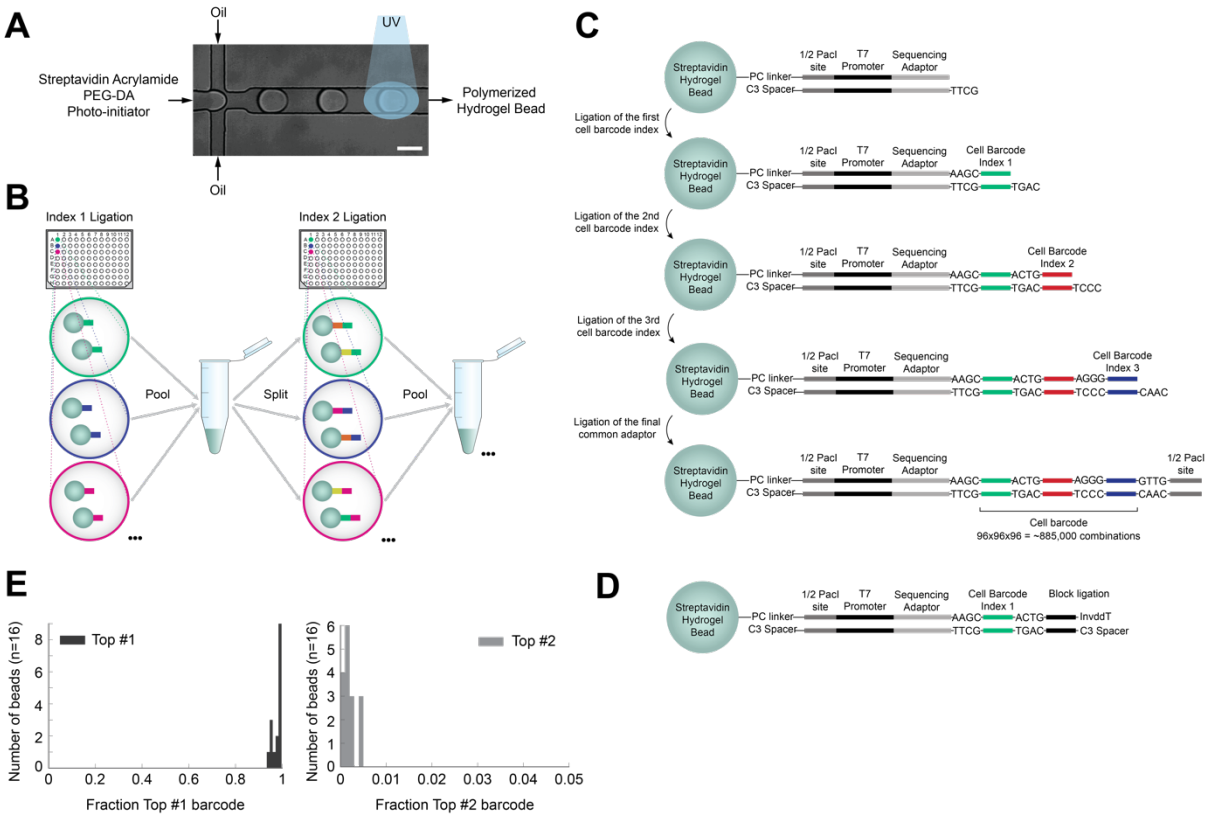


Fig. S3. Barcoded beads production and quality control. (**A**) Beads were produced in a 2-inlets microfluidic device by dispersing a mixture comprising PolyEthylene Glycol Di-Acrydrite (PEG-DA), Streptavidin Acrylamide and the photo-initiator. Flow rates were adjusted to produce 9 pl droplets, immediately exposed to UV light for polymerization of the hydrogel network (see Material and Methods). Scale bar is 25 µm (**B**) Split-and-pool synthesis principle. (**C**) Barcodes were synthesized by successive ligation of double-stranded indexes containing 5' overhang of 4 base pairs by three rounds of split-and-pool synthesis using 96 Index 1, 96 Index 2 and 96 Index 3. Barcodes were flanked at one end by common sequences comprising ½ Pac1 restriction site, the T7 promoter and the Illumina Read #2 sequencing primer, which were bound to the beads via a photocleavable linker (PC-linker). A 3' C3-spacer was added to the 3'end of the photocleaved site for directed ligation to the other end of the barcode comprising a second common sequence with ½ Pac1 restriction site ligated to the index 3. (**D**) Barcodes that failed in one of the three split-pool rounds were completed with a "block" oligonucleotide comprising a 5' C3-spacer and a 3' Inverted ddT to prevent ligation. (**E**) Single-bead sequencing results showing the fraction of the first two most abundant barcodes of 16 beads. In average, 97.7% of the barcodes corresponded to the same sequence and the second most abundant barcode represented 0.17% of all sequences.

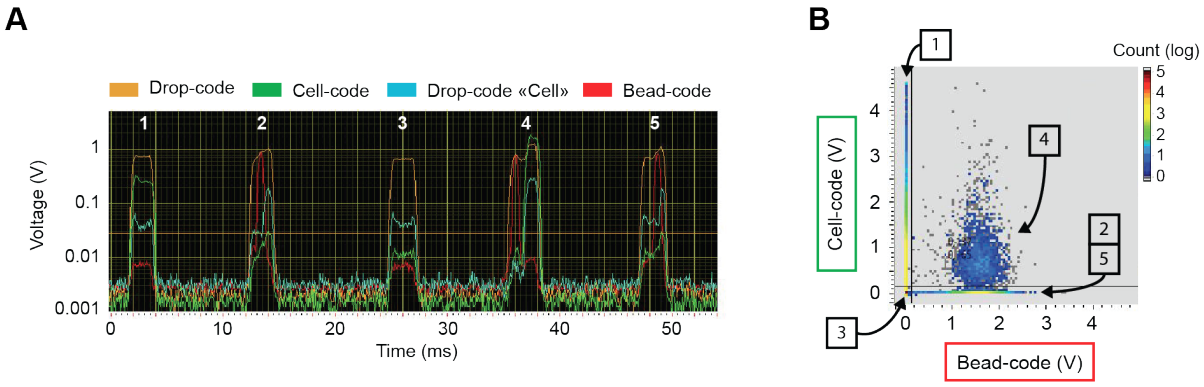


Fig. S4. Live monitoring of cells and barcoded hydrogel beads co-encapsulation in droplet microfluidic system. Droplets were scanned after fusion as they crossed a laser beam and their fluorescence intensity analyzed in real time. Sulforhodamine B (orange fluorescent dye) was used as common marker (drop-code) and Dye-405 was used as marker of the nucleosome-containing droplets (drop-code cell). Cells and beads were labeled with calcein AM cell-permeant dye (green fluorescence) and biotin-Cy5 (red fluorescence) respectively. (**A**) Time traces of 5 droplets showing all possible fusion events. Droplets #4 contained both one cell plus one bead, resulting in nucleosomes barcoding. (**B**) Scatter plot showing cell fluorescence intensity (green) versus bead fluorescence intensity (red) in each droplet allows precise counting of the number of cells co-encapsulated with a barcoded bead. Droplets from panel (A) are indicated as examples of the different droplet populations.

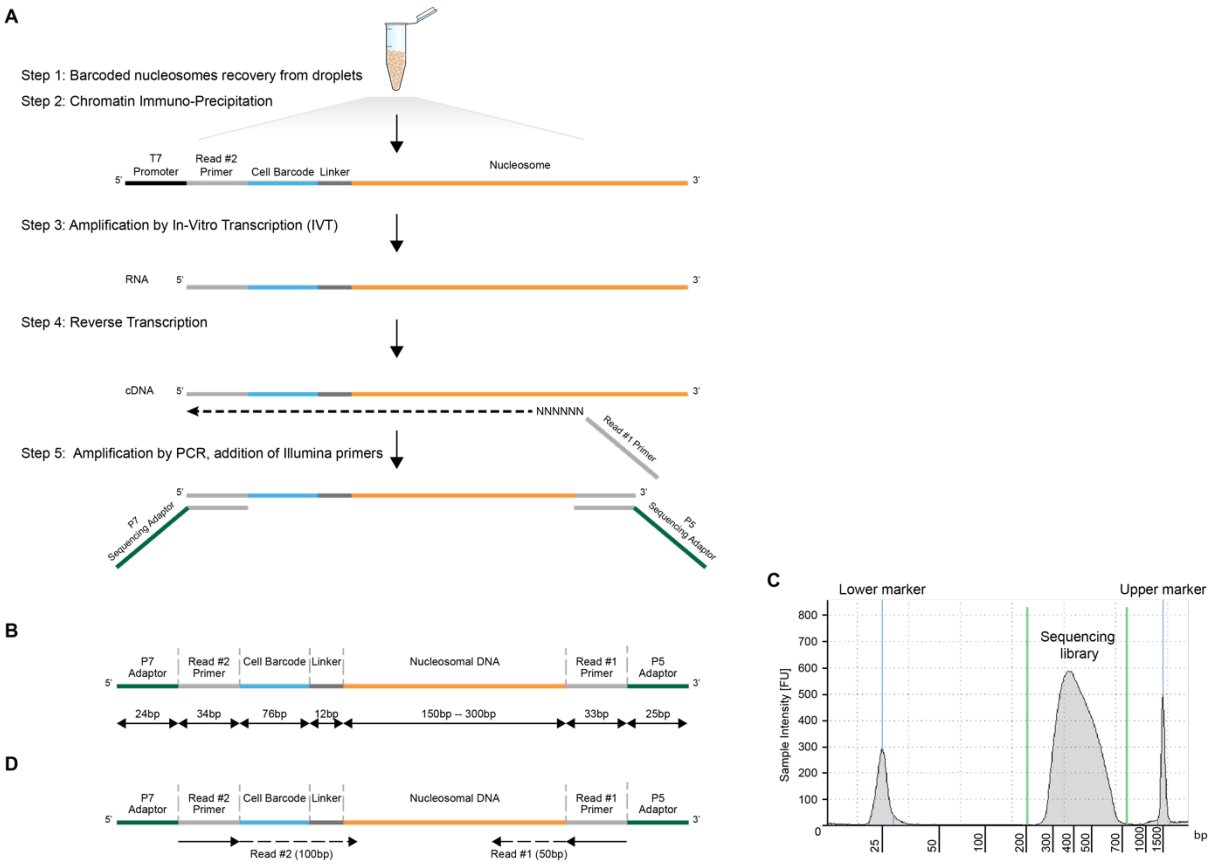


Fig. S5. Sequencing library preparation. (**A**) Enriched barcoded nucleosomes were linearly amplified in in-vitro transcription reaction. The amplified RNA were reverse transcribed into cDNA by random priming, appending the reverse complement of Illumina Read #1 sequencing primer. cDNA were amplified by PCR, appending the Illumina P7 and P5 sequences. (**B**) Schematic of the final sequencing product with size in bp of each element" constituting the sequence. (**C**) Electropherogram showing the size distribution of the final sequencing library. The smear ranging from 300 bp to 700 bp and corresponds to barcoded nucleosomes (profile obtained by Tapestation). (**D**) Single-cell ChIP-seq libraries were sequenced as follows: 50 bp were assigned to read the nucléosomal sequence and 100 bp were assigned to read the barcode.

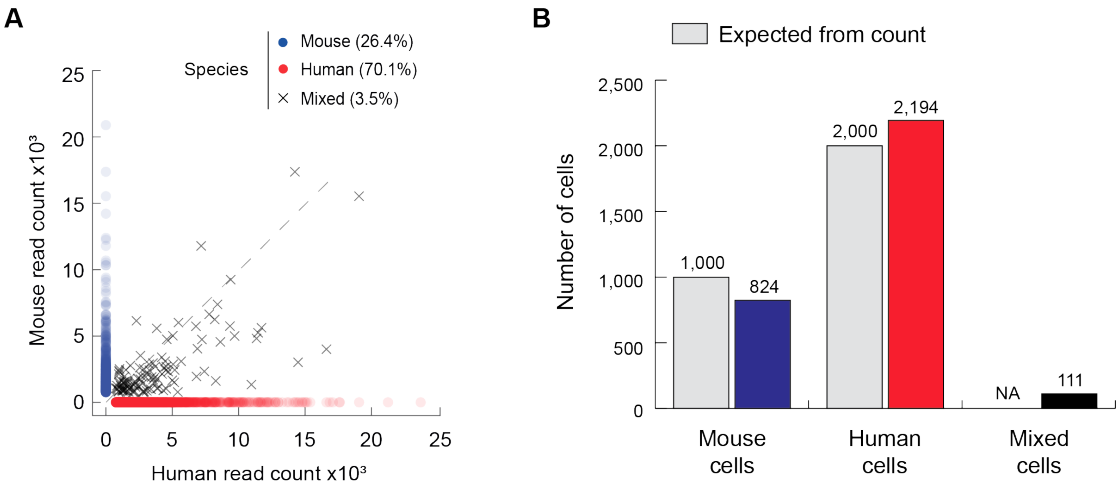


Fig. S6. Specie mixing experiment reveals specific mapping and single-cell resolution. (**A**) Scatter plot of number of reads per barcode aligning to the mouse versus human reference genome showing that 96.5% of the barcodes are specific to one specie (at least 95% of the reads mapping to one of the two specie). (**B**) Barplot showing the number of barcodes identified for each specie in comparison with the expected number of cells counted on the microfluidic station (3,000 in total from a mixture comprising 1/3 mouse cells and 2/3 of human cells).

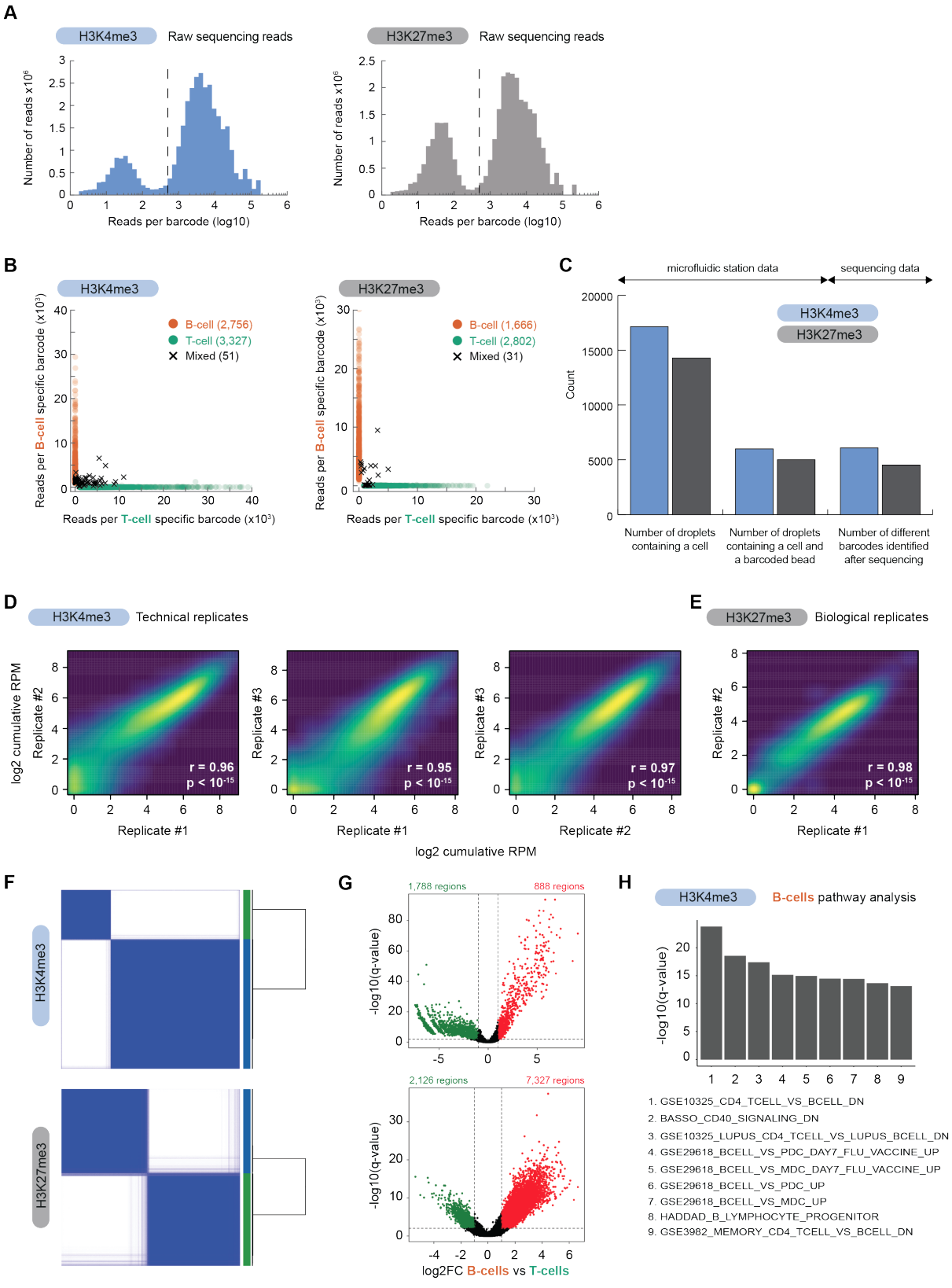


Fig. S7. Quality controls of single-cell ChIP-seq data. (**A**) Histograms of the distribution of scChIP-seq raw sequencing reads per cell in H3K4me3 and H3K27me3 single-cell ChIP-seq datasets. A threshold of 500 reads (dotted line) was applied to eliminate low read count

barcodes. (**B**) Scatter plot showing the number of raw reads aligned on B-cell specific barcode set versus T-cell specific barcode set. (**C**) Barplot showing the total number of cells and the number of cells co-encapsulated with a barcoded hydrogel bead detected by fluorescence on the microfluidic station in H3K4me3 and H3K27me3 single-cell ChIP-seq experiments. Analysis of sequencing data showed a number of identified barcodes closely related to the number of droplets containing both a cell and a bead counted on the microfluidic station, suggesting high overall efficacy of the system. (**D**) Technical replicates corresponded to three independent fractions of the same emulsion collected and processed in parallel. Correlation between replicates is calculated based on the cumulative count per million reads in 5 kb genomic bins across single-cells. Pearson correlation scores and p-values are computed genome-wide. (**E**) Biological replicates corresponded to two emulsions collected from different cell culture flasks and processed with different batches of barcoded hydrogel beads. Correlation between replicates is calculated based on the cumulative count per million reads in 50 kb genomic bins across single-cells. Pearson correlation score and p-value are computed genome-wide. (**F**) Hierarchical clustering and corresponding heatmap of cell to cell consensus clustering score for H3K4me3 (top panel) and H3K27me3 (bottom panel) scChIP-seq datasets. Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded and indicated between the heatmap and the dendrogram. (**G**) Volcano plot representing adjusted p-values (Wilcoxon rank's test) versus fold-changes for differential analysis comparing chromatin traits between B-cells and T-cells (thresholds of 0.01 for q-value and 1 for |log2FC|) for H3K4me3 (top panel) and H3K27me3 (bottom panel) scChIP-seq datasets. (**H**) Barplot displaying the log10 of adjusted p-values from pathway analysis in relation to B-cells in H3K4me3 scChIP-seq dataset. Most significant gene sets are indicated below the barplot.

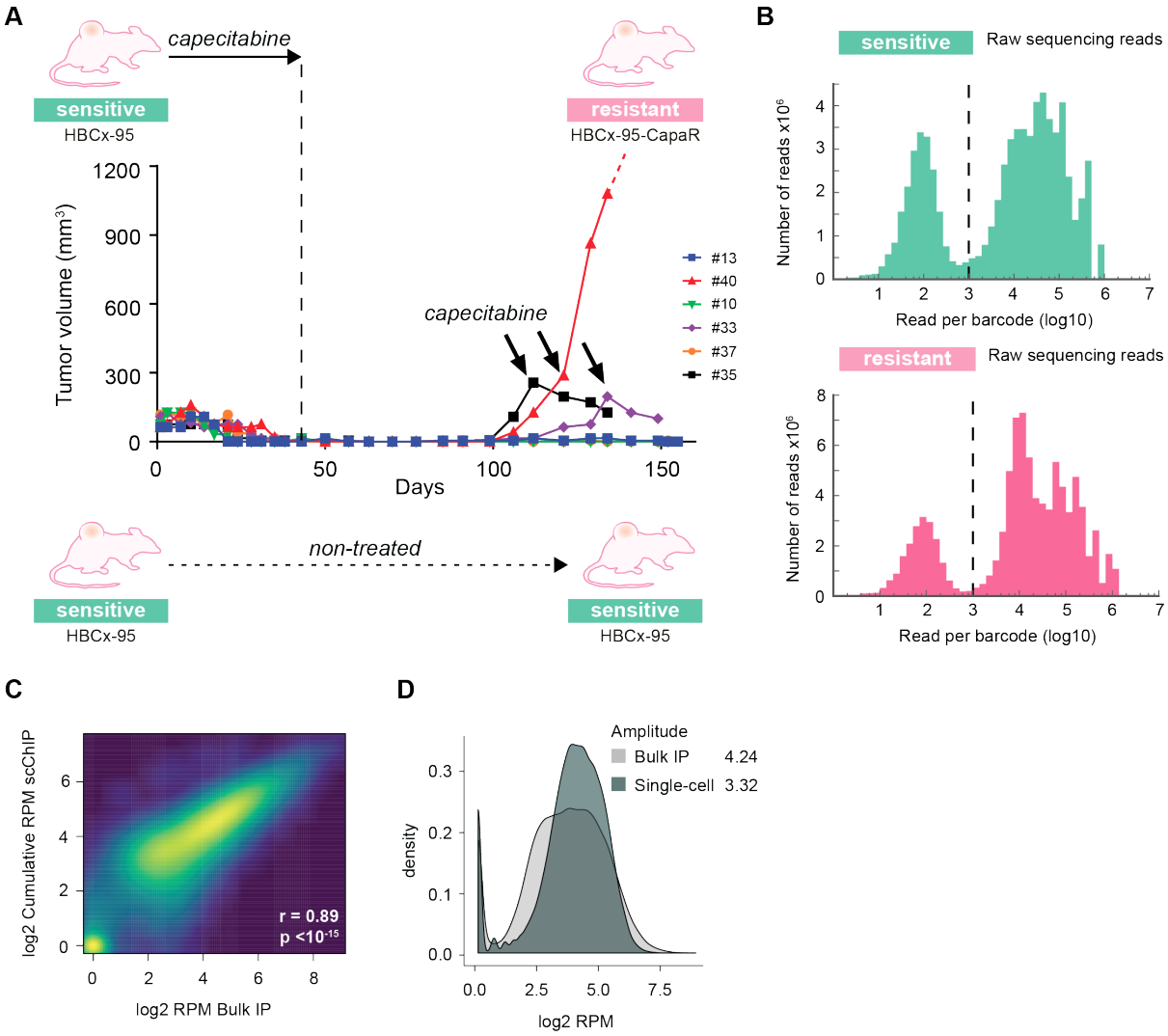


Fig. S8. Breast tumor model of acquired resistance to Capecitabine. (**A**) PDX mice responsive to Capecitabine (HBCx-95) were treated continuously for 50 days. Tumor relapses were treatment again with Capecitabine and mice that failed to respond were selected for the study (HBCx-95-CapaR). (**B**) Histograms of the distribution of scChIP-seq raw sequencing reads per cell in untreated HBCx-95 and Capecitabine-resistant HBCx-95-CapaR PDX. A threshold of 1,000 reads (dotted line) was applied to eliminate low read count barcodes. (**C**) Scatter plot displaying the log2 enrichments in cumulative single-cell versus bulk H3K27me3 ChIP-seq data, calculated based on the count per million of mapped reads in 50 kb genomic bins. Pearson correlation score and p-value are computed genome-wide. (**D**) Distribution of the log2 enrichments in 50 kb genomic bins for cumulative single-cell and bulk H3K27me3 ChIP-seq data.

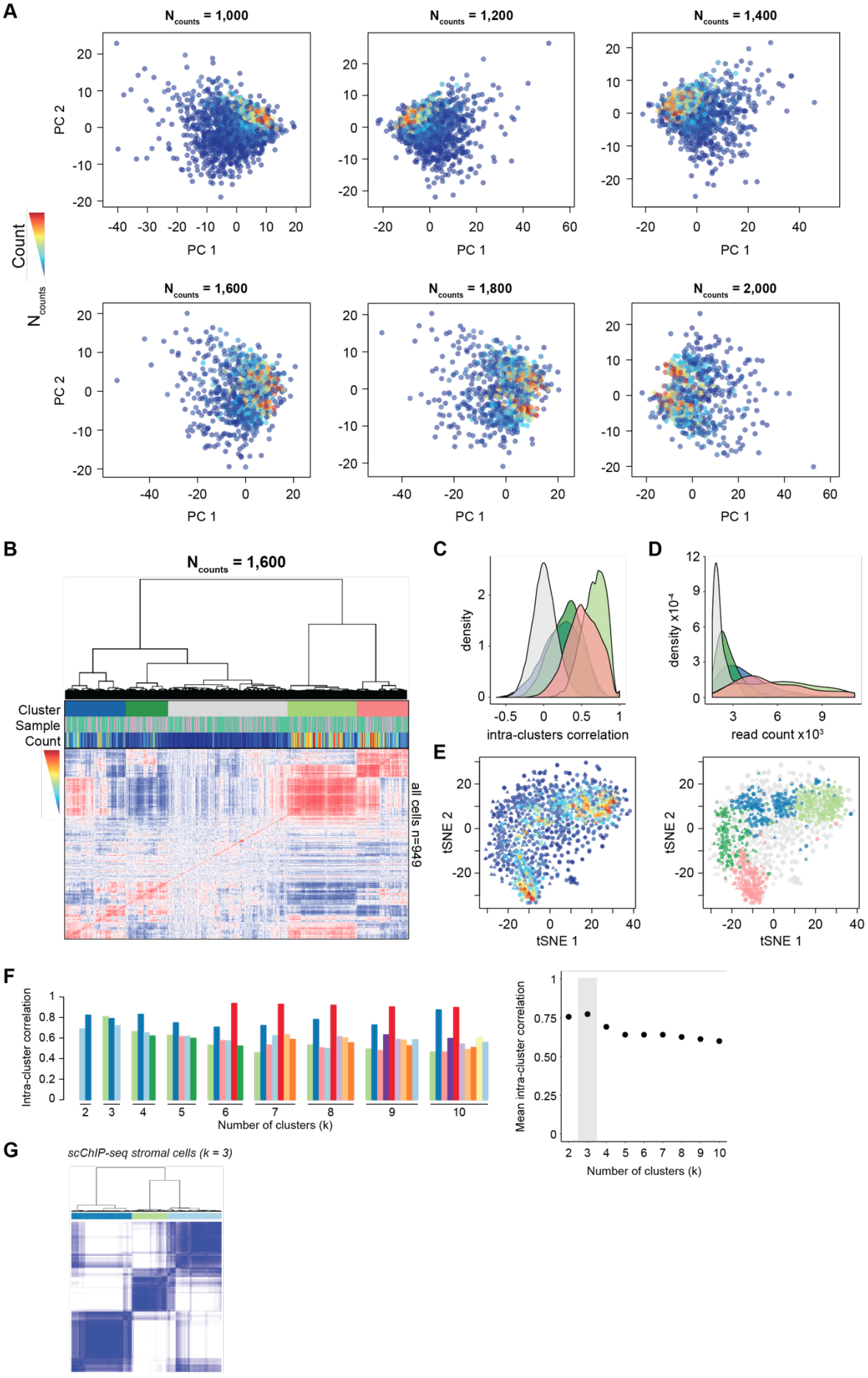

A
N_counts = 1,000
N_counts = 1,200
N_counts = 1,400
N_counts = 1,600
N_counts = 1,800
N_counts = 2,000
PC 1
PC 2
Count
N_counts
B
N_counts = 1,600
Cluster
Sample
Count
all cells n=949
C
density
intra-clusters correlation
D
density x10⁻⁴
read count x10³
E
tSNE 1
tSNE 2
F
Intra-cluster correlation
Number of clusters (k)
Mean intra-cluster correlation
G
scChIP-seq stromal cells (k = 3)

Fig. S9. Reliable identification of subpopulation is directly related to single-cell coverage. (**A**) PCA plots with varying minimum number of unique mapped reads per cell showing that PCA were mainly driven by cell coverage up to 1,600 reads per cell. (**B - D**) Correlation matrix of 949 mouse stromal cells with a minimum coverage of 1,600 unique mapped reads and based on their H3K27me3 profiles. Correlation-based clustering and the distribution of intra-cluster correlation score revealed a population of non-correlated cells (light grey cluster). These cells were characterized with lower cell coverage compared to other clusters. (**B**) Hierarchical clustering and corresponding heatmap of cell to cell Pearson correlation scores. Cluster membership from hierarchical clustering, sample of origin (green for HBCx-95 and pink for HBCx-95-CapaR) and unique read count are indicated above heatmap. (**C**) Distribution of the intra-cluster correlation scores for each cluster identified by hierarchical clustering. (**D**) Distribution of the number of reads per cluster identified by hierarchical clustering. (**E**) t-SNE plots representing scChIP-seq datasets, points are colored according to cell coverage and cluster membership. Non-correlated cells (light grey) are uniformly distributed within all stable clusters and removed for analysis. (**F**) Left: mean of all pairwise correlation score between cluster's members is plotted for *k* clusters ranging from 2 to 10. Right: mean intra-cluster correlation score for k clusters ranging from 2 to 10. At $k = \mathbf{3}$ clusters, the intra-cluster correlation is maximized. (**G**) Hierarchical clustering and corresponding heatmap of cell to cell consensus clustering score for scChIP-seq stromal cells. Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above heatmap.

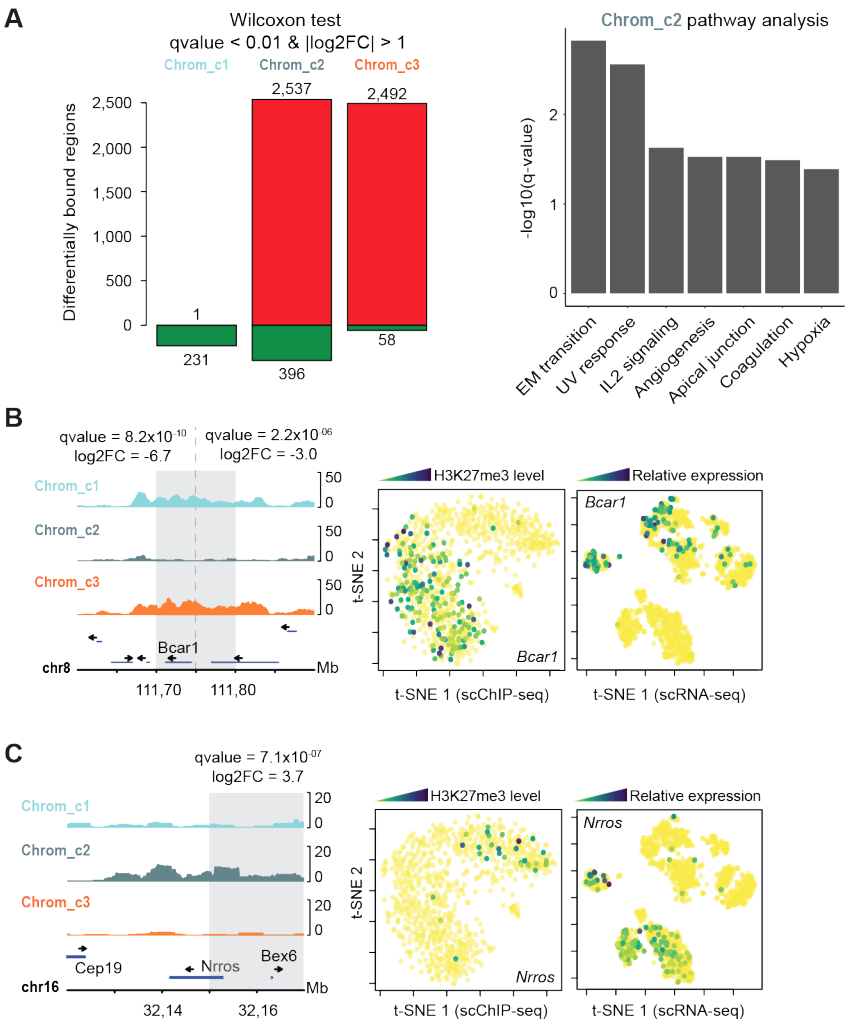


Fig. S10. Supervised analysis of single-cell ChIP-seq H3K27me3 profiles of mouse stromal cells. (**A**) Left: differentially bound regions identified by Wilcoxon signed-rank test. Genomic regions were considered enriched (red) or depleted (green) if the adjusted p-values were lower than 0.01 and the fold change greater than 2. Right: barplot displaying the log10 of adjusted p-values from pathway analysis in relation to Chrom_c2. (**B - C**) Left panels: snapshots of differentially enriched loci for Chrom_c3 (*Bacr1*) and Chrom_c2 (*Nrros*) versus others, with cumulative single-cell profiles for each cell type and cluster. Differentially bound regions identified by Wilcoxon signed-rank test are indicated in grey with the corresponding adjusted p-value and log2 fold change, calculated in respect to Chrom_c2 for both loci. Middle panel: t-SNE representation of scChIP-seq datasets, points are colored according to H3K27me3 enrichment signals in each cell for *Bcar1* and *Nrros* loci. Right panels: t-SNE representation of scRNA-seq datasets, points are colored according to expression signal for *Bcar1* and *Nrros* in each cell.

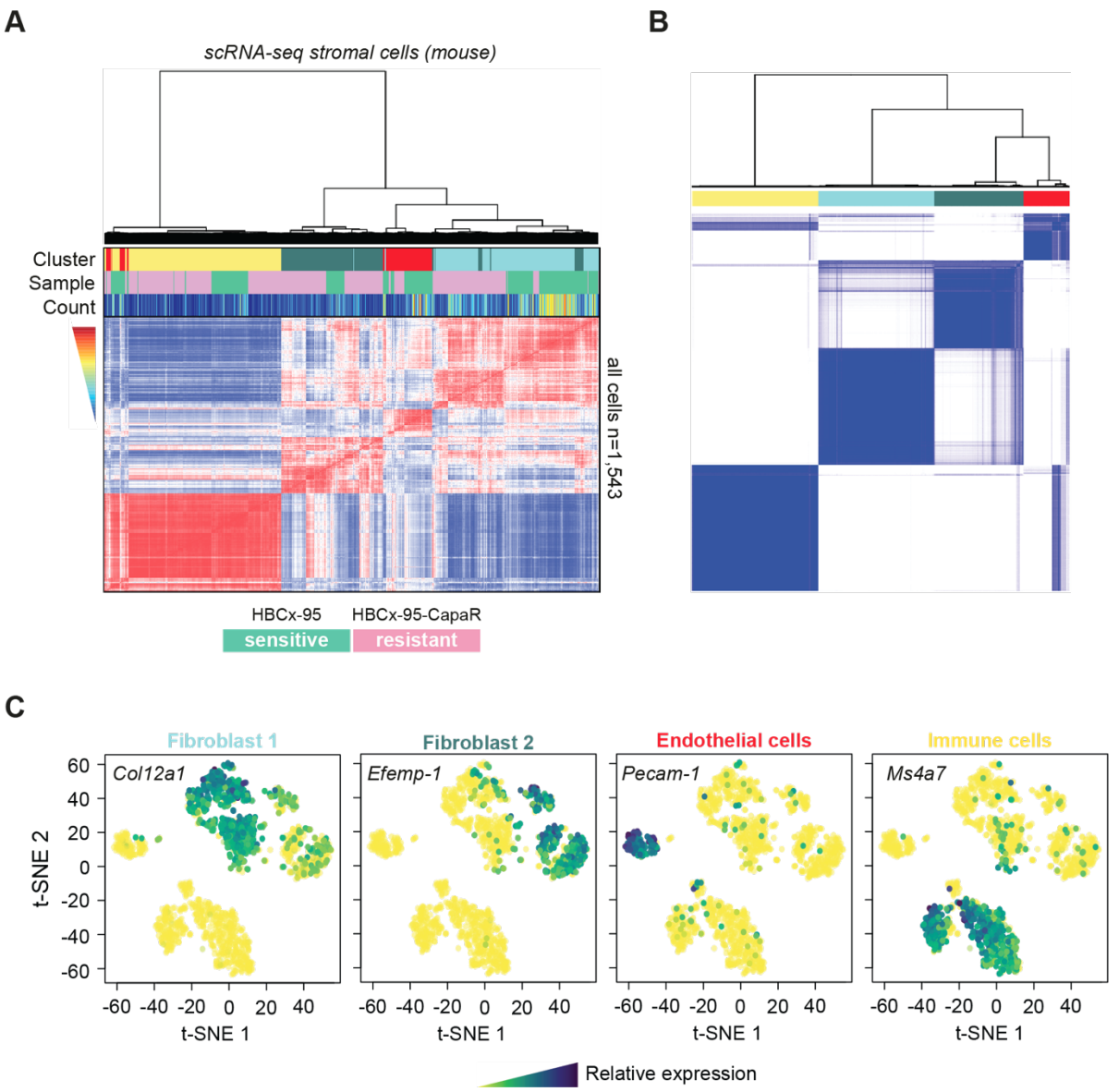


Fig. S11. Supervised analysis of single-cell RNA-seq of mouse stromal cells. (**A**) Hierarchical clustering and corresponding heatmap of cell to cell Pearson correlation scores. Cluster membership from consensus clustering, sample of origin (green for HBCx-95 and pink for HBCx-95-CapaR) and unique read count are indicated above heatmap. (**B**) Hierarchical clustering and corresponding heatmap of cell to cell consensus clustering score for scRNA-seq stromal cells. Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above heatmap. (**C**) t-SNE plots showing expression signal of marker genes for each cell types identified by differential gene expression analysis.

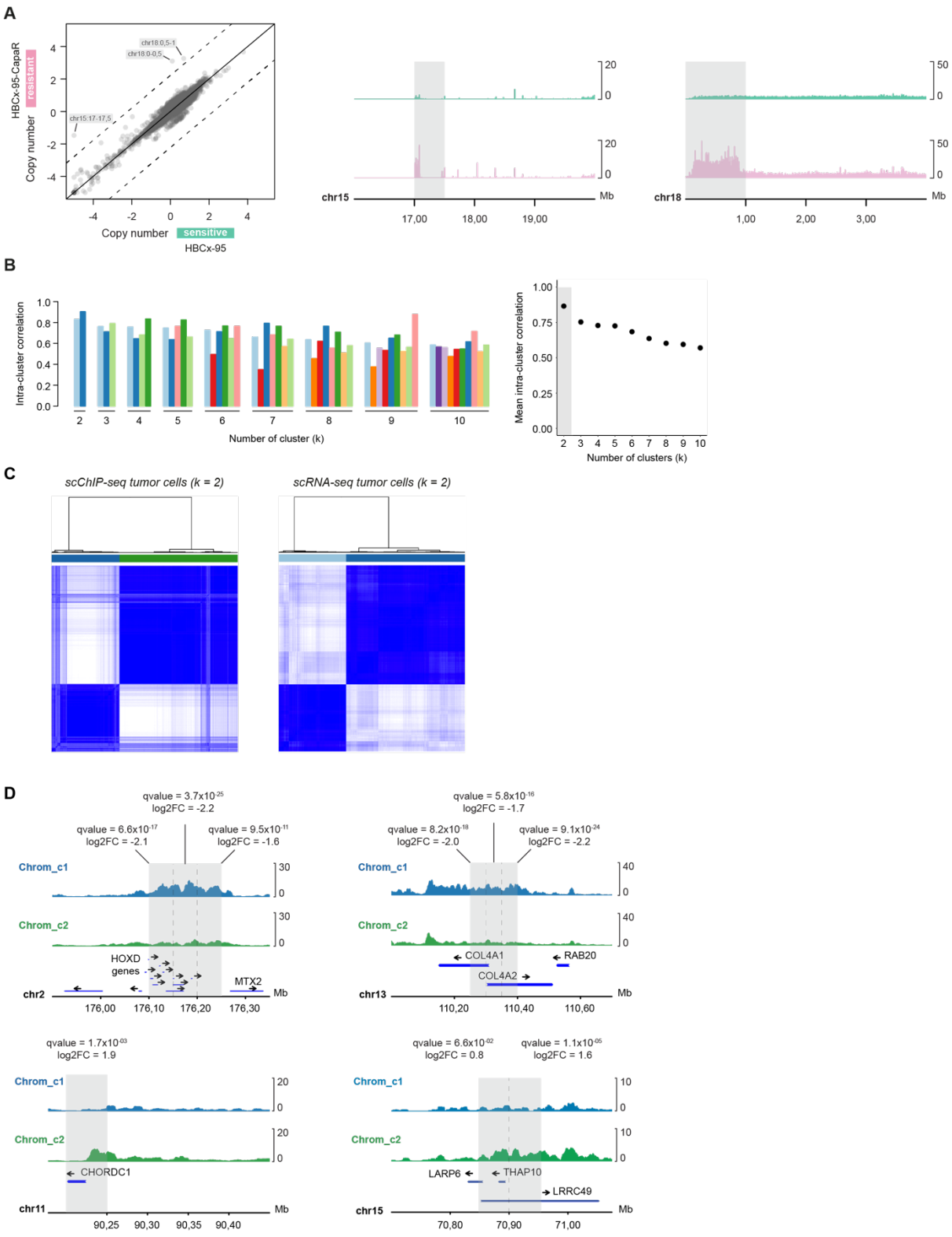


Fig. S12. Clustering of single-cell ChIP-seq profiles of human tumor cells (HBCx-95 and HBCx-95-CapaR PDXs). (**A**) Left: Copy number in 0.5 Mb non-overlapping regions plotted for bulk DNA profiles of Capecitabine-resistant PDX (HBCx-95-CapaR) versus untreated PDX (HBCx-95). Right: snapshots of loci affected by copy number variation for bulk DNA profiles of Capecitabine-resistance PDX and untreated PDX indicated in grey. (**B**) Left: mean

of all pairwise correlation score between cluster's members is plotted for *k* clusters ranging from 2 to 10. Right: mean intra-cluster correlation score for k clusters ranging from 2 to 10. At *k* = 2 clusters, the intra-cluster correlation is maximized. **(C)** Hierarchical clustering and corresponding heatmap of cell to cell consensus clustering score for scChIP-seq and scRNA-seq tumor cells (HBCx-95 and HBCx-95-CapaR PDXs). Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above heatmap. (**D**) Aggregated H3K27me3 chromatin profiles for Chrom_c1 and Chrom_c2 are shown for top significant loci identified by differential analysis. For each window indicated in grey, log2 fold-change and adjusted p-value are indicated.

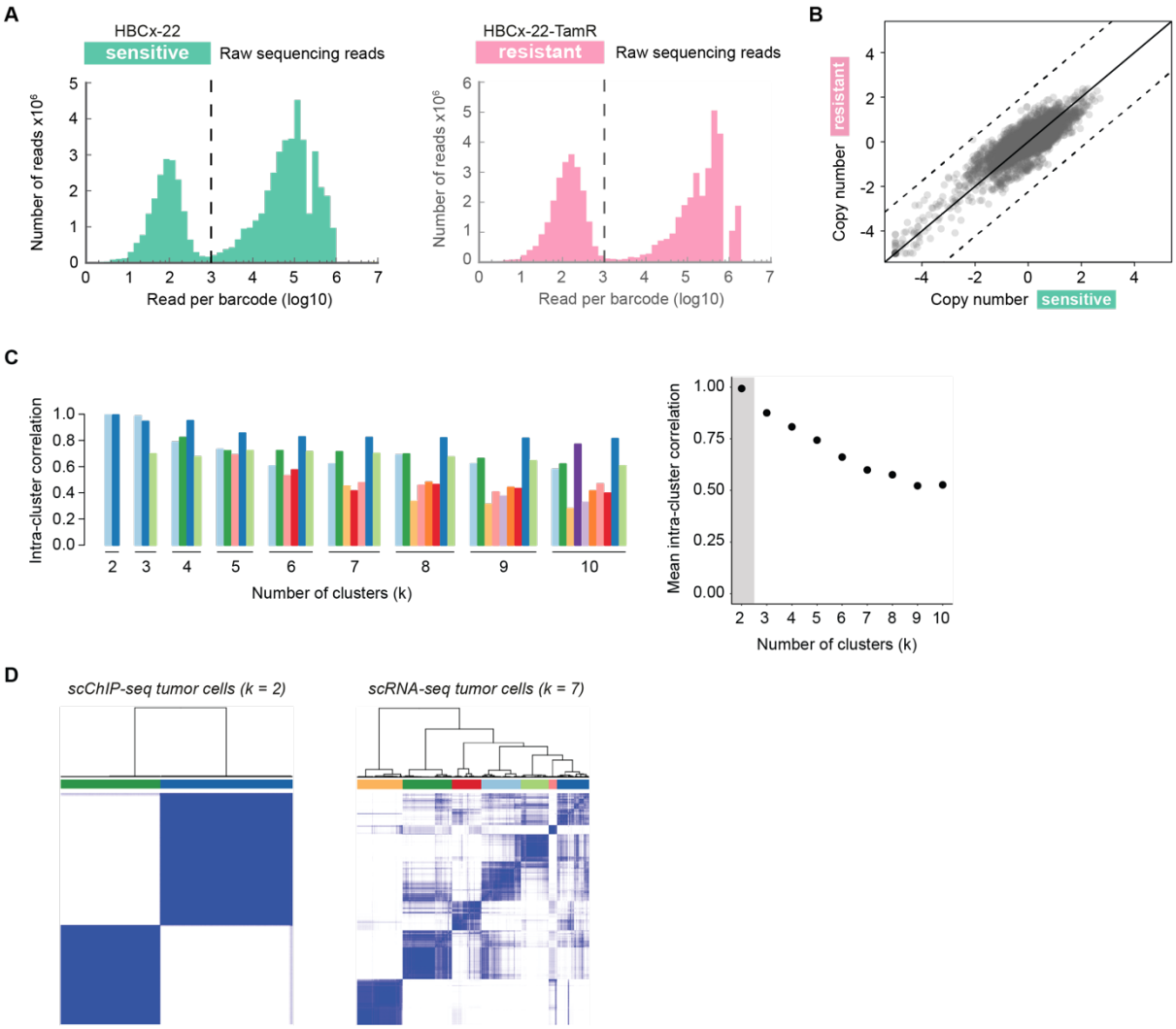


Fig. S13. Clustering of single-cell ChIP-seq profiles of human tumor cells (HBCx-22 and HBCx-22-TamR PDXs). (**A**) Histograms of the distribution of scChIP-seq raw sequencing reads per cell in untreated HBCx-22 and Tamoxifen-resistant HBCx-22-TamR PDX. A threshold of 1,000 reads (dotted line) was applied to eliminate low read count barcodes. (**B**) Copy number in 0.5 Mb non-overlapping regions plotted for bulk DNA profiles of Tamoxifen-resistant PDX (HBCx-22-TamR) versus untreated PDX (HBCx-22). No aberrant variation in copy number were identified in this xenograft model. (**C**) Left: mean of all pairwise correlation score between cluster's members is plotted for *k* clusters ranging from 2 to 10. Right: mean intra-cluster correlation score for k clusters ranging from 2 to 10. At $k = 2$ clusters, the intra-cluster correlation is maximized. (**D**) Hierarchical clustering and corresponding heatmap of cell to cell consensus clustering score for scChIP-seq and scRNA-seq tumor cells (HBCx-22 and HBCx-22-TamR PDXs). Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above heatmap.

Table. S1. Oligonucleotide sequences used for single-cell DNA barcode synthesis on hydrogel beads and sequencing library preparation.

Single-cell DNA barcode synthesis

SEQ1	top	/5PCBio/TTAAGAATTTAATACGACTCACTATAGGGAGAGTGA CTGGAGTTCAGACGTGTGCTCTTCCGATCT
	bottom	5P/CGAAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTC TCCCTATAGTGAGTCGTATTAAATTCTTAA/C3Spacer
SEQ2	top	/5Phos/CAACGTGATTGCTTGTGACTTAA
	bottom	/5Phos/TTAAGTCACAAGCAATCAC
SEQ3	Top	/5Phos/**-linker-**GATACCGTCGAC/3C6/
	bottom	/5InvddT/GTCGACGGTATC

Sequencing library preparation

SEQ4	RT primer	TACACGACGCTCTTCCGATCTNNNNNN
SEQ5	forward PCR primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC GACGCTCTTCCGATCT
	reverse PCR primer	CAAGCAGAAGACGGCATACGAGAT**-index-** GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

linker = index linker sequence (TTCG, TGAC or TCCC)
index = Illumina indexed primer

Discussion

Chromatin modifications play a central role in the regulation of gene expression and cell-type specific functions. Genome-wide mapping of histone modification patterns have already revealed altered chromatin states in tumor cells through which therapeutic resistance might emerge. However, comprehensive profiling of histone modifications in the context of normal and disease cell states is hampered by heterogeneity in chromatin states. Current profiling methods only yield averaged "snapshots" of the distribution of histone marks, principally driven by the dominant cell type. Studying the heterogeneity of chromatin alterations at the single-cell level is mandatory to understand their contribution to the tumor evolution. Such measurements at single-cell resolution have the potential to assess multiple facets of cancer biology including intratumoral heterogeneity, subpopulation characterisation, response to therapeutic treatment and the emergence of drug resistance.

Single-cell chromatin profiles recapitulate cell-type specific chromatin states with high accuracy

The power to identify cell type-specific chromatin states relies on measuring coherent variations between single-cell profiles. A high single-cell coverage is a must-have in order to reveal rich biological information and distinguish underlying patterns of variability in complex heterogeneous samples.

During this thesis, we conceived an alternative single-cell ChIP-seq platform to the previously reported Drop-ChIP method [Rotem 15a]. Although Drop-ChIP was able to identify distinct subpopulations based on their chromatin profiles, the low coverage per cell (only hundreds of reads per cell) limits its applicability to complex and heterogeneous samples (e.g tumor specimens). Our single-cell ChIP-seq approach relies on droplet-based microfluidic with live monitoring of droplet production, which enables chromatin indexing from isolated single-cells in droplets using DNA barcodes that are unique to a single cell. After indexing, the emulsion is broken and the content of merged droplets is used for immunoprecipitation of nucleosomes carrying the post-translational modification of interest. The barcoded immuno-precipitated DNA is amplified and sequenced using Next Generation Sequencing.

We identified the chromatin indexing in droplets as a critical step of the microfluidic workflow towards the generation of reliable and high coverage single-cell profiles. Particularly, the fine-tuning of enzymatic activities and the design of the DNA barcodes are directly related to the "amount of information" recovered per cell. First, we showed that chromatin from individual cells is efficiently digested by micrococcal nuclease in droplets to generate mostly DNA fragments of a size

of a nucleosome. Importantly, MNase activity is in turn synchronized between droplets, paused and inactivated on-demand to ensure consistency between single-cells. The barcoding strategy is a major weakness of Drop-ChIP, in part due to the complexity of the microfluidic procedure used to generate them. In our system, we developed hydrogel beads grafted with ~5×10^7 copies of the same and unique DNA sequence generated by split-pool synthesis. The structure of the barcodes has been optimized to minimize non-specific ligation events and improve the overall quality of the sequencing libraries. Yet, we anticipate that having more DNA barcodes in droplets available for chromatin indexing would further increase the cell coverage. Also, our scChIP-seq technology might benefit from replacing the hydrogel beads by an alternative barcoding approach which would make the workflow easier to handle and minimize the potential loss of information. For example, *in-situ* barcoding in combinatorial cellular indexing method could be adapted to the scChIP-seq procedure as already reported in single-cell chromatin accessibility assay [Cusanovich 15] or single-cell transcriptomic profiling from tens of thousands of cells [Cao 17, Rosenberg 18].

A high specificity and accuracy are essential to distinguish coherent patterns of variability in complex heterogeneous samples and classify single-cells according to their distinctive chromatin landscapes. Following this direction, we benchmarked the system in a serie of proof of concept experiments intented to simultaneously measure the specificity and accuracy of the procedure.

First, to evaluate the accuracy of the microfluidic workflow to produce chromatin profiles at the single-cell resolution, we profiled a mixed population of mouse and human cells. Post sequencing, we confirmed that 96.5% of the identified barcodes were unambiguously assigned to a single species, indicating minimal cross contamination between droplets and in the amplification steps of the procedure.

Next, we evaluated the specificity of the scChIP-seq procedure to classify single-cells according to their distinctive chromatin landscapes. We separately processed two human cell lines (B and T-cells) in the droplet-microfluidic workflow using two SETS of single-cell barcodes carrying an additional cell type-specific sequence. We profiled chromatin modifications associated with active transcription (H3K4me3) or repressive chromatin state (H3K27me3). For both chromatin marks, an unsupervised clustering approach revealed two well separated clusters which identities were confirmed by the cell-type specific sequence. In average, >99% of the barcodes belonging to one cluster were originating from the same cell type. Also, aggregated scChIP-seq profiles recapitulated cell-type specific chromatin states with high accuracy (Pearson correlation score: 0.93 [H3K4me3] and 0.97 [H3K27me3] with p-value $< 10^{-15}$).

Finally, the scChIP-seq was applied to profile a mixed suspension of two hu-

man cell lines (B and T-cells). The single-cell profiles obtained unambiguously clustered into two populations which closely matched with B and T-cell profiles obtained from conventional bulk ChIP-seq assays, confirming the capability of our scChIP-seq platform to effectively resolve heterogeneous samples.

In terms of single-cell coverage, our system enables the identification of 10^4 enriched loci per cell in comparison with hundreds of loci obtained with Drop-ChIP. This significant 10-fold improvement allows a more accurate profiling of complex biological samples as well as a higher sensitivity to decompose and detect rare cell populations. Importantly, the robustness of our system is supported by the close correlation between the number of droplets containing a cell plus a barcoded bead counted on the microfluidic station and the number of barcodes identified in the sequencing data. This indicates a good overall efficacy of the system in which the introduction of noise and the loss of material are limited. Altogether, these results highlight the performance of the system developed over existing Drop-ChIP method.

Computational analysis of single-cell ChIP-seq sequencing data is challenging because of the sparse coverage of the single-cell profiles and the binary measurement of histone modifications within individual cells (the presence or absence of reads constitutes the only readout of the system). In this thesis, we provide a framework for bioinformatic analysis of such data, using an unsupervised correlation-based clustering approach. This approach presents the advantage to "clean" the datasets by removing uncorrelated single-cells arising from the stochasticity of the chromatin indexing in droplets. Alternative strategies reported in the literature could be used to account for sparse coverage. First, analysis could be restricted to a set of signatures comprising specific genomic regions such as promoters or enhancers depending on the research question [Farlik 15, Buenrostro 15, Rotem 15a]. The use of a selected set of regions might improve the clustering of subpopulations but conversely it might also require preliminary knowledge about the cell types comprising the input sample. A second easy-to-implement solution consists in aggregating single-cell profiles sharing similar chromatin signature within identified clusters [Hou 16], at the cost that the single-cell resolution is lost in favour of an averaged epigenomic information at the level of the cluster. More sophistically, model-based algorithms are also being developed to predict chromatin states from missing data and provide a way around incomplete single-cell information [Ernst 15]. For example, Angermueller et al used a deep learning computational approach to identify known and *de novo* DNA methylation sites from sparse human and mouse single-cell methylation profiles [Angermueller 17]. Nonetheless, we anticipate that more adapted tools will be developed as single-cell epigenomic

technologies in general become popular.

Attractiveness of a new technology depends on its performance but also on the diversity of potential applications. Chromatin profiling is a versatile mean for mapping histone modifications but also DNA binding proteins (e.g transcription factors) and systematically characterize regulatory elements of the genome [Ernst 11]. In conventional chromatin profiling procedure, DNA and proteins are first crosslinked before fragmentation and an antibody is used to specifically enriched genomic regions where the proteins of interest are bound. Following this direction, we demonstrated that the MNase digestion of crosslinked chromatin in droplets is efficient and mostly yielded mono-nucleosomes under optimized conditions (data not shown in this manuscript). Unfortunately, single-cell profiles obtained from fixed human B-cell and T-cell lines were low coverage (from only tens to hundreds enriched loci detected per cell) suggesting low efficacy of chromatin barcoding in droplets as compared to previously reported indexing of fixed chromatin in bulk [Lara-Astiaso 14]. Alternatively, cells might also be replaced by extracted nuclei for single-nucleus chromatin profiling. The use of nuclei is of interest as soon as single-cell suspensions are difficult to obtain (for example in brain studies as reported in single-nucleus RNA-seq [Habib 17]) or from clinical samples that can't be readily dissociated. Preliminary data from frozen fixed nuclei showed efficient chromatin digestion but similarly to fixed cells, further optimizations would be necessary for effective barcoding in droplets (data not shown in this manuscript).

Altogether, these results obtained in the proof of concept study confirm our scChIP-seq procedure as a robust method to profile chromatin landscape at the single-cell level, to classify single-cells with a high accuracy and to identify specific chromatin features between cell populations. Also, the encouraging preliminary data obtained from fixed samples and extracted nuclei suggest that our scChIP-seq procedure is versatile, which offers exciting opportunities for epigenomic studies at single-cell resolution.

Single-cell ChIP-seq identified rare populations of cells in untreated, drug-sensitive tumors with chromatin features that match those of all resistant cells after treatment

Investigating the dynamic evolution of chromatin modifications in response to therapeutic treatment is mandatory to understand the potential emergence of drug resistance and therapeutic failure. Using patient-derived xenograft models of breast cancer, we investigated the heterogeneity of chromatin modifications in the context of acquired resistance to cancer therapy. We applied our scChIP-seq procedure to

profile a pair of PDX samples, one responsive to Tamoxifen and a second one with acquired resistance to Tamoxifen [Cottu 12, Cottu 14]. We identified a rare population of cells in the untreated, drug-sensitive tumor with a H3K27me3 landscape comparable to that of resistant cells after treatment, which might drive therapeutic resistance. Resistant-like cells are characterized by a loss of H3K27me3 enrichment in EGFR loci, a gene known to be involved in Tamoxifen resistance [Massarweh 08, Chung 17]. We observed by single-cell RNA-seq that EGFR is expressed in resistant cells, consistently with loss of H3K27me3 enrichment, but also in a transcriptional sub-clone within the sensitive tumor, suggesting that both chromatin and transcriptional features common to resistant cells are already found in the sensitive tumor. Interestingly, these resistant-like cells were not immediately detected by scRNA-seq. These findings reflect previous single-cell chromatin profiling studies suggesting that chromatin states could more precisely define cell identity and disease evolution than transcriptomic profiles obtained by single-cell RNA-seq [Rotem 15a, Corces 16].

To further complement our study, we profiled H3K27me3 enrichment of a second pair of PDX samples, one responsive to Capecitabine [Marangoni 07] and a second one with acquired resistance to Capecitabine. Resistant-like cells are also present, to a lesser extent, in the drug-sensitive tumor suggesting again the emergence of therapeutic resistance from pre-existing "epigenetic" sub-clones that expand upon treatment.

In addition, analysis and clustering of single-cell chromatin profiles of stromal mouse cells revealed subpopulations of fibroblast-like cells and immune-like cells consistent with previously reported "contaminant" cell types surrounding tumors in PDX models [Derose 11, Whittle 15]. Even if the interest in profiling mouse stromal cells from PDX mice is limited, the identification of distinct subpopulations strengthens the capability of our scChIP-seq system to decompose complex and heterogeneous samples. It also opens new avenues to interrogate the chromatin states of primary tumor "ecosystem" including malignant cells and tumor microenvironnement [Tirosh 16, Puram 17]. For example, a recent single-cell transcriptomic profiling of stromal cells in the lung tumor microenvironment revealed novel cell subtypes and highlighted transcription factors underlying distinct gene expression programs [Lambrechts 18]. Applying our single-cell ChIP-seq procedure would provide additional distinctive chromatin features which might complement transcriptomic characterisation of subpopulations and provide valuable information in the design of adapted therapy.

Fundamentally, our single-cell ChIP-seq system could be leveraged in combina-

tion with single-cell RNA-seq to establish connections between the epigenetic and the transcriptomic heterogeneity at single-cell resolution. Also, the dynamic of epigenetic modifications and the regulation of transcription remains largely uncharted and can be seen as analogous to the "chicken & egg" paradigm. What's coming first? Following this direction, our parallel study of H3K27me3 chromatin landscape and gene expression revealed that "epigenetic sub-clones" don't necessarily match with sub-clones identified on the basis of their transcriptomic profiles. Loss of H3K27me3 enrichment modifies the chromatin structure to a permissive state, where transcription can happen. To fully investigate this correlation, simultaneaous single-cell profiling of transcriptomic and chromatin states from the same cell is necessary. It seems reasonable to envision in a near future the technical feasibility of such study as the combination of certain single-cell sequencing techniques have already been reported in the literature. For example, the parallel sequencing of single-cell genome and transcriptome [Macaulay 15], methylome and transcriptome [Angermueller 16], or a combination of the three [Hou 16], open new avenues to simultaneously explore multi -*omics* layer at single-cell resolution (for review [Kelsey 17, Macaulay 17, Colomé-Tatché 18]).

Appendix

Appendix A

Material & Methods related to [Grosselin et al, in preparation]

Appendix A is dedicated to the material and methods related to [Grosselin et al, in preparation]. Figures referenced in this Appendix are available in Chapter 5, section 5.2 on page 97.

The overall workflow of the single-cell ChIP-seq procedure is shown in fig. S1.

Cell lines

Jurkat cells (ATCC, T18-125), an immortalized line of human T lymphocytes and Ramos cells (ATCC, CRL-1596), an immortalized line of human B lymphocytes, were grown in RPMI medium (Gibco, LifeTechnologies) supplemented with 10% heat inactivated bovine serum and 1% Pen/Strep (ThermoFisher Scientific). Mouse M300.19 parental cells (generous gift from B. Moser), an immortalized line of mouse pre-B lymphocytes, were grown in RPMI 1640 medium supplemented with 10% fetal calf serum, 1% Pen/Strep, 1% glutamine and 5×10^{-5} M β-MeEtOH.

Patient-derived xenografts.

Patient-derived xenografts of luminal breast cancer (HBCx-95 and HBCx-22) were established and treated as previously described [Cottu 14, Marangoni 18] to generate xenografts with acquired resistance to Capecitabine (HBCx-95-CapaR) and Tamoxifen (HBCx-22-TamR).

Microfluidic chips.

Four microfluidic chips were used: i) to compartmentalize single cells with lysis reagents and MNase in droplets; ii) to produce hydrogel beads; iii) to compartmentalize single hydrogel beads in droplets, and iv) for one-to-one fusion of droplets containing digested nucleosomes (from single lysed cells) with droplets containing

single hydrogel beads (fig. S2). All chips were fabricated using soft-photolithography in poly-dimethylsiloxane [Duffy 98] (PDMS, Sylgard) as described [Mazutis 13]. Masters were made using one layer of SU-8 photoresist (MicroChem). List depth of the photoresist layer for device i was 40.8 ± 1 µm, for device ii was 30 ± 1 µm and for device iii was 34 ± 1 µm. For device iv, layer depth was 45 ± 1 µm and electrodes were prepared by melting a 51In 32.5Bi 16.5Sn alloy (Indium Corporation of America) into the electrode channels [Siegel 07]. Microfluidic devices were treated the day of the experiment with 1% v/v 1H,1H,2H,2H-perfluorodecyltrichlorosilane (ABCR) in Novec HFE7100 fluorinated oil (3M) to prevent droplets wetting the channel walls.

Microfluidic operations.
Droplet formation, fusion and fluorescence analysis was performed on a dedicated droplet microfluidic station, similar to [Mazutis 13]. The continuous oil phase for all droplet microfluidics experiments was Novec HFE7500 fluorinated oil (3M) containing 2% w/w 008-FluoroSurfactant (RAN Biotechnologies).

scChIP-seq: cell compartmentalization and chromatin digestion.
Cells were centrifuged (300 g, 5 min at 4°C), labeled by 20 min incubation with 1 µM Calcein AM (ThermoFisher Scientific, C3099) and resuspended in cell suspension buffer, comprising DMEM/F12 (LifeTechnologies) supplemented with 30% Percoll (Sigma), 0.1% Pluronic F68 (LifeTechnologies), 25 mM Hepes pH 7.4 and 50 mM NaCl. Cells were resuspended to give an average number of cells per droplets of $\lambda = 0.1$, resulting in 90.48% of empty droplets, 9.05% of droplets containing one cell and only 0.46% containing two or more cells due to Poisson distribution of the cells in droplets [Clausell-Tormos 08]. The cells were co-flowed in a microfluidic chip (fig. S2) with digestion buffer containing lysis buffer (107.5 mM Tris-HCl pH 7.4, 322.5 mM NaCl, 2.15% Triton Tx-100, 0.215% DOC and 10.75 mM $CaCl_2$), 2 µM Sulforhodamine B (Sigma, #S1402-5G), 4 µM DY405 (Dyomics, #405-00), Protease Inhibitor cocktail and 0.2 U/µl MNase enzyme (ThermoFisher Scientific, EN0181). Droplets were produced by hydrodynamic flow-focusing [Anna 03] with a nozzle of 25 µm wide, 40 µm deep and 40 µm long. The flow rates (150 µl/hr for both aqueous phases, 850 µl/hr for the continuous oil phase) were calibrated to produce 45 pl droplets. The droplets were collected in a collection tube (1.5 ml Eppendorf tube filled with HFE-7500 fluorinated oil) and then incubated at 37°C for 20 min.

scChIP-seq: Production of hydrogel beads.
Hydrogel beads carrying barcoded DNA adaptors were produced using a split-and-mix synthesis as described [Klein 15, Zilionis 17], with minor modifications. Briefly, polyethylene diacrylate (PEG-DA) hydrogel beads containing streptavidin acrylamide were produced and barcoded primers were added to the beads by split-

and-pool synthesis using ligation. PEG-DA hydrogel beads were produced using the microfluidic device indicated in fig. S2. The 9 pl droplets were produced at 4.5 kHz frequency and were exposed at 200 mW/cm^2 with a 365 nm UV light source (OmniCure ac475-365) to trigger gel bead polymerization. Recovered gel beads were washed 10 times with washing buffer (100 mM Tris pH 7.4, 0.1% v/v Tween 20).

scChIP-seq: DNA barcode synthesis on beads.
PEG-DA beads were incubated for 1h at room temperature with a photo-cleavable biotinylated dsDNA oligonucleotide (see SEQ1 in Table S1) and then distributed into 96-wells plate, each well containing a double-stranded DNA with a specific first index (index 1), for split pool mediated ligation using the T7 DNA ligase (NEB) according to the manufacturer's instructions. At each round of split and pool, the hydrogel beads were pooled, washed as described [Zilionis 17]. Repeating this splitting and pooling process 3 times in total (adding 3 index) results in 96^3 combinations, which generates ~8.8×10^5 different barcodes. After adding the last index, the beads were pooled, and a common double-stranded DNA oligo (SEQ2 in Table S1) was ligated to the beads. Each bead carries in average ~5×10^7 copies of a unique barcode (see fig. S3 for quality controls of the single-cell barcodes).

scChIP-seq: compartmentalization of hydrogel beads.
The barcoded hydrogel beads were labeled by 30 min incubation with 10 µM Cy5-PEG3 biotin (Bioscience Interchim, FP-1M1220) and washed with washing buffer (100 mM Tris pH 7.4, 0.1% v/v Tween 20), then suspended in bead mix (62.5 mM EGTA, 2 mM dNTPs, 1 mM ATP, 0.5 µM Sulforhodamine B). Barcoded hydrogel beads were co-flowed using the microfluidic device indicated in fig. S2, with ligation mix (2x ligation buffer, 2 mM ATP, 1 µM Sulforhodamine B, 100 mM EGTA, 0.38 U/µl Fast-link DNA ligase) and EndRepair mix (4x ligation buffer, 4 mM dNTPs, 1 µM Sulforhodamine B, 0.08 U/µl Fast-link DNA ligase, 0.15x ENDit repair mix). The re-injection of the barcoded hydrogel beads in a close packed ordering [Abate 09b] resulted in 70 ± 5% of the droplets containing a single bead. The flow rates (150 µl/hr for the beads, 75 µl/hr for both ligation and EndRepair buffers, 150 µl/hr for the continuous oil phase) were calibrated to produce 100 pl droplets.

scChIP-seq: Barcode-Cell droplets fusion.
Droplets containing fragmented chromatin and droplets containing barcoded hydrogel beads were re-injected into a microfluidic device with two aqueous inlets and one oil inlet for droplet fusion (fig. S2). The paired droplets were electro coalesced [Chabert 05] using an electrical field generated by applying 100V ac (square wave) at 5 kHz across electrodes embedded in the microfluidic device. 75 ± 5% of the droplets were correctly paired and fuse.

scChIP-seq: Nucleosomes Barcoding in droplets.
Fused droplets were collected and exposed to UV for 90 seconds at 200 mW/cm^2 with a 365 nm UV light source (OmniCure ac475-365). The ligation was performed at 16°C overnight. The emulsion was then broken by addition of 1 volume of 80/20 v/v HFE-7500/1H,1H,2H,2H perfluoro-1-octanol (Sigma, 370533). The aqueous phase containing barcoded-nucleosomes was diluted by addition of 10 volumes of lysis dilution buffer (50 mM Tris-HCl pH 7.4, 1% Triton Tx-100, 0.1% DOC, 37.5 mM EDTA, 37.5 mM EGTA, 262.5 mM NaCl and 1.25 mM $CaCl_2$) and centrifuged 10 min at 10,000 g at 4°C. The soluble aqueous phase was used for the chromatin immunoprecipitation.

scChIP-seq: Barcoded-nucleosomes immunoprecipitation.
Protein-A magnetic particles (ThermoFisher Scientific, 10001D) were washed in blocking buffer comprising phosphate buffered saline (PBS) supplemented with 0.5% Tween 20, 0.5% BSA fraction V and incubated 4 hours at 4°C in 1 ml blocking buffer with 2 µg of antibody (anti-H3K4me3 [Millipore, #07-473] and anti-H3K27me3 [Cell Signaling Technology, #9733]). After incubation, the particles were suspended with the barcoded-nucleosomes and incubated at 4°C overnight. Magnetic particles were washed as described [Rotem 15a] and immediately processed to prepare the sequencing library.

scChIP-seq: Sequencing Library Preparation.
Concatemers of barcodes were digested by Pac1 restriction enzyme (NEB, R0547), as described by supplier. Immuno-precipitated chromatin was then treated with RNAse A (ThermoFisher Scientific, EN0531) and with of Proteinase K (ThermoFisher Scientific, EO0491). DNA was eluted from the magnetic particles with 1 volume of elution buffer (1% SDS, 10 mM Tris-HCl pH 8, 600 mM NaCl and 10 mM EDTA). Eluted DNA was purified with 1x AMPure XP beads (Beckman, A63881) and eluted with RNAse/DNase free water. Barcoded-nucleosomes were amplified by in vitro transcription using the T7 MegaScript kit (Ambion, AM1334). The resulting amplified RNA was purified using 1x RNAClean XP beads (Beckman, A66514) and reverse transcribed using SEQ4 (Table S1). After RNA digestion, DNA was amplified by PCR using SEQ5 (Table S1 and fig. S5). The final product was size-selected by gel electrophoresis.

Sequencing.
Single-cell ChIP-seq libraries were sequenced on Illumina NextSeq 500 MidOutput 150 cycles. Cycles were distributed as follows: 50 bp (Read #1) were assigned for the genomic sequence and 100 bp (Read #2) were assigned to the barcode. The 4-first cycles of the Read #2 were dark-cycles to prevent low complexity failure during clusters identification. Bulk ChIP-seq librairies and single-cell RNA-seq librairies were sequenced on Illumina HiSeq 2500 in Rapid run mode PE100.

Sequencing data analysis.

Sequencing data were analyzed with custom Python and R scripts.

Single-cell deconvolution.

Barcodes were extracted from Reads #2 by first searching for the constant 4 bp linkers found between the 20-mer indices of the barcode allowing up to 1 mismatch in each linker (see fig. S2 for barcode structure). If the correct linkers were identified, the three interspersed 20-mer indices were extracted and concatenated together to form a 60 bp non-redundant barcode sequence. A library of all 884,736 combinations of the 3 sets of 96 indices (96^3) was used to map barcode sequences using the sensitive read mapper Cushaw3 [Liu 14]. Each set of indices is error-correcting because it takes more than an edit-distance of 3 to convert one index into another. We therefore set a total mismatch threshold of 3 across the entire barcode, with two or less per index to avoid mis-assigning sequences to the wrong barcode Id. In a second, slower step, sequences that could not be mapped to the Cushaw3 index-library were split into their individual indices and each index compared against the set of 96 possible indices, allowing up to 2 mismatches in each individual index. Any sequences not assigned to a barcode Id by these two steps were discarded. The distribution of raw reads per barcode were fitted as two Gaussian distributions and a threshold was set to eliminate background barcodes with few reads (fig. S7A, fig. S8B and fig. S13A). Reads #1 were aligned to mouse mm10 and human hg38 reference genomes using bowtie-1.2.2 [Langmead 09] by keeping only reads having no more than one reportable alignments. For each barcode, aligned reads were extended to 150 bp and all the reads falling in the same 150 bp window were stacked into one as reads possibly originating from PCR duplicates or from the same nucleosome. Barcodes having at least 500 but no more 10,000 uniquely mapped reads were considered for subsequent analysis (Fig. 1B).

Unsupervised clustering of H3K4me3 and H3K27me3 single-cell profiles of B and T-lymphocytes.

Reads for each barcode were binned in non-overlapping 5 kb genomic bins (50 kb for H3K27me3) spanning the genome to generate a n x m coverage matrix with n barcodes and m genomic bins. We filtered out the coverage matrix eliminating genomic regions not represented in at least 1% of all cells. The coverage matrix was normalized by dividing counts by the total number of reads per barcode and multiplying by the average number of reads across all barcodes. Filtered matrix was reduced by principal component analysis and subpopulations were identified by consensus clustering using the R package ConsensusClusterPlus [Wilkerson 10]. The optimal number of clusters (k) was determined by clustering random selection of 80% of the cells over 1,000 repetitions and k was chosen to maximize

intra-cluster correlation score and clusters' stability. Then, clustering results were visualized in t-SNE plots [Van Der Maaten 08]. To visualize chromatin profiles of subpopulations, we aggregated reads of single-cells within each cluster and created enrichment maps using the R package Sushi [Phanstiel 14] (Fig. 1D).

Unsupervised clustering of H3K27me3 single-cell profiles of mouse stromal and human tumor cells.

Conversely to in-vitro cultured cell lines, PCA of single-cell H3K27me3 profiles of mouse stromal and human tumor cells were driven by confounding factor such as cell coverage. By increasing the number of uniquely mapped reads per barcode, a minimum of 1,600 reads per barcode were necessary to identify subpopulations independently to the cell coverage (fig. S9A). Then, we computed Pearson pairwise correlation and highlighted a subpopulation of non-correlated cells. By plotting the distribution of pairwise Pearson correlation score of the single-cell data and a random distribution, we set a threshold at 0.38 to eliminate non-correlated cells from the coverage matrix (fig. S9B-E). The coverage matrix was then processed as previously described with the optimal number of clusters determined by consensus clustering (fig. S9F-G).

Differential analysis of single-cell ChIP-seq profiles.

To identify differentially enriched regions across single-cells, we performed a non-parametric Wilcoxon signed-rank test. Genomic regions were considered enriched or depleted if the adjusted p-values were lower than 0.01 and the fold change greater than 1.

Correcting tumor copy number profiles.

We used the R package HMMcopy [Lai 16] to correct for copy number variation in non-treated versus resistant xenograft models. Reads from bulk input ChIP-seq samples were binned in 0.5 Mb non-overlapping regions spanning the genome (fig. S12A and fig. S13B).

Appendix B

Bioinformatic pipeline for single-cell ChIP-seq data analysis

This Appendix B is dedicated to the introduction of the bioinformatic tools developed during this thesis for the pre-processing of raw sequencing data from single-cell ChIP-seq experiments. Section B.1 provides an overview of the bioinformatic pipeline as well as the averaged features of sequencing run completion status. Section B.2 details and gives examples of typical results observed at each stage of the pipeline.

B.1 Introduction to the bioinformatic pipeline

The majority of the programs presented in Fig. B.1 have been developed during this thesis. The input files of the pipeline are Illumina raw BCL files generated by the sequencer and the output is a matrix summarizing the number of reads per genomic region and per cell that is be used in downstream analysis specific to each experiment.

Quality control of single-cell ChIP-seq sequencing data

Single-cell ChIP-seq samples were sequenced using the NextSeq system (Illumina) in Paired-End mode, allowing reading from both ends of the molecules.

The structure of sequencing libraries is presented as a reminder in Fig. B.2. Briefly, each molecule is framed by the Illumina adapters P5 and P7 needed for pairing on the flow-cell and cluster formation by bridge amplification. Primers Read #1 and Read #2 initiate sequencing and are assigned to nucleosomal DNA

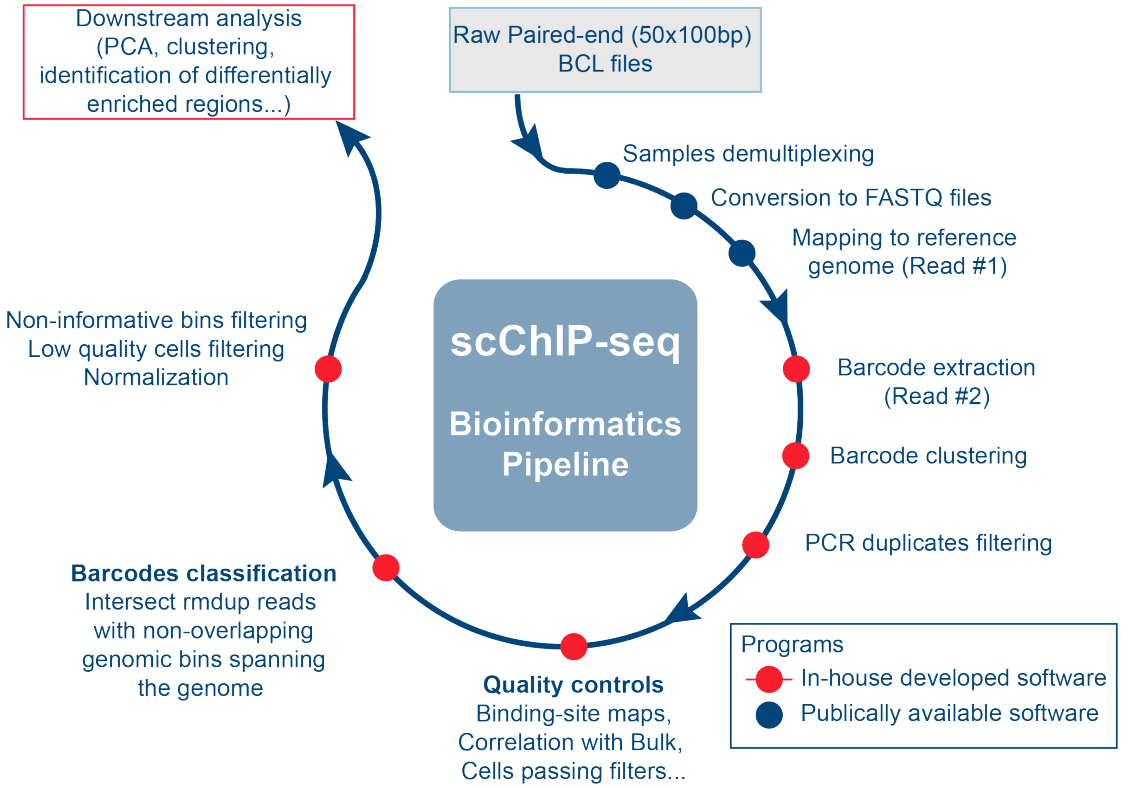


Figure B.1: **Bioinformatic pipeline for processing of raw scChIP-seq data**.

and single-cell barcodes respectively. Also, Read #2 primer can be used optionally to read a 6 bp Illumina index, which enables multiplexing different samples on the same sequencing run.

Illumina systems require some diversity during the first read cycles of Read #1 and Read #2 to accurately determine the position of each cluster on the flow-cell. Too low diversity might negatively impact clusters identification by the detection system, dropping quality of the data. In order to overcome this problem, Illumina recommends to artificially increase the diversity and spike-in known molecules from the PhiX genome virus.

In the context of sequencing single-cell ChIP-seq samples, there is no diversity at all for the first 4 base pairs of Read #2. Indeed, the latter 4 bp are common to all molecules as they are used in the synthesis of the single-cell barcodes (see Chapter 5, see Supplementary Materials). Illumina recommends the addition of 10% to 50% of PhiX on a NextSeq system but we opted for an alternative solution,

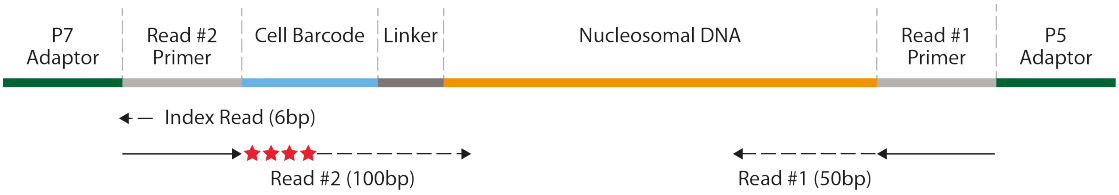


Figure B.2: **Decomposition of sequencing libraries.**
Sequencing libraries are flanked by the Illumina P5 and P7 adapters and Read #1 and Read #2 sequencing primers. The Read #1 (50 bp) is associated with the reading of the nucleosomal DNA and the Read #2 (100 bp) is assigned to the reading of the single-cell barcode sequences. The first 4 cycles of Read #2 (red stars in the figure) are masked to avoid reading failure related to low sequence complexity.

which avoid "losing" up to half of the reads. This solution consists in masking the first 4 base pairs of the Read #2, in other words, not to read the fluorescence signals during these 4 cycles[1]. The latter alternative solution was preferred as the data quality was comparable to expected quality in standard NextSeq runs and only 10% of PhIX was spiked-in.

NextSeq 550 MidOutput 150 cycles V2		
Number of reads	190 ± 50 millions	About 130 M reads after passing filter (Illumina)
Cluster density	250 ± 100 k/mm^2	Between 130 and 165 k/mm^2 (Illumina)
Passing filter	85% ± 15%	Indicates the "purity" of the signal by comparing the fluorescence intensity of the bases
% > Q30	85% ± 10%	Indicates the base percentage with a score greater than 30 (less than 1 error per 1000 bases)

Table B.1: **Sequencing run metrics.**

The table B.1 summarizes the averaged metrics observed over the sequencing runs performed. In average, 190 ± 50 million reads were obtained per run, which is significantly higher than the specifications provided by Illumina. This can be explained by a higher cluster density (250 instead of 150 k/mm^2). Nonetheless,

[1]Masking sequencing cycles requires a "custom recipe" which modifies the program of the sequencer (only provided by Illumina Tech Support)

the quality of the data is not impacted as suggested by high passing filter percentages and Q30 quality score. The latters make it possible to estimate the quality of the data after sequencing: (1) Passing filter is a measurement related to the fluorescence intensity and gives an indication of the purity of each cluster; (2) Q30 score indicates the probability of having a sequencing error over 1000 base pairs (or 99.9% accuracy).

B.2 Overview of raw reads processing steps

This section describes step-by-step the processing of the raw sequencing data. When applicable, averaged results or statistics are given as examples.

Sample demultiplexing and conversion to fastq format

Raw sequencing data were downloaded as BCL files format from the sequencing platform server. The conversion to the FASTQ format and the demultiplexing of the reads were carried out using `bcl2fastq` software available from Illumina website. The program requires a SampleSheet recapitulating each sample multiplexed in the sequencing run, which was automatically generated by a custom python script.

```
$ CreateSampleSheet_bcl2fastq [list of Illumina Index]
$ bcl2fastq -R [run_folder] -o [output_folder] --no-lane-splitting
```

Mapping to the reference genome

Read #1 were aligned to the reference genome (here human hg38; mouse mm9) using `bowtie` [Langmead 09] and keeping only uniquely mapped reads.

```
$ bowtie-build hg38.fa hg38
$ bowtie -m 1 --sam hg38 -q Read1.fastq > Read1_hg38.sam
$ samtools view -F 4 -S Read1_hg38.sam > Read1_hg38_mapped.sam
```

A read is discarded if there is more than one valid alignment for this read (-m option). Samtools were then used to only retain mapped reads in the alignment file in SAM format [Li 09a]. Typically, the global alignement rate was 80% ± 5% but 10% ± 2% of the reads were removed due to multiple alignements.

Deconvolution of barcode reads

The deconvolution of the reads associated with the barcode sequence is already detailled in Chapter 5 (see Supplementary Materials of the paper manuscript).

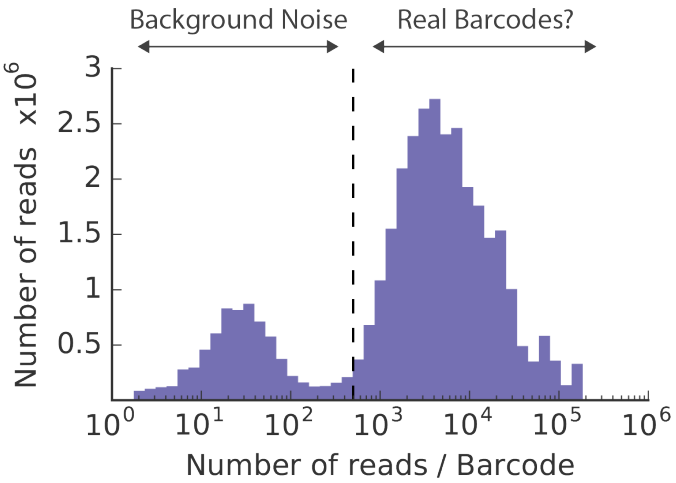


Figure B.1: **Distribution of the number of reads per barcode.**

The reads associated with the barcodes (Reads #2) were decomposed and the 3 indexes were aligned to a reference database containing all possible barcode combinations. The distribution of the number of reads per barcode is illustrated in Fig. B.1 and can be decomposed into 2 normal distributions. The first contains barcodes having between 1 to 100 reads and the latter are most likely associated with background noise. The second distribution groups barcodes having more than 500 reads, which can probably match with the captured cells. This barcode distribution is used to set a first threshold for the elimination of the background noise (here 500 reads, delimited by the black-dotted line).

Identification of PCR duplicates

In the final stage of the preparation of sequencing libraries, DNA were purified on gel and the band between 300 bp and 600 bp were excised and purified. Thus, we can estimate that the sequenced nucleosomal DNA fragment should be equivalent to one nucleosome in length. This estimation was confirmed by the analysis of the sequencing data. Indeed, the length of Read #2 makes it possible to read the single cell barcode plus fifteen base pairs which were then aligned on the reference genome in parallel with the Read #1. Thanks to this double information, the length distribution of entire DNA fragments that were sequenced can be reconstructed (Fig. B.2a).

The size of the sequenced fragments is mainly of the order of the nucleosome and we have based on this observation to identify the reads potentially derived from the same nucleosome. For each single-cell barcode, all reads contained in a 150 bp window relative to the beginning of another read are considered as coming

from the same nucleosome (Fig. B.2b).

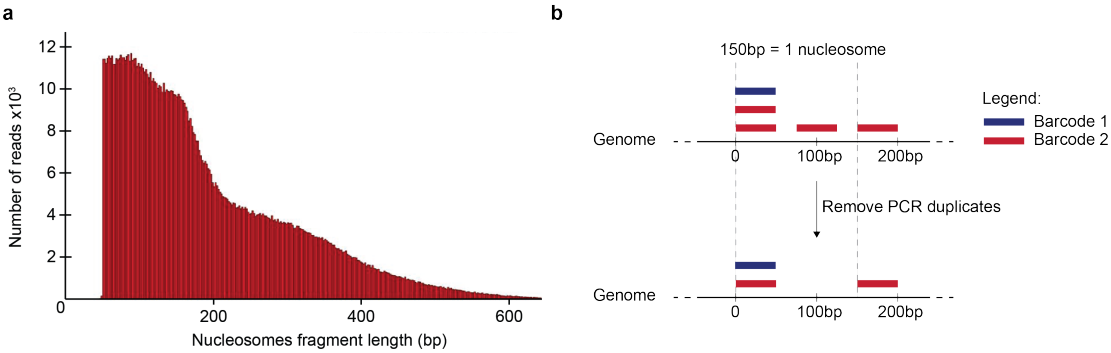


Figure B.2: **Strategy for the identification of duplicate reads.** (a) Length distribution of the DNA fragments sequenced. Most fragments are in the size of one nucleosome (~150 bp). (b) Schematic of the strategy used to identify duplicate reads. For each barcode, all the reads falling in the same 150 bp window are considered originating from the same nucleosome and are removed for subsequent analysis.

A custom python script was developed to flag duplicate reads and output filtered SAM alignment file and filtered Barcode list file.

```
$ PCRduplicateRemoval [SAM file] [Barcode file] [read threshold]
```

Statistics related to the identification of duplicates are illustrated in Fig. B.3 and Fig. B.4. For each read of each barcode, three scenarios are presented:

1. The Read #2 doesn't have its equivalent Read #1 aligned on the genome. The latter read is deleted and flagged as "Unmapped Read #1".

2. The read is identified as duplicate. The latter read is deleted and flagged as "Duplicate reads".

3. If the first two conditions are false, the read is retained in the alignment file for downstream analysis and flagged as "Remaining reads".

The first figure illustrates the impact of sequencing depth on statistics related to duplicates identification. The same sample was sequenced on Illumina MiSeq platform (~15M of reads) and NextSeq (~85M of reads). The y-axis of the left represents the number of reads per barcode for each of the 3 scenarios described above. Each point corresponds to a barcode and is colored according to the type of sequencing platform. Increasing the initial number of reads by a factor of 5

leads to an increase of the number of useful reads per barcode by a factor 2.5. The number of reads with no equivalent Read #1 aligned to the genome increases proportionally as a function of the total number of reads (factor 5). Conversely, duplicate reads explode and are multiplied by a factor ~10.

The second ordinate axis on the right side of the graph translates these increases into an average fraction observed on all barcodes of the initial number of reads after a MiSeq (black round) or NextSeq (black triangle) sequencing. As expected, the percentage of reads without its counterpart aligned to the genome is equivalent for both types of sequencing (~38% of initial reads). It is interesting to note that only $1/4^{\text{th}}$ of the initial reads are useful and kept for subsequent analysis (i.e "remaining read"), unlike duplicate reads whose proportion goes from a fifth to more than a third. In conclusion, for this sample, the sequencing limit is probably reached.

Fig. B.4 presents the average fraction of the number of inital reads observed on the set of barcodes for different epigenetic marks and at comparable sequencing depth (~75M of reads per sample). The trend for H3K4me3 and H3K27me3 samples is similar with $1/3^{\text{rd}}$ duplicate reads and about 40% remaining reads. The main difference is the sample H3K27ac so the proportion of remaining reads exceeds 45% for only 28% of duplicate reads. This difference illustrates a greater complexity of sequencing libraries but results in lower data specificity and higher background noise.

These figures demonstrate the importance of identifying PCR duplicates and reads that may be derived from the same nucleosome. Their proportion by barcode is not negligible and these artefacts could introduce a bias during the clustering of cells and the search for cellular subpopulations.

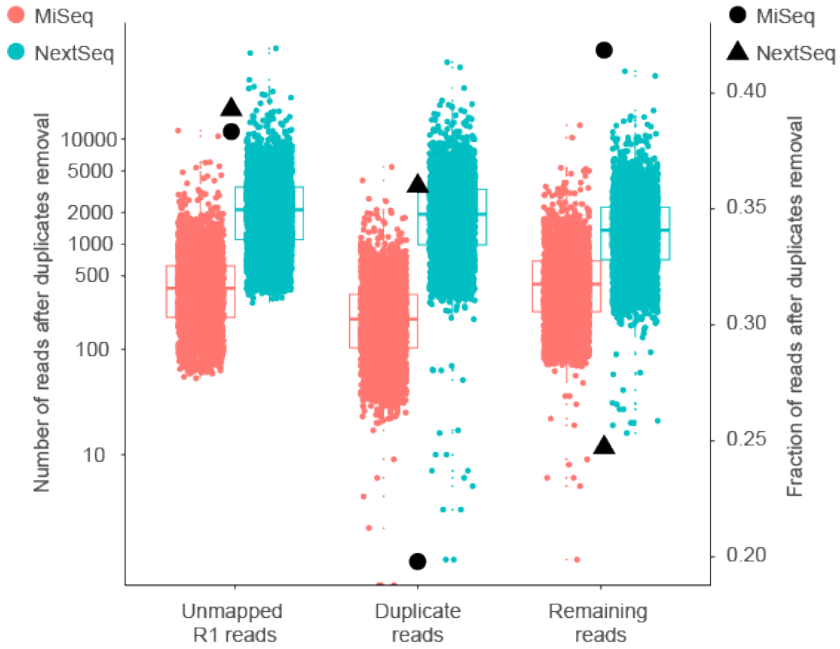


Figure B.3: **Impact of the sequencing depth on the total number of reads per barcode.**
The distribution of the number of reads per label is represented for the 3 possible scenarios when identifying PCR duplicates (y1 axis on the left). When the number of initial reads is multiplied by 5, the number of useful reads is increased by a factor of 2.5 and the number of duplicates by 10. The axis on the right corresponds to the average fraction observed compared to the number of reads initial. 25% of the reads are useful against $> 40\%$ indicating sequencing close to the saturation limit.

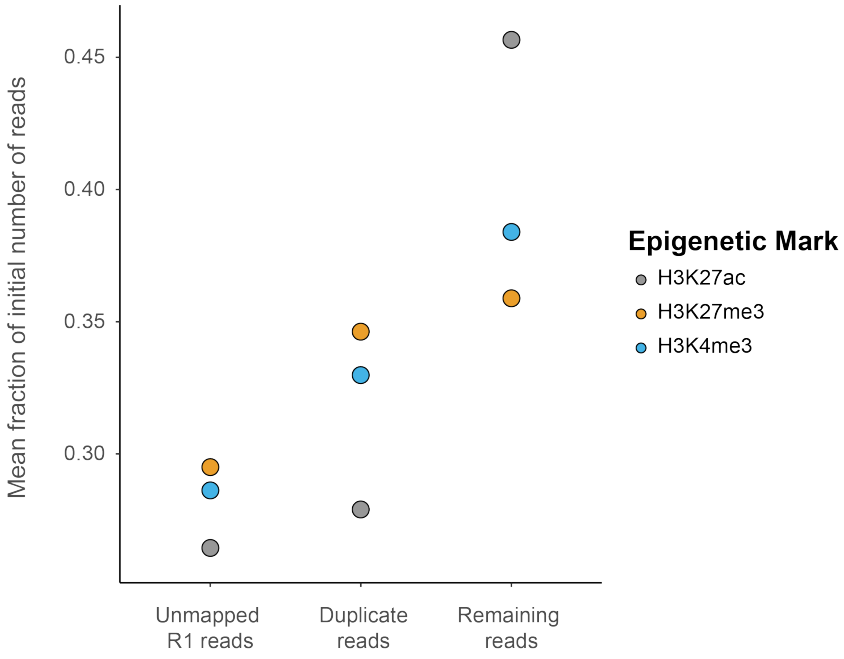


Figure B.4: **Statistics of duplicate reads identification for 3 distinct histone marks.**
The observed average fraction relative to the number of initial reads is represented for 3 distinct epigenetic marks. The proportion of reads for each category is similar for H3K4me3 and H3K27me3.

Generation of the coverage matrix

Unlike the single-cell RNA-seq where reads are associated with genes that can be easily compared, the reads of the single-cell ChIP-seq are distributed on the entire genome. So we need to create a comparison unit from one cell to another and summarize the data. The method we used is to "split" the genome into regions of defined size that do not overlap. These genomic regions can then be used, like the RNA-seq genes, to define common epigenetic profiles between cells.

First, non overlapping regions spanning the genome were created from a list of the chromosome size (available in Ensembl or UCSC websites) using `bedtools makewindows` [Quinlan 10].

```
$ bedtools makewindows -w 50000 -g hg38.chrom.sizes > hg38_50kb.bed
$ head hg38_50kb.bed
chr1 0 50000
chr1 50000 100000
chr1 100000 150000
...
```

Then, reads for each barcode are intersected within the genomic regions using `bedtools coverage` [Quinlan 10] implemented in a custom python script. The output of each barcode is aggregated in a coverage matrix.

```
$ scChIP_classification --sam-file [RMDUP SAM file]
--barcode-clusters-file [RMDUP Barcode file] --nb-reads [read
threshold] --bed-regions hg38_50kb.bed --multiplexing [number of
threads]
```

The program generates a matrix $m \times n$, where m is the number of regions and n is the number of cells. The value c_{ij} of the matrix corresponds to the number of reads of the cell j intersected with the region i.

Filtration and normalization of the coverage matrix

The matrix previously generated was filtered to eliminate bad-quality cells (or aberrant cells) and non-informative regions. The criteria imposed for filtering were the following:

- Barcodes: the total read count per barcode must be higher than 500 (threshold depending on the barcode distribution)
- Regions: if the sum of the reads across all cells is zero, the region was deleted

- Regions: For H3K4me3 datasets, only regions represented in more than 1% of the barcodes were conserved

The resulting filtered matrix was normalized by the total read count per cell to account for the variability in the sequencing depth. With C, the classification matrix, each term c_{ij} was normalized according to B.2.1:

$$\widehat{c_{ij}} = c_{ij}\frac{\overline{M}}{M_i} \; with \; M_i = \sum_j c_{ij} \text{ and } \overline{M} \text{ average of } M_i \tag{B.2.1}$$

The filtered and normalized matrix was then used as a starting point for the analyzes specific to each experiment such as clustering, subpopulation identification, identification of differentially enriched regions, etc.

Bibliography

Bibliography

[Abate 09a] A. R. Abate, A. Poitzsch, Y. Hwang, J. Lee, J. Czerwinska & D. A. Weitz. *Impact of inlet channel geometry on microfluidic drop formation.* Physical Review E, vol. 80, no. 2, page 026310, aug 2009. www

[Abate 09b] Adam R Abate, Chia-Hung Chen, Jeremy J Agresti & David a Weitz. *Beating Poisson encapsulation statistics using close-packed ordering.* Lab on a chip, vol. 9, no. 18, pages 2628–2631, sep 2009. www

[Abate 11] Adam R. Abate & David A. Weitz. *Faster multiple emulsification with drop splitting.* Lab on a Chip, vol. 11, no. 11, page 1911, jun 2011. www

[Adli 10] Mazhar Adli, Jiang Zhu & Bradley E Bernstein. *Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors.* Nature methods, vol. 7, no. 8, pages 615–8, 2010. www

[Agresti 10] Jeremy J Agresti, Eugene Antipov, Adam R Abate, Keunho Ahn, Amy C Rowat, J.-C. Baret, Manuel Marquez, Alexander M Klibanov, Andrew D Griffiths & David A Weitz. *Ultrahigh-throughput screening in drop-based microfluidics for directed evolution.* Proceedings of the National Academy of Sciences, vol. 107, no. 9, pages 4004–4009, mar 2010. www

[Ahn 06] Keunho Ahn, Charles Kerbage, Tom P. Hunt, R. M. Westervelt, Darren R. Link & D. A. Weitz. *Dielectrophoretic manipulation of drops for high-speed microfluidic sorting devices.* Applied Physics Letters, vol. 88, no. 2, pages 1–3, jan 2006. www

[Albayrak 16] Cem Albayrak, Christian A Jordi, Christoph Zechner, Jing Lin, Colette A Bichsel, Mustafa Khammash & Savaş Tay. *Digital Quantification of Proteins and mRNA in Single Mammalian Cells.* Molecular Cell, vol. 61, no. 6, pages 914–924, mar 2016. www

[Allfrey 64] V G Allfrey, R Faulkner & A E Mirsky. *Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis.* Proceedings of the National Academy of Sciences of the United States of America, vol. 51, no. 1938, pages 786–94, 1964. www

[Almendro 14] Vanessa Almendro, Yu Kang Cheng, Amanda Randles, Shalev Itzkovitz, Andriy Marusyk, Elisabet Ametller, Xavier Gonzalez-Farre, Montse Muñoz, Hege G Russnes, Åslaug Helland, Inga H. Rye, Anne Lise Borresen-Dale, Reo Maruyama, Alexander VanOudenaarden, Mitchell Dowsett, Robin L. Jones, Jorge Reis-Filho, Pere Gascon, Mithat Gönen, Franziska Michor & Kornelia Polyak. *Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity.* Cell Reports, vol. 6, no. 3, pages 514–527, 2014. www

[Angerer 17] Philipp Angerer, Lukas Simon, Sophie Tritschler, F. Alexander Wolf, David Fischer & Fabian J. Theis. *Single cells make big data: New challenges and opportunities in transcriptomics.* Current Opinion in Systems Biology, vol. 4, pages 85–91, aug 2017. www

[Angermueller 16] Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle & Wolf Reik. *Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity.* Nature Methods, vol. 13, no. October 2015, 2016. www

[Angermueller 17] Christof Angermueller, Heather J. Lee, Wolf Reik & Oliver Stegle. *DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning.* Genome Biology, vol. 18, no. 1, page 67, dec 2017. www

[Anna 03] Shelley L. Anna, Nathalie Bontoux & Howard A. Stone. *Formation of dispersions using "flow focusing" in microchannels*. Applied Physics Letters, vol. 82, no. 3, pages 364–366, jan 2003. www

[Bannister 11] Andrew J Bannister & Tony Kouzarides. *Regulation of chromatin by histone modifications*. Cell Research, vol. 21, no. 3, pages 381–395, mar 2011. www

[Baret 09] Jean-Christophe Baret, Oliver J. Miller, Valerie Taly, Michaël Ryckelynck, Abdeslam El-Harrak, Lucas Frenz, Christian Rick, Michael L. Samuels, J. Brian Hutchison, Jeremy J. Agresti, Darren R. Link, David A. Weitz & Andrew D. Griffiths. *Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity*. Lab on a Chip, vol. 9, no. 13, page 1850, jul 2009. www

[Barski 07] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev & Keji Zhao. *High-Resolution Profiling of Histone Methylations in the Human Genome*. Cell, vol. 129, no. 4, pages 823–837, 2007.

[Bedford 09] Mark T Bedford & Steven G Clarke. *Protein Arginine Methylation in Mammals: Who, What, and Why*. Molecular Cell, vol. 33, no. 1, pages 1–13, jan 2009. www

[Beneyton 16] Thomas Beneyton, I. Putu Mahendra Wijaya, Prexilia Postros, Majdi Najah, Pascal Leblond, Angelique Couvent, Estelle Mayot, Andrew D. Griffiths & Antoine Drevelle. *High-throughput screening of filamentous fungi using nanoliter-range droplet-based microfluidics*. Scientific Reports, vol. 6, no. 1, page 27223, jul 2016. www

[Beneyton 17] Thomas Beneyton, Stephane Thomas, Andrew D. Griffiths, Jean Marc Nicaud, Antoine Drevelle & Tristan Rossignol. *Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica*. Microbial Cell Factories, vol. 16, no. 1, page 18, dec 2017. www

[Bernstein 05] Bradley E. Bernstein, Michael Kamal, Kerstin Lindblad-Toh, Stefan Bekiranov, Dione K. Bailey, Dana J. Huebert, Scott McMahon, Elinor K. Karlsson, Edward J. Kulbokas, Thomas R. Gingeras, Stuart L. Schreiber & Eric S. Lander. *Genomic maps and comparative analysis of histone modifications in human and mouse.* Cell, vol. 120, no. 2, pages 169–181, jan 2005. www

[Bernstein 07] Bradley E Bernstein, Alexander Meissner & Eric S Lander. *The Mammalian Epigenome.* Cell, vol. 128, no. 4, pages 669–681, 2007. www

[Bird 85] Adrian Bird, Mary Taggart, Marianne Frommer, Orlando J Miller & Donald Macleod. *A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA.* Cell, vol. 40, no. 1, pages 91–99, jan 1985. www

[Bird 99] Adrian P Bird & Alan P Wolffe. *Methylation-induced repression-belts, braces, and chromatin.* Cell, vol. 99, no. 5, pages 451–454, nov 1999. www

[Bird 02] Adrian Bird. *DNA methylation patterns and epigenetic memory.* Genes and Development, vol. 16, no. 1, pages 6–21, jan 2002. www

[Blanc 17] Roméo S Blanc & Stéphane Richard. *Arginine Methylation: The Coming of Age.* Molecular Cell, vol. 65, no. 1, pages 8–24, jan 2017. www

[Bo Zheng 04] Bo Zheng, Joshua D. Tice & Rustem F. Ismagilov. *Formation of Droplets of Alternating Composition in Microfluidic Channels and Applications to Indexing of Concentrations in Droplet-Based Assays.* Analytical chemistry, vol. 76, no. 17, pages 4977–4982, 2004. www

[Bose 15] Sayantan Bose, Zhenmao Wan, Ambrose Carr, Abbas H. Rizvi, Gregory Vieira, Dana Pe'er & Peter A. Sims. *Scalable microfluidics for single-cell RNA printing and sequencing.* Genome Biology, vol. 16, no. 1, page 120, jun 2015. www

[Brennecke 13] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni & Marcus G Heisler. *Accounting for*

technical noise in single-cell RNA-seq experiments. Nature Methods, vol. 10, no. 11, pages 1093–1098, nov 2013. www

[Brind'Amour 15] Julie Brind'Amour, Sheng Liu, Matthew Hudson, Carol Chen, Mohammad M. Karimi & Matthew C. Lorincz. *An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations*. Nature Communications, vol. 6, page 6033, 2015. www

[Brownell 96a] James E. Brownell & C. David Allis. *Special HATs for special occasions: Linking histone acetylation to chromatin assembly and gene activation*. Current Opinion in Genetics and Development, vol. 6, no. 2, pages 176–184, apr 1996. www

[Brownell 96b] James E Brownell, Jianxin Zhou, Tamara Ranalli, Ryuji Kobayashi, Diane G Edmondson, Sharon Y Roth & C. D. Allis. *Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation*. Cell, vol. 84, no. 6, pages 843–851, 1996. www

[Buenrostro 15] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang & William J Greenleaf. *Single-cell chromatin accessibility reveals principles of regulatory variation*. Nature, vol. 523, no. 7561, pages 486–490, 2015. www

[Cao 02] Ru Cao, Liangjun Wang, Hengbin Wang, Li Xia, Hediye Erdjument-Bromage, Paul Tempst, Richard S Jones & Yi Zhang. *Role of histone H3 lysine 27 methylation in polycomb-group silencing*. Science, vol. 298, no. 5595, pages 1039–1043, nov 2002. www

[Cao 17] Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell & Jay Shendure. *Comprehensive single-cell transcriptional profiling of a multicellular organism*. Science, vol. 357, no. 6352, pages 661–667, aug 2017. www

[Chabert 05] Max Chabert, Kevin D. Dorfman & Jean-Louis Viovy. *Droplet fusion by alternating current (AC) field electroco-*

alescence in microchannels. ELECTROPHORESIS, vol. 26, no. 19, pages 3706–3715, oct 2005. www

[Chabert 08] Max Chabert & J.-L. Viovy. *Microfluidic high-throughput encapsulation and hydrodynamic self-sorting of single cells*. Proceedings of the National Academy of Sciences, vol. 105, no. 9, pages 3191–3196, mar 2008. www

[Chaipan 17] Chawaree Chaipan, Anna Pryszlak, Hansi Dean, Pascal Poignard, Vladimir Benes, Andrew D Griffiths & Christoph A Merten. *Single-Virus Droplet Microfluidics for High-Throughput Screening of Neutralizing Epitopes on HIV Particles*. Cell Chemical Biology, vol. 24, no. 6, pages 751–757.e3, jun 2017. www

[Chattopadhyay 14] Pratip K Chattopadhyay, Todd M Gierahn, Mario Roederer & J Christopher Love. *Single-cell technologies for monitoring immune systems*, feb 2014.

[Chung 17] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, Zhengyan Kan, Wonshik Han & Woong-Yang Park. *Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer*. Nature Communications, vol. 8, page 15081, may 2017. www

[Clark 16] Stephen J. Clark, Heather J. Lee, Sébastien A. Smallwood, Gavin Kelsey & Wolf Reik. *Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity*. Genome Biology, vol. 17, no. 1, page 72, dec 2016. www

[Clausell-Tormos 08] Jenifer Clausell-Tormos, Diana Lieber, Jean-Christophe Baret, Abdeslam El-Harrak, Oliver J. Miller, Lucas Frenz, Joshua Blouwolff, Katherine J. Humphry, Sarah Köster, Honey Duan, Christian Holtze, David A. Weitz, Andrew D. Griffiths & Christoph A. Merten. *Droplet-Based Microfluidic Platforms for the Encapsulation and Screening of Mammalian Cells and Multicellular Organisms*. Chemistry & Biology, vol. 15, no. 5, pages 427–437, may 2008. www

[Clements 03] Adrienne Clements, Arienne N Poux, Wan Sheng Lo, Lorraine Pillus, Shelley L Berger & Ronen Marmorstein. *Structural basis for histone and phosphohistone binding by the GCN5 histone acetyltransferase.* Molecular Cell, vol. 12, no. 2, pages 461–473, aug 2003. www

[Collins 15] David J. Collins, Adrian Neild, Andrew DeMello, Ai-Qun Liu & Ye Ai. *The Poisson distribution and beyond: methods for microfluidic droplet production and single cell encapsulation.* Lab on a Chip, vol. 15, no. 17, pages 3439–3459, aug 2015. www

[Colomé-Tatché 18] M. Colomé-Tatché & F. J. Theis. *Statistical single cell multiomics integration*, feb 2018. `https://www.sciencedirect.com/science/article/pii/S2452310018300039{#}bib9`

[Corces 16] M Ryan Corces, Jason D Buenrostro, Beijing Wu, Peyton G Greenside, Steven M Chan, Julie L Koenig, Michael P Snyder, Jonathan K Pritchard, Anshul Kundaje, William J Greenleaf, Ravindra Majeti & Howard Y Chang. *Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.* Nature Genetics, vol. 48, no. 10, pages 1193–1203, oct 2016. www

[Cottu 12] P. Cottu, E. Marangoni, F. Assayag, P. De Cremoux, A. Vincent-Salomon, Ch. Guyader, L. De Plater, C. Elbaz, N. Karboul, J. J. Fontaine, S. Chateau-Joubert, P. Boudou-Rouquette, S. Alran, V. Dangles-Marie, D. Gentien, M. F. Poupon & D. Decaudin. *Modeling of response to endocrine therapy in a panel of human luminal breast cancer xenografts.* Breast Cancer Research and Treatment, vol. 133, no. 2, pages 595–606, 2012. www

[Cottu 14] Paul Cottu, Ivan Bièche, Franck Assayag, Rania El Botty, Sophie Chateau-Joubert, Aurélie Thuleau, Thomas Bagarre, Benoit Albaud, Audrey Rapinat, David Gentien, Pierre De La Grange, Vonick Sibut, Sophie Vacher, Rana Hatem, Jean Luc Servely, Jean Jacques Fontaine, Didier Decaudin, Jean Yves Pierga, Sergio Roman-Roman & Elisabetta Marangoni. *Acquired resistance to endocrine treatments is associated with tumor-specific molecular changes in patient-derived luminal breast cancer xenografts.* Clinical

Cancer Research, vol. 20, no. 16, pages 4314–4325, aug 2014. www

[Courtois 08] Fabienne Courtois, Luis F. Olguin, Graeme Whyte, Daniel Bratton, Wilhelm T. S. Huck, Chris Abell & Florian Hollfelder. *An Integrated Device for Monitoring Time-Dependent in vitro Expression From Single Genes in Picolitre Droplets.* ChemBioChem, vol. 9, no. 3, pages 439–446, feb 2008. www

[Cramer 04] Carsten Cramer, Peter Fischer & Erich J Windhab. *Drop formation in a co-flowing ambient fluid.* Chemical Engineering Science, vol. 59, pages 3045–3058, 2004. www

[Cusanovich 15] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell & Jay Shendure. *Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing.* Science, vol. 348, no. 6237, pages 910–914, may 2015. www

[DeKosky 13] Brandon J DeKosky, Gregory C Ippolito, Ryan P Deschner, Jason J Lavinder, Yariv Wine, Brandon M Rawlings, Navin Varadarajan, Claudia Giesecke, Thomas Dörner, Sarah F Andrews, Patrick C Wilson, Scott P Hunicke-Smith, C Grant Willson, Andrew D Ellington & George Georgiou. *High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire.* Nature biotechnology, vol. 31, no. 2, pages 166–169, feb 2013. www

[Delley 18] Cyrille L. Delley, Leqian Liu, Maen F. Sarhan & Adam R. Abate. *Combined aptamer and transcriptome sequencing of single cells.* Scientific Reports, vol. 8, no. 1, page 2919, dec 2018. www

[Derose 11] Yoko S. Derose, Guoying Wang, Yi Chun Lin, Philip S Bernard, Saundra S Buys, Mark T.W. Ebbert, Rachel Factor, Cindy Matsen, Brett A Milash, Edward Nelson, Leigh Neumayer, R Lor Randall, Inge J Stijleman, Bryan E. Welm & Alana L Welm. *Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes.* Nature Medicine, vol. 17, no. 11, pages 1514–1520, nov 2011. www

[Du 15] Jiamu Du, Lianna M Johnson, Steven E Jacobsen & Dinshaw J Patel. *DNA methylation pathways and their crosstalk with histone methylation.* Nature Reviews Molecular Cell Biology, vol. 16, no. 9, pages 519–532, 2015. www

[Duffy 98] David C. Duffy, J. Cooper McDonald, Olivier J.A. Schueller & George M. Whitesides. *Rapid prototyping of microfluidic systems in poly(dimethylsiloxane).* Analytical Chemistry, vol. 70, no. 23, pages 4974–4984, 1998. www

[Edd 08] Jon F. Edd, Dino Di Carlo, Katherine J. Humphry, Sarah Köster, Daniel Irimia, David A. Weitz & Mehmet Toner. *Controlled encapsulation of single-cells into monodisperse picolitre drops.* Lab on a Chip, vol. 8, no. 8, page 1262, jul 2008. www

[Elowitz 02] Michael B Elowitz, Arnold J Levine, Eric D Siggia & Peter S Swain. *Stochastic gene expression in a single cell.* Science, vol. 297, no. 5584, pages 1183–1186, aug 2002. www

[Epigenomics 15] Consortium Epigenomics. *Integrative analysis of 111 reference human epigenomes.* Nature, vol. 518, no. 7539, pages 317–329, feb 2015. www

[Ernst 11] Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis & Bradley E. Bernstein. *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, vol. 473, no. 7345, pages 43–49, may 2011. www

[Ernst 15] Jason Ernst & Manolis Kellis. *Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues.* Nature Biotechnology, vol. 33, no. 4, pages 364–376, apr 2015. www

[Esteller 07] Manel Esteller. *Epigenetic gene silencing in cancer: The DNA hypermethylome.* Human Molecular Genetics, vol. 16, no. R1, pages R50–R59, apr 2007. www

[Eyer 17] Klaus Eyer, Raphaël C L Doineau, Carlos E Castrillon, Luis Briseño-Roa, Vera Menrath, Guillaume Mottet, Patrick

England, Alexei Godina, Elodie Brient-Litzler, Clément Nizak, Allan Jensen, Andrew D Griffiths, Jérôme Bibette, Pierre Bruhns & Jean Baudry. *Single-cell deep phenotyping of IgG-secreting cells for high-resolution immune monitoring.* Nature Biotechnology, vol. 35, no. 10, pages 977–982, sep 2017. www

[Fan 03] Yuhong Fan, Tatiana Nikitina, Elizabeth M Morin-Kensicki, Jie Zhao, Terry R Magnuson, Christopher L Woodcock & Arthur I Skoultchi. *H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo.* Molecular and cellular biology, vol. 23, no. 13, pages 4559–72, jul 2003. www

[Fan 11] H Christina Fan, Jianbin Wang, Anastasia Potanina & Stephen R Quake. *Whole-genome molecular haplotyping of single cells.* Nature Biotechnology, vol. 29, no. 1, pages 51–59, jan 2011. www

[Fan 15] H. Christina Fan, Glenn K. Fu & Stephen P. A. Fodor. *Combinatorial labeling of single cells for gene expression cytometry.* Science, vol. 347, no. 6222, 2015.

[Farlik 15] Matthias Farlik, Nathan C Sheffield, Angelo Nuzzo, Paul Datlinger, Andreas Schönegger, Johanna Klughammer & Christoph Bock. *Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics.* Cell Reports, vol. 10, no. 8, pages 1386–1397, mar 2015. www

[Felsenfeld 03] Gary Felsenfeld & Mark Groudine. *Controlling the double helix.* Nature, vol. 421, no. 6921, pages 448–453, jan 2003. www

[Fischle 05] Wolfgang Fischle, Shan Tseng Boo, Holger L. Dormann, Beatrix M. Ueberheide, Benjamin A. Garcia, Jeffrey Shabanowitz, Donald F. Hunt, Hironori Funabiki & C. David Allis. *Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation.* Nature, vol. 438, no. 7071, pages 1116–1122, dec 2005. www

[Flavahan 17] William A. Flavahan, Elizabeth Gaskell & Bradley E. Bernstein. *Epigenetic plasticity and the hallmarks of cancer.* Science, vol. 357, no. 6348, 2017. www

[Flemming 82] Walther Flemming. *Zellsubstanz, kern und zelltheilung*. Leipzig, F. C. W. Vogel, 1882.

[Fraga 05] Mario F Fraga, Esteban Ballestar, Ana Villar-Garea, Manuel Boix-Chornet, Jesus Espada, Gunnar Schotta, Tiziana Bonaldi, Claire Haydon, Santiago Ropero, Kevin Petrie, N Gopalakrishna Iyer, Alberto Pérez-Rosado, Enrique Calvo, Juan A Lopez, Amparo Cano, Maria J Calasanz, Dolors Colomer, Miguel Ángel Piris, Natalie Ahn, Axel Imhof, Carlos Caldas, Thomas Jenuwein & Manel Esteller. *Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer*. Nature Genetics, vol. 37, no. 4, pages 391–400, apr 2005. www

[Franke 09] Thomas Franke, Adam R. Abate, David A. Weitz & Achim Wixforth. *Surface acoustic wave (SAW) directed droplet flow in microfluidics for PDMS devices*. Lab on a Chip, vol. 9, no. 18, page 2625, sep 2009. www

[Frei 16] Andreas P Frei, Felice Alessio Bava, Eli R Zunder, Elena W.Y. Hsieh, Shih Yu Chen, Garry P Nolan & Pier Federico Gherardini. *Highly multiplexed simultaneous detection of RNAs and proteins in single cells*. Nature Methods, vol. 13, no. 3, pages 269–275, mar 2016. www

[Frenz 09] Lucas Frenz, Kerstin Blank, Eric Brouzes & Andrew D. Griffiths. *Reliable microfluidic on-chip incubation of droplets in delay-lines*. Lab Chip, vol. 9, no. 10, pages 1344–1348, may 2009. www

[Friedensohn 17] Simon Friedensohn, Tarik A Khan & Sai T Reddy. *Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires*. Trends in Biotechnology, vol. 35, no. 3, pages 203–214, mar 2017. www

[Fu 15] Yusi Fu, Chunmei Li, Sijia Lu, Wenxiong Zhou, Fuchou Tang, X Sunney Xie & Yanyi Huang. *Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification*. Proceedings of the National Academy of Sciences, vol. 112, no. 38, pages 11923–11928, sep 2015. www

[Fuerstman 07] Michael J Fuerstman, Ann Lai, Meghan E Thurlow, Sergey S Shevkoplyas, Howard A Stone & George M

Whitesides. *The pressure drop along rectangular microchannels containing bubbles.* Lab on a Chip, vol. 7, no. 11, page 1479, 2007. www

[Fuks 03] François Fuks, Paul J Hurd, Daniel Wolf, Xinsheng Nan, Adrian P Bird & Tony Kouzarides. *The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation.* Journal of Biological Chemistry, vol. 278, no. 6, pages 4035–4040, feb 2003. www

[Gaudet 03] François Gaudet, J Graeme Hodgson, Amir Eden, Laurie Jackson-Grusby, Jessica Dausman, Joe W Gray, Heinrich Leonhardt & Rudolf Jaenisch. *Induction of tumors in mice by genomic hypomethylation.* Science, vol. 300, no. 5618, pages 489–492, apr 2003. www

[Georgiou 14] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann & Stephen R Quake. *The promise and challenge of high-throughput sequencing of the antibody repertoire.* Nature Biotechnology, vol. 32, no. 2, pages 158–168, jan 2014. www

[Gierahn 17] Todd M Gierahn, Marc H Wadsworth, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love & Alex K Shalek. *Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput.* Nature Methods, vol. 14, no. 4, pages 395–398, apr 2017. www

[Grigaityte 17] Kristina Grigaityte, Jason A Carter, Stephen J Goldfless, Eric W Jeffery, Ronald J Hause, Yue Jiang, David Koppstein, Adrian W Briggs, George M Church, Francois Vigneault & Gurinder S Atwal. *Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the T cell repertoire.* bioRxiv, page 213462, nov 2017. www

[Grindberg 13] Rashel V Grindberg, Joyclyn L Yee-Greenbaum, Michael J McConnell, Mark Novotny, Andy L O'Shaughnessy, Georgina M Lambert, M. J. Arauzo-Bravo, Jun Lee, Max Fishman, Gillian E Robbins, Xiaoying Lin, Pratap Venepally, Jonathan H Badger, David W Galbraith, Fred H Gage & Roger S Lasken. *RNA-sequencing from single nuclei.* Proceed-

ings of the National Academy of Sciences, vol. 110, no. 49, pages 19802–19807, dec 2013. www

[Habib 16] Naomi Habib, Yinqing Li, Matthias Heidenreich, Lukasz Swiech, Inbal Avraham-Davidi, John J Trombetta, Cynthia Hession, Feng Zhang & Aviv Regev. *Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons.* Science, vol. 353, no. 6302, pages 925–928, jul 2016. www

[Habib 17] Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R Choudhury, François Aguet, Ellen Gelfand, Kristin Ardlie, David A Weitz, Orit Rozenblatt-Rosen, Feng Zhang & Aviv Regev. *Massively parallel single-nucleus RNA-seq with DroNc-seq.* Nature Methods, vol. 14, no. 10, pages 955–958, aug 2017. www

[Hansen 89] Jeffrey C. Hansen, Juan Ausio, Valerie H. Stanik & K. E. van Holde. *Homogeneous Reconstituted Oligonucleosomes, Evidence for Salt-Dependent Folding in the Absence of Histone H1.* Biochemistry, vol. 28, no. 23, pages 9129–9136, nov 1989. www

[Hansen 08] Klaus H. Hansen, Adrian P. Bracken, Diego Pasini, Nikolaj Dietrich, Simmi S. Gehani, Astrid Monrad, Juri Rappsilber, Mads Lerdrup & Kristian Helin. *A model for transmission of the H3K27me3 epigenetic mark.* Nature Cell Biology, vol. 10, no. 11, pages 1291–1300, nov 2008. www

[Hashimshony 12] Tamar Hashimshony, Florian Wagner, Noa Sher & Itai Yanai. *CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification.* Cell Reports, vol. 2, no. 3, pages 666–673, 2012. www

[Hassan 02] Ahmed H. Hassan, Philippe Prochasson, Kristen E. Neely, Scott C. Galasinski, Mark Chandy, Michael J. Carrozza & Jerry L. Workman. *Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes.* Cell, vol. 111, no. 3, pages 369–379, nov 2002. www

[Hellman 07] Asaf Hellman & Andrew Chess. *Gene body-specific methylation on the active X chromosome.* Science, vol. 315, no. 5815, pages 1141–1143, feb 2007. www

[Hergeth 15] Sonja P Hergeth & Robert Schneider. *The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle*. EMBO reports, vol. 16, no. 11, pages 1439–1453, nov 2015. www

[Hewish 73] Dean R. Hewish & Leigh A. Burgoyne. *Chromatin substructure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease*. Biochemical and Biophysical Research Communications, vol. 52, no. 2, pages 504–510, may 1973. www

[Hindson 11] Benjamin J. Hindson, Kevin D Ness, Donald A Masquelier, Phillip Belgrader, Nicholas J Heredia, Anthony J Makarewicz, Isaac J Bright, Michael Y Lucero, Amy L Hiddessen, Tina C Legler, Tyler K Kitano, Michael R Hodel, Jonathan F Petersen, Paul W Wyatt, Erin R Steenblock, Pallavi H Shah, Luc J Bousse, Camille B Troup, Jeffrey C Mellen, Dean K Wittmann, Nicholas G Erndt, Thomas H Cauley, Ryan T Koehler, Austin P So, Simant Dube, Klint A Rose, Luz Montesclaros, Shenglong Wang, David P Stumbo, Shawn P Hodges, Steven Romine, Fred P Milanovich, Helen E White, John F Regan, George A Karlin-Neumann, Christopher M Hindson, Serge Saxonov & Bill W Colston. *High-throughput droplet digital PCR system for absolute quantitation of DNA copy number*. Analytical Chemistry, vol. 83, no. 22, pages 8604–8610, nov 2011. www

[Holbert 05] Marc A Holbert & Ronen Marmorstein. *Structure and activity of enzymes that remove histone modifications*. Current Opinion in Structural Biology, vol. 15, no. 6, pages 673–680, 2005. www

[Hong 03] Jong Wook Hong & Stephen R. Quake. *Integrated nanoliter systems*. Nature Biotechnology, vol. 21, no. 10, pages 1179–1183, oct 2003. www

[Hosokawa 17] Masahito Hosokawa, Yohei Nishikawa, Masato Kogawa & Haruko Takeyama. *Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics*. Scientific Reports, vol. 7, no. 1, page 5199, dec 2017. www

[Hotchkiss 48] Rollin D Hotchkiss. *the Quantitative Separation of Purines, Pyrimidines, Snd Nucleosides By Paper Chromatography.* J. Biol. Chem, vol. 175, pages 315–332, 1948. www

[Hou 16] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang & Jirun Peng. *Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas.* Cell Research, vol. 26, no. 3, pages 304–319, mar 2016. www

[Huebner 08] Ansgar Huebner, Luis F. Olguin, Daniel Bratton, Graeme Whyte, Wilhelm T. S. Huck, Andrew J. de Mello, Joshua B. Edel, Chris Abell & Florian Hollfelder. *Development of Quantitative Cell-Based Enzyme Assays in Microdroplets.* Analytical Chemistry, vol. 80, no. 10, pages 3890–3896, may 2008. www

[Islam 11] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg & Sten Linnarsson. *Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq.* Genome research, vol. 21, no. 7, pages 1160–7, jul 2011. www

[Ito 08] Yoko Ito, Thibaud Koessler, Ashraf E.K. Ibrahim, Sushma Rai, Sarah L Vowler, Sayeda Abu-amero, Ana Luisa Silva, Ana Teresa Maia, Joanna E Huddleston, Santiago Uribelewis, Kathryn Woodfine, Maja Jagodic, Raffaella Nativio, Alison Dunning, Gudrun Moore, Elena Klenova, Sheila Bingham, Paul D.P. Pharoah, James D Brenton, Stephan Beck, Manjinder S Sandhu & Adele Murrell. *Somatically acquired hypomethylation of IGF2 in breast and colorectal cancer.* Human Molecular Genetics, vol. 17, no. 17, pages 2633–2643, sep 2008. www

[Ito 11] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He & Yi Zhang. *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, vol. 333, no. 6047, pages 1300–1303, sep 2011. www

[Jaitin 14] D. a. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, a. Mildner, N. Cohen, S. Jung, a. Tanay

& I. Amit. *Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types*. Science, vol. 343, no. 6172, pages 776–779, 2014. www

[Kelly 10] Theresa K Kelly, Daniel D De Carvalho & Peter A Jones. *Epigenetic modifications as therapeutic targets*. Nature Biotechnology, vol. 28, no. 10, pages 1069–1078, oct 2010. www

[Kelsey 17] Gavin Kelsey, Oliver Stegle & Wolf Reik. *Single-cell epigenomics: Recording the past and predicting the future*. Science, vol. 358, no. 6359, pages 69–75, 2017. www

[Kidder 11] Benjamin L Kidder, Gangqing Hu & Keji Zhao. *ChIP-Seq: Technical considerations for obtaining high-quality data*. Nature Immunology, vol. 12, no. 10, pages 918–922, 2011. www

[Kim 07] Jin-Woong Kim, Andrew S. Utada, Alberto Fernandez-Nieves, Zhibing Hu & David A. Weitz. *Fabrication of Monodisperse Gel Shells and Functional Microgels in Microfluidic Devices*. Angewandte Chemie International Edition, vol. 46, no. 11, pages 1819–1822, mar 2007. www

[Kim 13] Kyunghwan Kim, Bomi Lee, Jaehoon Kim, Jongkyu Choi, Jin Man Kim, Yue Xiong, Robert G Roeder & Woojin An. *Linker histone H1.2 cooperates with Cul4A and PAF1 to drive H4K31 ubiquitylation-mediated transactivation*. Cell Reports, vol. 5, no. 6, pages 1690–1703, dec 2013. www

[Kintses 10] Balint Kintses, Liisa D van Vliet, Sean RA Devenish, Florian Hollfelder, Corresponding Author, by Adam Woolley & Andrew J DeMello. *Microfluidic droplets: new integrated workflows for biological experiments This review comes from a themed issue on Nanotechnology and Miniaturization Edited*. Current Opinion in Chemical Biology, vol. 14, pages 548–555, 2010. www

[Klein 15] Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz & Marc W. Kirschner. *Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells*. Cell, vol. 161, no. 5, pages 1187–1201, 2015. www

[Knoechel 14] Birgit Knoechel, Justine E Roderick, Kaylyn E Williamson, Jiang Zhu, Jens G Lohr, Matthew J Cotton, Shawn M Gillespie, Daniel Fernandez, Manching Ku, Hongfang Wang, Federica Piccioni, Serena J Silver, Mohit Jain, Daniel Pearson, Michael J Kluk, Christopher J Ott, Leonard D Shultz, Michael a Brehm, Dale L Greiner, Alejandro Gutierrez, Kimberly Stegmaier, Andrew L Kung, David E Root, James E Bradner, Jon C Aster, Michelle a Kelliher & Bradley E Bernstein. *An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia.* Nature Genetics, vol. 46, no. 4, pages 364–370, apr 2014. www

[Kornberg 74] R D Kornberg & J O Thomas. *Chromatin Structure: Oligomers of the Histones.* Science, vol. 184, no. 4139, pages 865–868, may 1974. www

[Kornberg 77] R D Kornberg. *Structure of Chromatin.* Ann. Rev. Biochem, vol. 46, pages 931–54, 1977. www

[Köster 08] Sarah Köster, Francesco E. Angilè, Honey Duan, Jeremy J. Agresti, Anton Wintner, Christian Schmitz, Amy C. Rowat, Christoph A. Merten, Dario Pisignano, Andrew D. Griffiths & David A. Weitz. *Drop-based microfluidic devices for encapsulation of single cells.* Lab on a Chip, vol. 8, no. 7, page 1110, jun 2008. www

[Kouzarides 07] Tony Kouzarides. *Chromatin Modifications and Their Function.* Cell, vol. 128, no. 4, pages 693–705, feb 2007. www

[Lacar 16] Benjamin Lacar, Sara B. Linker, Baptiste N. Jaeger, Suguna Krishnaswami, Jerika Barron, Martijn Kelder, Sarah Parylak, Apuã Paquola, Pratap Venepally, Mark Novotny, Carolyn O'Connor, Conor Fitzpatrick, Jennifer Erwin, Jonathan Y. Hsu, David Husband, Michael J. McConnell, Roger Lasken & Fred H. Gage. *Nuclear RNA-seq of single neurons reveals molecular signatures of activation.* Nature Communications, vol. 7, page 11022, apr 2016. www

[Lachner 01] Monika Lachner, Dónal O'Carroll, Stephen Rea, Karl Mechtler & Thomas Jenuwein. *Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins.* Nature, vol. 410, no. 6824, pages 116–20, mar 2001. www

[Lai 16] Daniel Lai, Gavin Ha & Sohrab Shah. *HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data*, 2016.

[Lake 16] Blue B Lake, Rizi Ai, Gwendolyn E Kaeser, Neeraj S Salathia, Yun C Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian Bing Fan, Wei Wang, Jerold Chun & Kun Zhang. *Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain.* Science, vol. 352, no. 6293, pages 1586–1590, jun 2016. www

[Lambrechts 18] Diether Lambrechts, Els Wauters, Bram Boeckx, Sara Aibar, David Nittner, Oliver Burton, Ayse Bassez, Herbert Decaluwe, Andreas Pircher, Kathleen Van den Eynde, Birgit Weynand, Erik Verbeken, Paul De Leyn, Adrian Liston, Johan Vansteenkiste, Peter Carmeliet, Stein Aerts & Bernard Thienpont. *Phenotype Moulding of Stromal Cells in the Lung Tumour Microenvironment.* Nature, pages 1–23, jul 2018. www

[Lan 16] Freeman Lan, John R. Haliburton, Aaron Yuan & Adam R. Abate. *Droplet barcoding for massively parallel single-molecule deep sequencing.* Nature Communications, vol. 7, page 11784, jun 2016. www

[Lan 17] Freeman Lan, Benjamin Demaree, Noorsher Ahmed & Adam R Abate. *Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding.* Nature Biotechnology, vol. 35, no. 7, pages 640–646, may 2017. www

[Langmead 09] Ben Langmead, Cole Trapnell, Mihai Pop & Steven L Salzberg. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, vol. 10, no. 3, page R25, mar 2009. www

[Lara-Astiaso 14] David Lara-Astiaso, Assaf Weiner, Erika Lorenzo-Vivas, Irina Zaretsky, Diego Adhemar Jaitin, Eyal David, Hadas Keren-Shaul, Alexander Mildner, Deborah Winter, Steffen Jung, Nir Friedman & Ido Amit. *Chromatin state dynamics*

during blood formation. Science, vol. 345, no. 6199, pages 943–949, 2014.

[Leman 15] Marie Leman, Faris Abouakil, Andrew D. Griffiths & Patrick Tabeling. *Droplet-based microfluidics at the femtolitre scale.* Lab on a Chip, vol. 15, no. 3, pages 753–765, jan 2015. www

[Leung 16] Kaston Leung, Anders Klaus, Bill K Lin, Emma Laks, Justina Biele, Daniel Lai, Ali Bashashati, Yi-Fei Huang, Radhouane Aniba, Michelle Moksa, Adi Steif, Anne-Marie Mes-Masson, Martin Hirst, Sohrab P Shah, Samuel Aparicio & Carl L Hansen. *Robust high-performance nanoliter-volume single-cell multiple displacement amplification on planar substrates.* Proceedings of the National Academy of Sciences, vol. 113, no. 30, pages 8484–8489, jul 2016. www

[Li 92] En Li, Timothy H Bestor & Rudolf Jaenisch. *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell, vol. 69, no. 6, pages 915–926, 1992. www

[Li 93] En Li, Caroline Beard & Rudolf Jaenisch. *Role for DNA methylation in genomic imprinting.* Nature, vol. 366, no. 6453, pages 362–365, dec 1993. www

[Li 07] Bing Li, Michael Carey & Jerry L Workman. *The role of chromatin during transcription.* Cell, vol. 128, no. 4, pages 707–19, feb 2007. www

[Li 09a] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin & 1000 Genome Project Data Processing 1000 Genome Project Data Processing Subgroup. *The Sequence Alignment/Map format and SAMtools.* Bioinformatics (Oxford, England), vol. 25, no. 16, pages 2078–9, aug 2009. www

[Li 09b] Meng Li, Wei Dong Chen, Nickolas Papadopoulos, Steven N Goodman, Niels Christian Bjerregaard, Søren Laurberg, Bernard Levin, Hartmut Juhl, Nadir Arber, Helen Moinova, Kris Durkee, Kerstin Schmidt, Yiping He, Frank Diehl, Victor E Velculescu, Shibin Zhou, Luis A Diaz, Kenneth W

Kinzler, Sanford D Markowitz & Bert Vogelstein. *Sensitive digital quantification of DNA methylation in clinical samples.* Nature Biotechnology, vol. 27, no. 9, pages 858–863, sep 2009. www

[Li 11a] G W Li & X. Sunney Xie. *Central dogma at the single-molecule level in living cells.* Nature, vol. 475, no. 7356, pages 308–315, jul 2011. www

[Li 11b] Xiao-Bin Li, Feng-Chen Li, Juan-Cheng Yang, Haruyuki Kinoshita, Masamichi Oishi & Marie Oshima. *Study on the mechanism of droplet formation in T-junction microchannel.* Chemical Engineering Science, vol. 69, pages 340–351, 2011. www

[Li 15] Z. Li, A. M. Leshansky, L. M. Pismen & P. Tabeling. *Step-emulsification in a microfluidic device.* Lab on a Chip, vol. 15, no. 4, pages 1023–1031, feb 2015. www

[Link 04] D R Link, S L Anna, D A Weitz & H A Stone. *Geometrically Mediated Breakup of Drops in Microfluidic Devices.* Physical Review Letters, vol. 92, no. 5, page 4, 2004. www

[Link 06] Darren R. Link, Erwan Grasland-Mongrain, Agnes Duri, Flavie Sarrazin, Zhengdong Cheng, Galder Cristobal, Manuel Marquez & David A. Weitz. *Electric control of droplets in microfluidic devices.* Angewandte Chemie - International Edition, vol. 45, no. 16, pages 2556–2560, apr 2006. www

[Liu 14] Yongchao Liu, Bernt Popp & Bertil Schmidt. *CUSHAW3: Sensitive and Accurate Base-Space and Color-Space Short-Read Alignment with Hybrid Seeding.* PLoS ONE, vol. 9, no. 1, page e86869, jan 2014. www

[Love 06] J Christopher Love, Jehnna L Ronan, Gijsbert M Grotenbreg, Annemarthe G. Van Der Veen & Hidde L Ploegh. *A microengraving method for rapid selection of single cells producing antigen-specific antibodies.* Nature biotechnology, vol. 24, no. 6, pages 703–707, jun 2006. www

[Lowe 98] Kenneth C Lowe, Michael R Davey & J Brian Power. *Perfluorochemicals: Their applications and benefits to cell cul-*

ture. Trends in Biotechnology, vol. 16, no. 6, pages 272–278, jun 1998. www

[Luger 97] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent & Timothy J. Richmond. *Crystal structure of the nucleosome core particle at 2.8 Å resolution*. Nature, vol. 389, no. 6648, pages 251–260, sep 1997. www

[Luo 17] Chongyuan Luo, Christopher L Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R Nery, Justin P Sandoval, Brian Bui, Terrence J Sejnowski, Timothy T Harkins, Eran A Mukamel, M Margarita Behrens & Joseph R Ecker. *Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex*. Science, vol. 357, no. 6351, pages 600–604, aug 2017. www

[Ma 18] Sai Ma, Yuan-Pang Hsieh, Jian Ma & Chang Lu. *Low-input and multiplexed microfluidic assay reveals epigenomic variation across cerebellum and prefrontal cortex*. Science Advances, vol. 4, no. April, page eaar8187, apr 2018. www

[Macaulay 15] Iain C Macaulay, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M Shirley, Miriam Smith, Niels Van der Aa, Ruby Banerjee, Peter D Ellis, Michael A Quail, Harold P Swerdlow, Magdalena Zernicka-Goetz, Frederick J Livesey, Chris P Ponting & Thierry Voet. *G&T-seq: parallel sequencing of single-cell genomes and transcriptomes*. Nature Methods, vol. 12, no. 6, pages 519–522, apr 2015. www

[Macaulay 17] Iain C Macaulay, Chris P Ponting & Thierry Voet. *Single-Cell Multiomics: Multiple Measurements from Single Cells*, 2017. `http://www.ncbi.nlm.nih.gov/pubmed/28089370http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5303816`

[Macosko 15] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev & Steven A. McCarroll. *Highly*

Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell, vol. 161, no. 5, pages 1202–1214, 2015. www

[Mahler 15] Lisa Mahler, Miguel Tovar, Thomas Weber, Susanne Brandes, Martin Michael Rudolph, Josef Ehgartner, Torsten Mayr, Marc Thilo Figge, Martin Roth & Emerson Zang. *Enhanced and homogeneous oxygen availability during incubation of microfluidic droplets*. RSC Adv., vol. 5, no. 123, pages 101871–101878, nov 2015. www

[Marangoni 07] Elisabetta Marangoni, Anne Vincent-Salomon, Nathalie Auger, Armelle Degeorges, Franck Assayag, Patricia De Cremoux, Ludmilla De Plater, Charlotte Guyader, Gonzague De Pinieux, Jean Gabriel Judde, Magali Rebucci, Carine Tran-Perennou, Xavier Sastre-Garau, Brigitte Sigal-Zafrani, Olivier Delattre, Véronique Diéras & Marie France Poupon. *A new model of patient tumor-derived breast cancer xenografts for preclinical assays*. Clinical Cancer Research, vol. 13, no. 13, pages 3989–3998, jul 2007. www

[Marangoni 18] Elisabetta Marangoni, Cécile Laurent, Florence Coussy, Rania El Botty, Sophie Chateau-Joubert, Jean-Luc Servely, Ludmilla de Plater, Franck Assayag, Ahmed Dahmani, Elodie Montaudon, Fariba Némati, Justine Fleury, Sophie Vacher, David Gentien, Audrey Rapinat, Pierre Foidart, Nor Eddine Sounni, Agnès Noël, Anne Salomon, Marick Lae, Didier Decaudin, Sergio Roman-Roman, Ivan Bièche, Martine Piccard & Fabien Reyal. *Capecitabine efficacy is correlated with TYMP and RB expression in PDX established from triple-negative breast cancers*. Clinical Cancer Research, vol. 24, no. 11, pages 2605–2615, jun 2018. www

[Massarweh 08] Suleiman Massarweh, C Kent Osborne, Chad J Creighton, Lanfang Qin, Anna Tsimelzon, Shixia Huang, Heidi Weiss, Mothaffar Rimawi & Rachel Schiff. *Tamoxifen resistance in breast tumors is driven by growth factor receptor signaling with repression of classic estrogen receptor genomic function*. Cancer Research, vol. 68, no. 3, pages 826–833, feb 2008. www

[Mazutis 09a] Linas Mazutis, Ali Fallah Araghi, Oliver J. Miller, Jean Christophe Baret, Lucas Frenz, Agnes Janoshazi,

Valérie Taly, Benjamin J. Miller, J. Brian Hutchison, Darren Link, Andrew D. Griffiths & Michael Ryckelynck. *Droplet-based microfluidic systems for high-throughput single DNA molecule isothermal amplification and analysis.* Analytical Chemistry, vol. 81, no. 12, pages 4813–4821, jun 2009. www

[Mazutis 09b] Linas Mazutis, Jean-Christophe Baret & Andrew D. Griffiths. *A fast and efficient microfluidic system for highly selective one-to-one droplet fusion.* Lab on a Chip, vol. 9, no. 18, page 2665, sep 2009. www

[Mazutis 09c] Linas Mazutis & Andrew D. Griffiths. *Preparation of monodisperse emulsions by hydrodynamic size fractionation.* Applied Physics Letters, vol. 95, no. 20, page 204103, nov 2009. www

[Mazutis 13] Linas Mazutis, John Gilbert, W Lloyd Ung, David a Weitz, Andrew D Griffiths & John a Heyman. *Single-cell analysis and sorting using droplet-based microfluidics.* Nature protocols, vol. 8, no. 5, pages 870–91, may 2013. www

[McDaniel 16] Jonathan R McDaniel, Brandon J DeKosky, Hidetaka Tanno, Andrew D Ellington & George Georgiou. *Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes.* Nature Protocols, vol. 11, no. 3, pages 429–442, 2016. www

[Merrifield 63] R. B. Merrifield. *Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide.* Journal of the American Chemical Society, vol. 85, no. 14, pages 2149–2154, jul 1963. www

[Mikkelsen 07] Tarjei S. Mikkelsen, Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae Kyung Kim, Richard P. Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S. Lander & Bradley E. Bernstein. *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, vol. 448, no. 7153, pages 553–560, aug 2007. www

[Najah 12] Majdi Najah, Andrew D. Griffiths & Michael Ryckelynck. *Teaching single-cell digital analysis using droplet-based mi-*

crofluidics. Analytical Chemistry, vol. 84, no. 3, pages 1202–1209, feb 2012. www

[Nan 98] Xinsheng Nan, H. H. Ng, Colin A. Johnson, Carol D. Laherty, Bryan M. Turner, Robert N. Eisenman & Adrian Bird. *Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex*. Nature, vol. 393, no. 6683, pages 386–389, may 1998. www

[Navin 11] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks & Michael Wigler. *Tumour evolution inferred by single-cell sequencing*. Nature, vol. 472, no. 7341, pages 90–95, apr 2011. www

[Ng 09] S. S. Ng, W. W. Yue, U. Oppermann & R. J. Klose. *Dynamic protein methylation in chromatin biology*. Cellular and Molecular Life Sciences, vol. 66, no. 3, pages 407–422, feb 2009. www

[Niu 08] Xize Niu, Shelly Gulati, Joshua B. Edel & Andrew J. DeMello. *Pillar-induced droplet merging in microfluidic circuits*. Lab on a Chip, vol. 8, no. 11, page 1837, nov 2008. www

[Niu 09] Xize Niu, Fabrice Gielen, Andrew J. DeMello & Joshua B. Edel. *Electro-coalescence of digitally controlled droplets*. Analytical Chemistry, vol. 81, no. 17, pages 7321–7325, sep 2009. www

[Obexer 17] Richard Obexer, Alexei Godina, Xavier Garrabou, Peer R.E. Mittl, David Baker, Andrew D. Griffiths & Donald Hilvert. *Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase*. Nature Chemistry, vol. 9, no. 1, pages 50–56, aug 2017. www

[Olins 74] A. L. Olins & D E Olins. *Spheroid Chromatin Units (nu bodies)*. Science, vol. 183, no. 4122, pages 330–332, jan 1974. www

[Olins 03] Donald E. Olins & Ada L. Olins. *Chromatin history: Our view from the bridge.* Nature Reviews Molecular Cell Biology, vol. 4, no. 10, pages 809–814, oct 2003. www

[Patel 14] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev & Bradley E. Bernstein. *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, vol. 344, no. 6190, pages 1396–1401, 2014. www

[Peterson 17] Vanessa M Peterson, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova & Joel A Klappenbach. *Multiplexed quantification of proteins and transcripts in single cells.* Nature Biotechnology, vol. 35, no. 10, pages 936–939, aug 2017. www

[Phanstiel 14] Douglas H Phanstiel, Alan P Boyle, Carlos L Araya & Michael P Snyder. *Sushi.R: Flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures.* Bioinformatics, vol. 30, no. 19, pages 2808–2810, oct 2014. www

[Pienta 84] K. J. Pienta & D. S. Coffey. *A structural analysis of the role of the nuclear matrix and DNA loops in the organization of the nucleus and chromosome.* Journal of cell science. Supplement, vol. 1, no. Supplement 1, pages 123–135, 1984. www

[Portela 10] Anna Portela & Manel Esteller. *Epigenetic modifications and human disease.* Nature Biotechnology, vol. 28, no. 10, pages 1057–1068, oct 2010. www

[Prakadan 17] Sanjay M. Prakadan, Alex K. Shalek & David A. Weitz. *Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices.* Nature Reviews Genetics, vol. 18, no. 6, pages 345–361, apr 2017. www

[Pray-Grant 05] Marilyn G. Pray-Grant, Jeremy A. Daniel, David Schieltz, John R. Yates & Patrick A. Grant. *Chd1 chromodo-*

main links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature, vol. 433, no. 7024, pages 434–438, jan 2005. www

[Preissl 18] Sebastian Preissl, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U. Gorkin, Yanxiao Zhang, Brandon C. Sos, Veena Afzal, Diane E. Dickel, Samantha Kuan, Axel Visel, Len A. Pennacchio, Kun Zhang & Bing Ren. *in developing mouse forebrain reveals.* Nature Neuroscience, vol. 21, no. 3, pages 432–439, mar 2018. www

[Puram 17] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, Daniel G Deschler, Mark A Varvares, Ravi Mylvaganam, Orit Rozenblatt-Rosen, James W Rocco, William C Faquin, Derrick T Lin, Aviv Regev & Bradley E Bernstein. *Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer.* Cell, vol. 171, no. 7, pages 1611–1624.e24, dec 2017. www

[Putney 81] Scott D Putney, S. J. Benkovic & Paul R Schimmel. *A DNA fragment with an alpha-phosphorothioate nucleotide at one end is asymmetrically blocked from digestion by exonuclease III and can be replicated in vivo.* Proceedings of the National Academy of Sciences, vol. 78, no. 12, pages 7350–7354, 1981. www

[Quinlan 10] Aaron R. Quinlan & Ira M. Hall. *BEDTools: A flexible suite of utilities for comparing genomic features.* Bioinformatics, vol. 26, no. 6, pages 841–842, mar 2010. www

[Ram 11] Oren Ram, Alon Goren, Ido Amit, Noam Shoresh, Nir Yosef, Jason Ernst, Manolis Kellis, Melissa Gymrek, Robbyn Issner, Michael Coyne, Timothy Durham, Xiaolan Zhang, Julie Donaghey, Charles B Epstein, Aviv Regev & Bradley E Bernstein. *Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells.* Cell, vol. 147, no. 7, pages 1628–1639, dec 2011. www

[Razin 80] A Razin & A. Riggs. *DNA methylation and gene function.* Science, vol. 210, no. 4470, pages 604–610, nov 1980. www

[Rea 00] Stephen Rea, Frank Eisenhaber, Dónal O'Carroll, Brian D. Strahl, Zu Wen Sun, Manfred Schmid, Susanne Opravil, Karl Mechtier, Chris P. Ponting, C. David Allis & Thomas Jenuwein. *Regulation of chromatin structure by site-specific histone H3 methyltransferases.* Nature, vol. 406, no. 6796, pages 593–599, aug 2000. www

[Reik 05] Wolf Reik & Annabelle Lewis. *Co-evolution of X-chromosome inactivation and imprinting in mammals.* Nature Reviews Genetics, vol. 6, no. 5, pages 403–410, may 2005. www

[Rosenberg 18] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, Suzie H Pun, Drew L Sellers, Bosiljka Tasic & Georg Seelig. *Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.* Science, mar 2018. www

[Rotem 15a] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A. Sperling, Alon Goren, David A. Weitz & Bradley E. Bernstein. *Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state.* Nature Biotechnology, vol. 33, no. 11, pages 1165–1172, 2015. www

[Rotem 15b] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A. Sperling, Michael Schnall-Levin, Huidan Zhang, Anindita Basu, Bradley E. Bernstein & David A. Weitz. *High-Throughput Single-Cell Labeling (Hi-SCL) for RNA-Seq Using Drop-Based Microfluidics.* PLOS ONE, vol. 10, no. 5, 2015. www

[Sarraf 04] Shireen A Sarraf & Irina Stancheva. *Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly.* Molecular Cell, vol. 15, no. 4, pages 595–605, aug 2004. www

[Schones 08] Dustin E. Schones & Keji Zhao. *Genome-wide approaches to studying chromatin modifications.* Nature Reviews Genetics, vol. 9, no. 3, pages 179–191, mar 2008. www

[Schwartzman 15] Omer Schwartzman & Amos Tanay. *Single-cell epigenomics: techniques and emerging applications.* Nature Reviews Genetics, vol. 16, no. 12, pages 716–726, dec 2015. www

[Sciambi 15] Adam Sciambi & Adam R. Abate. *Accurate microfluidic sorting of droplets at 30 kHz.* Lab Chip, vol. 15, no. 1, pages 47–51, dec 2015. www

[Shahi 17] Payam Shahi, Samuel C. Kim, John R. Haliburton, Zev J. Gartner & Adam R. Abate. *Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding.* Scientific Reports, vol. 7, page 44447, mar 2017. www

[Shapiro 03] Howard Shapiro. *Pratical Flow Cytometry, 4th edition.* Wiley-Liss, pages 1–733, 2003. www

[Sharma 10] Sreenath V. Sharma, Diana Y. Lee, Bihua Li, Margaret P. Quinlan, Fumiyuki Takahashi, Shyamala Maheswaran, Ultan McDermott, Nancy Azizian, Lee Zou, Michael A. Fischbach, Kwok Kin Wong, Kathleyn Brandstetter, Ben Wittner, Sridhar Ramaswamy, Marie Classon & Jeff Settleman. *A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations.* Cell, vol. 141, no. 1, pages 69–80, apr 2010. www

[Shen 13] Hui Shen & Peter W. Laird. *Interplay between the cancer genome and epigenome.* Cell, vol. 153, no. 1, pages 38–55, mar 2013. www

[Shi 04] Yujiang Shi, Fei Lan, Caitlin Matson, Peter Mulligan, Johnathan R Whetstine, Philip A Cole, Robert A Casero & Yang Shi. *Histone demethylation mediated by the nuclear amine oxidase homolog LSD1.* Cell, vol. 119, no. 7, pages 941–953, dec 2004. www

[Shogren-Knaak 06] Michael Shogren-Knaak, Haruhiko Ishii, Jian Min Sun, Michael J Pazin, Jamas R Davie & Craig L Peterson. *Histone H4-K16 acetylation controls chromatin structure and protein interactions.* Science, vol. 311, no. 5762, pages 844–847, feb 2006. www

[Siegel 07] Adam C. Siegel, Derek A. Bruzewicz, Douglas B. Weibel & George M. Whitesides. *Microsolidics: Fabrication of three-dimensional metallic microstructures in poly(dimethylsiloxane).* Advanced Materials, vol. 19, no. 5, pages 727–733, mar 2007. www

[Smallwood 14] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik & Gavin Kelsey. *Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity.* Nature Methods, vol. 11, no. 8, pages 817–820, aug 2014. www

[Ståhlberg 12] Anders Ståhlberg, Christer Thomsen, David Ruff & Pierre Åman. *Quantitative PCR analysis of DNA, RNAs, and proteins in the same single cell.* Clinical Chemistry, vol. 58, no. 12, pages 1682–1691, dec 2012. www

[Stoeckius 17] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija & Peter Smibert. *Simultaneous epitope and transcriptome measurement in single cells.* Nature Methods, vol. 14, no. 9, pages 865–868, jul 2017. www

[Strahl 00] Brian D. Strahl & C. David Allis. *The language of covalent histone modifications.* Nature, vol. 403, no. 6765, pages 41–45, jan 2000. www

[Svensson 18] Valentine Svensson, Roser Vento-Tormo & Sarah A Teichmann. *Exponential scaling of single-cell RNA-seq in the past decade.* Nature Protocols, vol. 13, no. 4, pages 599–604, mar 2018. www

[Tabeling 05] P. Tabeling. Introduction to microfluidics. Oxford University Press, 2005. www

[Tahiliani 09] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind & Anjana Rao. *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, vol. 324, no. 5929, pages 930–935, may 2009. www

[Tan 04] Yung-Chieh Tan, Jeffrey S. Fisher, Alan I. Lee, Vittorio Cristini & Abraham Phillip Lee. *Design of microfluidic channel geometries for the control of droplet volume, chemical concentration, and sorting.* Lab on a Chip, vol. 4, no. 4, page 292, jul 2004. www

[Tang 09] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao & M Azim Surani. *mRNA-Seq whole-transcriptome analysis of a single cell.* Nature Methods, vol. 6, no. 5, pages 377–382, may 2009. www

[Taunton 96] J Taunton, C A Hassig & S L Schreiber. *A Mammalian Histone Deacetylase Related to the Yeast Transcriptional Regulator Rpd3p.* Science, vol. 272, no. 5260, pages 408–411, apr 1996. www

[Thorsen 01] Todd Thorsen, Richard W. Roberts, Frances H. Arnold & Stephen R. Quake. *Dynamic Pattern Formation in a Vesicle-Generating Microfluidic Device.* Physical Review Letters, vol. 86, no. 18, pages 4163–4166, apr 2001. www

[Thorsen 02] Todd Thorsen, Sebastian J Maerkl & Stephen R Quake. *Microfluidic large-scale integration.* Science, vol. 298, no. 5593, pages 580–584, oct 2002. www

[Tice 03] Joshua D Tice, Helen Song, Adam D Lyon & Rustem F Ismagilov. *Formation of Droplets and Mixing in Multiphase Microfluidics at Low Values of the Reynolds and the Capillary Numbers.* Langmuir, vol. 19, pages 9127–9133, 2003. www

[Tirosh 16] Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G.K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Alexandra Chloé Villani, Cory M. Johannessen, Aleksandr Y. Andreev, Eliezer M. Van Allen, Monica Bertagnolli, Peter K. Sorger, Ryan J. Sullivan, Keith T. Flaherty, Dennie T. Frederick, Judit Jané-Valbuena, Charles H. Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev & Levi A. Garraway. *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* Science, vol. 352, no. 6282, pages 189–196, apr 2016. www

[Tschiersch 94] Bettina Tschiersch, Annemarie Hofmann, Veiko Krauss, Rainer Dorn, G Korge & Gunter Reuter. *The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes.* The EMBO journal, vol. 13, no. 16, pages 3822–3831, 1994. www

[Tsukada 06] Yu Ichi Tsukada, Jia Fang, Hediye Erdjument-Bromage, Maria E. Warren, Christoph H. Borchers, Paul Tempst & Yi Zhang. *Histone demethylation by a family of JmjC domain-containing proteins.* Nature, vol. 439, no. 7078, pages 811–816, feb 2006. www

[Turner 05] Bryan M. Turner. *Reading signals on the nucleosome with a new nomenclature for modified histones.* Nature Structural and Molecular Biology, vol. 12, no. 2, pages 110–112, 2005. www

[Unger 00] Marc A. Unger, Hou Pu Chou, Todd Thorsen, Axel Scherer & Stephen R Quake. *Monolithic microfabricated valves and pumps by multilayer soft lithography.* Science, vol. 288, no. 5463, pages 113–116, apr 2000. www

[Vakoc 05] Christopher R Vakoc, Sean A Mandat, Benjamin A Olenchock & Gerd A Blobel. *Histone H3 lysine 9 methylation and HP1γ are associated with transcription elongation through mammalian chromatin.* Molecular Cell, vol. 19, no. 3, pages 381–391, aug 2005. www

[Van Der Maaten 08] Laurens Van Der Maaten & Geoffrey Hinton. *Visualizing Data using t-SNE.* Journal of Machine Learning Research, vol. 9, pages 2579–2605, 2008. www

[van Galen 16] Peter van Galen, Aaron D Viny, Oren Ram, Russell J.H. Ryan, Matthew J Cotton, Laura Donohue, Cem Sievers, Yotam Drier, Brian B Liau, Shawn M Gillespie, Kaitlin M Carroll, Michael B Cross, Ross L Levine & Bradley E Bernstein. *A Multiplexed System for Quantitative Comparisons of Chromatin Landscapes.* Molecular Cell, vol. 61, no. 1, pages 170–180, jan 2016. www

[van Holde 89] Kensal E. van Holde. Chromatin. Springer Series in Molecular Biology. Springer New York, New York, NY, 1989. www

[Villani 17] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L De Jager, Orit Rozenblatt-Rosen, Andrew A Lane, Muzlifah Haniffa, Aviv Regev & Nir Hacohen. *Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.* Science (New York, N.Y.), vol. 356, no. 6335, page eaah4573, apr 2017. www

[Waddington 42] Conrad Waddington. *The Epigenotype.* Endeavour, vol. 1, pages 18–20, 1942.

[Wang 08] Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae Young Roh, Weiqun Peng, Michael Q Zhang & Keji Zhao. *Combinatorial patterns of histone acetylations and methylations in the human genome.* Nature Genetics, vol. 40, no. 7, pages 897–903, jul 2008. www

[Wang 14] Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Clonal evolution in breast cancer revealed by single nucleus genome sequencing.* Nature, vol. 512, no. 7513, pages 155–160, aug 2014. www

[Watt 88] Fujiko Watt & P. L. Molloy. *Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter.* Genes & development, vol. 2, no. 9, pages 1136–1143, 1988. www

[Wei 98] Y Wei, C A Mizzen, R G Cook, M A Gorovsky & C D Allis. *Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during mitosis and meiosis in Tetrahymena.* Proceedings of the National Academy of Sciences, vol. 95, no. 13, pages 7480–7484, jun 1998. www

[Whittle 15] James R Whittle, Michael T Lewis, Geoffrey J Lindeman & Jane E Visvader. *Patient-derived xenograft models of breast*

cancer and their predictive power. Breast cancer research, vol. 17, no. 1, page 17, dec 2015. www

[Wilkerson 10] Matthew D Wilkerson & D Neil Hayes. *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.* Bioinformatics (Oxford, England), vol. 26, no. 12, pages 1572–3, jun 2010. www

[Xia 98] Younan Xia & George M. Whitesides. *Soft Lithography.* Angewandte Chemie International Edition, vol. 37, no. 5, pages 550–575, mar 1998. www

[Zagnoni 09a] Michele Zagnoni, Charles N. Baroud & Jonathan M. Cooper. *Electrically initiated upstream coalescence cascade of droplets in a microfluidic flow.* Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, vol. 80, no. 4, page 046303, oct 2009. www

[Zagnoni 09b] Michele Zagnoni & Jonathan M. Cooper. *On-chip electrocoalescence of microdroplets as a function of voltage, frequency and droplet size.* Lab on a Chip, vol. 9, no. 18, page 2652, sep 2009. www

[Zhang 01] Y. Zhang & D. Reinberg. *Transcription regulation by histone methylation: Interplay between different covalent modifications of the core histone tails.* Genes and Development, vol. 15, no. 18, pages 2343–2360, 2001.

[Zhang 12] Wei Yun Zhang, Wenhua Zhang, Zhiyuan Liu, Cong Li, Zhi Zhu & Chaoyong James Yang. *Highly Parallel Single-Molecule Amplification Approach Based on Agarose Droplet Polymerase Chain Reaction for Efficient and Cost-Effective Aptamer Selection.* Analytical Chemistry, vol. 84, no. 1, pages 350–355, jan 2012. www

[Zhou 02] Ming Zhou, Terrence R. Barrette, Chandan Kumar-Sinha, Debashis Ghoshk, Sooryanarayana Varambally, Saravana M. Dhanasekaran, Ming Zhou, Terrence R. Barrette, Chandan Kumar-Sinha, Martin G. Sanda, Debashis Ghosh, Kenneth J. Pienta, Richard G. A. B. Sewalt, Arie P. Otte, Mark A. Rubin & Arul M. Chinnaiyan. *The polycomb group protein EZH2 is involved in progression of prostate cancer.*

Nature, vol. 419, no. OCTOBER, pages 388–390, oct 2002.
WWW

[Zilionis 17] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M. Klein & Linas Mazutis. *Single-cell barcoding and sequencing using droplet microfluidics*. Nature Protocols, vol. 12, no. 1, pages 44–73, 2017.
WWW

Résumé

La nature dynamique de la chromatine est un acteur majeur de la régulation de la transcription et est suspectée de contribuer à l'évolution tumorale. L'étude des modifications de la chromatine à l'échelle de la cellule unique est indispensable pour comprendre l'impact de la plasticité épigénétique au cours de la tumorigénèse.
Dans ce manuscrit, je décris le développement d'un système basé sur la microfluidique en gouttelettes permettant d'obtenir la cartographie des modifications de la chromatine à l'échelle de la cellule unique.
Le système a été évalué pour cartographier des modifications d'histones associées à un état transcriptionnel actif (H3K4me3) ou réprimé (H3K27me3) de cellules B et T humaines. Les données ont permis de classer >99% des cellules sur la base de leur profil épigénétique, définissant ainsi avec une grande précision des états de la chromatine propres à chaque type cellulaire.
A partir de xénogreffes dérivées de patient atteint du cancer du sein et ayant acquis une résistance thérapeutique, le système a permis la détection de sous-populations rares de cellules parmi les tumeurs non-traitées, présentant un profil chromatinien similaire aux cellules cancéreuses résistantes.
Cette étude démontre l'importance de l'hétérogénéité cellulaire sur la progression tumorale et met en évidence une signature épigénétique associée à la résistance et susceptible d'être la cible d'un traitement thérapeutique.

Mots Clés

cellule unique | microfluidique
hétérogénéité tumorale
épigénétique

Abstract

The dynamic nature of chromatin and transcriptional features play a critical role in normal differentiation and are expected to contribute to tumor evolution. Studying the heterogeneity of chromatin alterations with single-cell resolution is mandatory to understand the contribution of epigenetic plasticity in cancer.
In this thesis, I describe a droplet microfluidics approach to profile chromatin landscapes of thousands of cells at single-cell resolution, with an unprecedented coverage of 10,000 loci per cell.
The system was evaluated to profile histone modifications associated with active (H3K4me3) and inactive transcription (H3K27me3) of human B cells and T cells, and revealed that >99% of the cells were correctly assigned to one cell type, defining distinct chromatin states of immune cells with high accuracy.
In patient-derived xenograft (PDX) models of breast cancer with acquired drug resistance, the system enabled the detection of a rare subpopulation of cells in the untreated, drug-sensitive tumors with chromatin features characteristic of resistant cancer cells. These cells had lost chromatin marks (H3K27me3) associated with stable transcriptional repression for a number of genes known to promote resistance, potentially priming them for transcriptional activation.
These results highlight the potential selection of cells with specific chromatin marks in response and in resistance to cancer therapy.

Keywords

single-cell epigenomics | droplet-based microfluidics | tumor heterogeneity | drug resistance