Exploring the universe of single cells using multi-omic approaches

Buys Anton de Barbanson

The work described in this thesis was performed at the Hubrecht Institute for Developmental Biology and Stem Cell Research (the Royal Netherlands Academy of Arts and Sciences, KNAW) within the framework of the research school Cancer Stem cells & Developmental biology (CS&D), which is part of the Utrecht Graduate School of Life Sciences (Utrecht University).

ISBN-13: 978-94-6416-923-2
©2022 by Buys Anton de Barbanson All rights reserved. No parts of this book may be reproduced, stored in a retrieval system or transmitted in any form or by any means, without prior permission of the author.

Cover and illustrations by B. A. de Barbanson

Printed by Ridderprint | www.ridderprint.nl

Exploring the universe of single cells using multi-omic approaches

Het universum van enkele cellen verkennen met behulp van multimodale technieken
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. H. R. B. M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op

woensdag 2 februari 2022 des middags te 12.15 uur

door

Buys Anton de Barbanson

geboren op 11 april 1991
te Groningen

Promotor: Prof. dr. A. van Oudenaarden

Copromotor: Dr. J. de Ridder

Contents

1 Introduction 5
1.1 Single cell genomics 6
1.1.1 Single nucleotide variation 6
1.1.2 Copy number 12
1.2 Single cell epigenomics 16
1.2.1 Single cell DNA methylation detection 16
1.2.2 Histone modifications 19
1.3 Combining it all together:
single cell multi-omics 24
1.4 Thesis outline 26

2 Integration of multiple lineage measurements from the same cell reconstructs parallel tumor evolution 29
2.1 Summary 29
2.2 Introduction 30
2.3 Results 33
2.4 Discussion 41
2.5 Methods 43
2.6 Supplementary figures 57

3 Hierarchical chromatin regulation during blood formation uncovered by single-cell sortChIC 65
3.1 Summary 65
3.2 Introduction 66
3.3 Results 69
3.4 Discussion 86
3.5 Methods 88
3.6 Supplementary figures 101

4 scChIC-TAPS reveals histone modification specific DNA methylation dynamics during the cell cycle 111
4.1 Abstract 111

4.2 Introduction 112
4.3 Results 113
4.4 Discussion 122
4.5 Materials and methods 123
4.6 Supplementary Figures 133

5 Discussion 141

6 Addendum 165
6.1 Samenvatting 165
6.2 Curriculum Vitae 167
6.3 List of publications 168

List of acronyms

FACS	Fluorescence-activated cell sorting
FANS	Fluorescence-activated nuclei sorting
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	Messenger Ribonucleic Acid
RT	Reverse Transcription
IVT	In Vitro Transcription
PCR	Polymerase Chain Reaction
WGS	Whole Genome Sequencing
MDA	Multiple Displacement Amplification
UMI	Unique Molecular Identifier
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
CNV	Copy Number Variant
sSNV	Somatic Single Nucleotide Variant
gSNV	Germline Single Nucleotide Variant
VAF	Variant Allele Frequency
BAF	B-Allele Frequency
LOH	Loss Of Heterozygosity
LOF	Loss Of Function
CIN	Chromosomal instability
ITH	Intra Tumor heterogeneity
CRC	Colorectal Cancer
BM	Bone Marrow
HSPCs	Hematopoietic Stem and Progenitor Cells
HSCs	Hematopoietic stem cells
pA-MNase	protein-A-micrococcal nuclease
MNase	Micrococcal Nuclease
ChIP	Chromatin Immuno Precipitation
DNMT1	DNA Methyltransferase 1
PMDs	Partially Methylated Domains
DHU	dihydroxyuracil
WGBS	Whole Genome Bisulfite Sequencing

Chapter 1

Introduction

From unicellular to complex multi-cellular organisms, no single cell is alike. In humans and other animals, tissues are composed of a variety of cell types which are in themselves composed of heterogeneous cells at multiple molecular levels. Single cell research studies the differences in pheno- and geno-types and drivers of those differences between single cells. Variation is continuously being introduced in all cells, making cells diverge from each other and making every cell unique. The highest divergence rate happens during development when cells are differentiating leading to variations in, for example, histone modifications, methylation or the proteome. During tissue homeostasis, single cells balance environmental factors with their corresponding cell function. Variation between cells can be introduced during cell divisions in form of DNA mutations (SNVs/indels/structural), copy number changes, or methylation not being maintained. Diseases such as cancer can start out as a single cell, which quickly diverges from its ancestral cell and gives rise to a heterogeneous tumor.

In order to measure the molecular variation between single cells, sensitive sequencing technologies have been developed. The single cell sequencing field spawned with single cell DNA sequencing and has progressed to measure transcriptomics, DNA accessibility, histone modifications, locations of DNA-binding proteins, DNA and RNA methylation and lineage.

This thesis focuses on analyzing multiple combined modalities measured with single cell sequencing techniques, also known under the single cell multi-omics umbrella. Here, I will introduce the relevant topics for multi-omics data analysis of single cell genomics and epigenomics data.

Deoxyribonucleic Acid (DNA) is tightly regulated through multiple layers by the epigenome. At the DNA-nucleotide level, cytosines and adenines can be modified with a methyl group, which is known to regulate expression. DNA methylation is the lowest level of the *epigenetic modifications*, and is discussed in section 1.2.1. At a higher epigenetic level, DNA is packaged using nucleosome protein complexes, which can be chemically modified, which I will further discuss in section 1.2.2.

1.1 Single cell genomics

DNA is tightly regulated through multiple layers to control gene expression and to avoid errors during replication. However, nucleotides can be altered, most commonly during cell division, which we observe as mutations. Both large and small mutations in the DNA can be used to infer phylogenetic relationships, and the mutations can have consequences which are extensively studied.

1.1.1 Single nucleotide variation

Every cell division, the nuclear DNA is replicated such that both daughter cells inherit a copy of DNA. Nucleotides and, hence, DNA is affected by general wear and tear caused by many internal and external factors. This wear and tear causes DNA to gradually mutate over time, resulting in an ever-increasing deviation from the original germline sequence.

Mutations can be subdivided into different classes. One class is the single base substitution, *Single Nucleotide Variant* (SNV). There are many processes which can cause the appearance of SNVs. Environmental factors such as mutagenic substances or radiation, transcription and deamination of methylated cytosines resulting in thymine bases. Also, during DNA replication a small amount of mistakes is made, which cause lineages of cells to gradually deviate from each other at the level of SNVs. This also implies that the amount of SNVs is correlated to the amount of cell divisions.

Germline single nucleotide variation

Approximately 35 new SNVs appear during the cell divisions making up a human gamete[1] [49]. When SNVs appear in the germline the offspring will inherit them, making these SNVs germline (gSNVs). When the SNV abundance across randomly sampled individuals is higher than 1%, the SNV is classified as an *Single Nucleotide Polymorphism* (SNP). For a single individual, approximately 1 in 2000 bases contains such an SNP [144]. The alleles of an SNV

---

[1] The cells which contain the genetic information passed to offspring: the egg cell and sperm cell

refers to the set of bases detected at the location of the variant. A set of alleles which have been inherited from a single parent are called a *haplotype*. The set of all SNPs on a single chromosome makes up the *diplotype* of either paternal or maternal chromosome. When multiple alleles are present for a single locus, the variant is called to be *heterozygous*. When both maternal and paternal copies are identical, the variant is called *homozygous*.

SNVs that are spatially close to each other on a chromosome are very unlikely to separate by DNA recombination, which is referred to as linkage-disequilibrium. Hence, by knowing a few SNVs, nearby variants can be inferred. Groups of such highly correlated SNVs are called haplotype blocks. Additionally, haplotype information can help to reconstruct the diplotype of a genomic region, as haplotype blocks are subsets of the diplotype. There are numerous efforts which aim to document human haplotypes such as, for example, HapMap [84] or the 1000 genomes project [10].

Single nucleotide variation effects

A *Somatic Single Nucleotide Variant* (sSNV) is a variant which occurs in a somatic cell[2]. Very small changes in the DNA sequence of a cell can result in a large and abnormal phenotypic effects. For example, SNVs in the TP53, BRCA, APC or SMAD4 genes can lead to cancer. Variants which cause cancer are called driver mutations.

Most of the single base substitutions changes are not harmful; they affect regions of the genome where the exact sequence is not important or, in case of a protein coding gene, the base-substitution does not result in a (functionally) different amino-acid. Additionally, more and still effective copies of the same gene can be available so that the loss of function of one gene does not cause haploinsufficiency. Often, cancer develops from a sequential order of mutations. Therefore, studying the order of mutations across time in cancer samples is of importance, and many examples are available that study the order in which mutations occur across tumor development [140, 20, 192, 147]. The order of driver mutations can be determined by performing single cell phylogenetic inference on single cell DNA sequencing, or by deep bulk sequencing followed by phylogenetic inference and de-convolution of the ’clones’ present in a sample. The set of all somatic variants detected in a sample result in a

---

[2] Any cell other than a gamete, germ cell, gametocyte or undifferentiated stem cell

mutation profile, which is a histogram containing the variant base itself and the two bases flanking the substitution (3bp context). A mutation profile is influenced by the mutational processes and active mutagens, as well as by a cell's activated DNA repair mechanisms. An sSNV can result in a growth advantage of a particular cell. This phenomenon leads to *Intra Tumor heterogeneity* (ITH) which plays a key role in cancer development.

Single cell single nucleotide variant analysis

Disease associated mutations present in the germline are usually detected through bulk sequencing, but can also be detected using single cell DNA sequencing (table 1). Somatic mutations (of which some could be driver mutations) can be detected in bulk as well as single cell data. Linking somatic variants with confidence to specific clones is only possible using single cell data. Somatic SNVs can be used to estimate the past and current mutagenic processes in different clones. *Loss Of Heterozygosity* (LOH) can be determined where one haplotype is lost and only SNPs from a single haplotype can be detected for a region in a single cell.

The general computational workflow of single cell DNA analysis can be separated into 6 steps: Demultiplexing, Trimming, Mapping, Variant Calling, Genotyping and finally Clustering and Imputation. For protocols where multiple samples are sequenced from the same pool of DNA, the *demultiplexing* step assigns reads to a cell of origin. During *trimming*, low quality bases and remains of sequencing adapters are removed. The trimmed reads are *mapped*, which detects the most optimal alignment of the read to the reference genome and assigns a confidence score to this alignment. Commonly used mappers are BWA-mem [104], and Bowtie2 [100]. During variant calling, all mapped reads are scanned and locations in the genome which likely contain an SNV are selected.

Unfortunately, the amplification of single cell libraries is currently more error-prone than bulk libraries. The vast majority of variants present in raw sequencing data is technical variation caused by amplification, library preparation and sequencing. The main driver of a higher error rate is the difference in the amount of input DNA. A bulk library has much more input DNA than a single cell library ($>$100 ng vs 6.6 pg). This means that in order to get to a sufficient DNA concentration required for sequencing, more amplification is required which will introduce more artifacts. Secondly, only a single template

Method	Description
DOP-PCR	High error rate but the coverage is relatively uniform [169]
Cut-seq	More suitable for combining protocols, but might be more error prone due to intermediate *In Vitro Transcription* (IVT) amplification and *Reverse Transcription* (RT) [198]
MDA	Due to the lower error rate more suitable for SNVs calling, but the MDA protocol suffers from severe allelic dropout and allelic bias [52, 68, 157, 82]
MALBAC	Whole genome amplification method with reduced amplification bias compared to MDA approach [200]
LIANTI / Nuc-seq	Tn5-tagmentation based method, where a transposon is used to obtain fragments with adapters, followed by the use of IVT and RT for amplification [41, 183].

Table 1: Overview of common single cell DNA-sequencing methods

molecule is available, which is a single point of failure. When the template is lost before amplification, or not prepared for amplification, it will be impossible to detect during sequencing. When a mutation occurs in a template molecule before or during amplification, the error will be passed on to all its amplified DNA molecules which are sequenced. As a result most or all the sequenced reads will contain the technically induced variant. Hence, an error introduced before amplification will result in a high *Variant Allele Frequency* (VAF), making it cumbersome or impossible to distinguish a technical variant from a biological variant [58].

There are various approaches to reduce the effect of noise and sparsity on variant calls. As most of the single cell amplification methods have relatively high technical mutation rates, correction of the base-calling confidence values are beneficial. This can be performed using, for example, GATK base-calibration [54]. A common approach for selecting a set of high confidence variants is to only take into account variants that occur within plausible haplotypes [78, 30] (Figure 1). Haplotypes are also taken into account in the more accurate bulk variant callers, like HaploTypeCaller [11] and DeepVariant [134]. Single cell variant callers are, however, more stringent as they require the assignment of a haplotype block identified in a germline sample of the same individual to all reads covering that variant. This method is especially effective in a hybrid organism with a lot of heterozygous SNPs and not effective in inbred organisms. To filter potential artifactual variant calls, machine learning is used which is trained on real and false positive variants found in single cell libraries

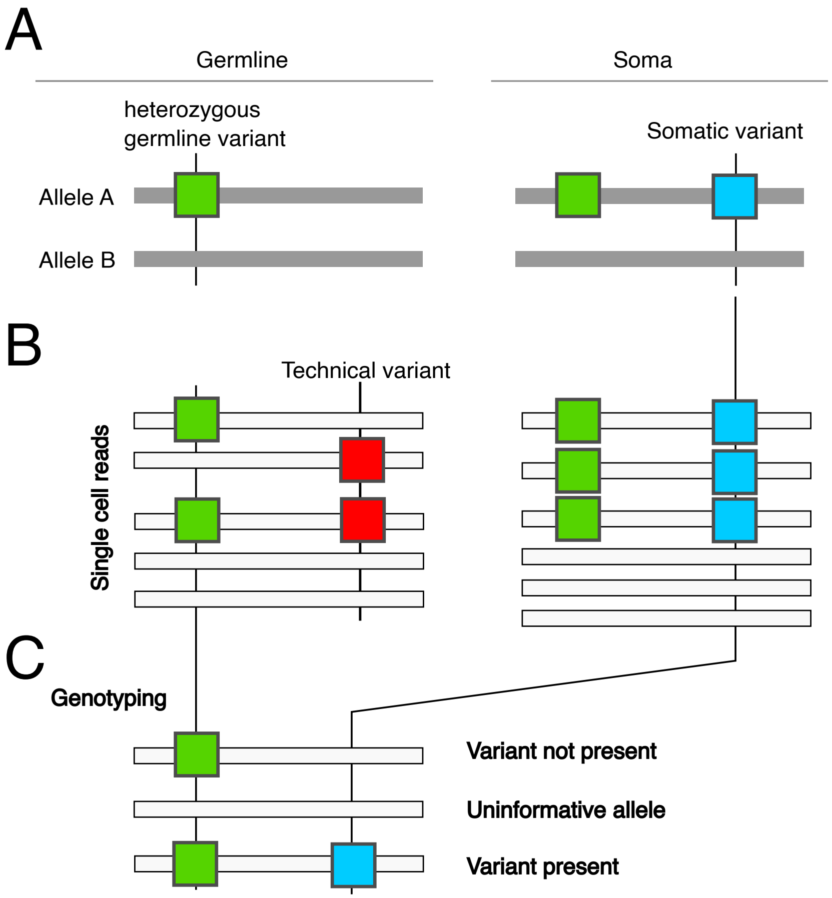


Figure 1: Usage of heterozygous germline variants to detect true somatic variants and genotyping in a high dropout regime. A) Non-inbred diploid organisms contain many heterozygous germline mutations. B) Technical variants are not phase properly with a nearby germline variant, while true somatic variants are perfectly phased. C) During genotyping, the presence and absence of the somatic variant can be estimated using only a single read covering the in-cis heterozygous variant and location of the somatic variant.

in order to select more true positive variants [11]. Integration with other modalities which have different error-profiles can be beneficial for cross-validating mutation calls as well as to genotype cells with, for example, complementary scRNA [168] or bulk sequencing libraries [78].

Both single cell DNA sequencing variant calling and genotyping is mainly challenging due to data sparsity. The sparsity is caused by allelic dropout where not all alleles present are detected and site dropout where a complete site is missed. For the purpose of variant calling, a bulk sample is sequenced with 30x coverage, meaning that on average each position in the genome is covered with 30 and each allele with roughly 15 reads. To get similar quality information from a single cell compared to a bulk library, every cell would need to be sequenced with at least 30x coverage. Due to the very high sequencing costs this is in practice rarely done, or only for very few cells.

Single cell genotyping

Variant calling results in a set of locations in the genome which contain a SNV. Apart from the location of the variant, the reference and alternative allele(s) are determined. Usually, to answer the posed biological questions, it is necessary to, not only know the total set of variants, but to also find what variants are present for each individual cell. When, for example, deciphering what clones are present in a sample and to which clone every cell belongs, single cell resolution is required. The process which identifies what variant is present in each cell is called genotyping. During genotyping it is determined which of the detected alleles are present in each cell. Due to allelic and site dropout it is hard to genotype single cells, resulting in locations where ambiguity arises because either the homolog reference allele is sequenced or the variant is not truly present. Multiple single cell variant callers are available, such as Monovar [194], SCcaller [58], Single cell genotyper [143] and SCAN-SNV [109] which takes advantage of the allelic amplification bias in *Multiple Displacement Amplification* (MDA).

Because single cell genotype data is usually sparse (the coverage is around 0.05x per cell), there are efforts to impute missing genotype data. Imputation of the genotypes is possible when phylogenetic relationships between cells are assumed. As an organism grew out from a single cell, the root of the tree is the germline genotype, and a rooted tree structure can be used with the germline genotype as root. For computational tractability, a model is often used where no back-mutations are permitted, and with a fixed error rate of false variant discoveries, false positives and false negatives.

A common application of single cell genotyping is to estimate the phylogenetic relationships between the cells. Integration of both genotyping and phylogenetic inference solves these problems at the same time, and recent implementations of such approaches have been published [193, 159].

Another option to reduce sparsity is to group cells by another modality such as copy number or lineage marker during phylogenetic reconstruction and/or imputation [147]. For copy number, (co)-clustering SNV information with copy number information is especially beneficial and the interplay, order and causality between sSNVs and copy number aberrations are an active line of research. A main difficulty of integrating copy number and SNVs is that it is not trivial to merge the phylogenetic sSNV edit distance between the cells and the copy number distance (or distances in any other modality) because biologically relevant weights between the modalities are not yet available and might not hold in tumors or vary across cell types [99].

Detecting known SNPs in single cells can be performed on virtually all single cell protocols by verifying which base is present at the SNP location. Not all of these SNPs will be phased with a *Germline Single Nucleotide Variant* (gSNV). For such variants it will not be possible to show whether the variant is not present, especially in a high dropout regime, such as shallow single cell DNA sequencing. Phasing with a gSNV can help detecting which alleles are present in a cell and which are lost. More over, phasing is used in *Ribonucleic Acid* (RNA) sequencing to detect allele specific expression, and in DNA methylation sequencing protocols to detect allele specific methylation.

1.1.2 Copy number

In a mitotic cell division, a cell goes through the cell cycle. In the S phase of the cell cycle, the chromosomes are replicated, resulting in two identical copies, one for each daughter cell. The copies are, in normal circumstances, equally divided between the two daughter cells during M phase. When something goes wrong in distributing the chromosomes over the daughter cells, a process called mis-segregation, the resulting daughter cells will have either lost or gained one or more chromosomes. Such cells with an abnormal amount of copies are called aneuploid cells. Chromosome missegregations occur in a small fraction of normal cell divisions, and are common during embryogenesis

[176]. The frequency of missegregations is increased in cells with *Chromosomal instability* (CIN). Causes of CIN include degenerate telomeres, problems in DNA damage response and defective cell cycle checkpoints. In 68% of cancer patients aneuploidy is detected, and 61% of first trimester miscarriages contain copy number aberrations, with a higher rate when the age of the mother increases [47].

Missegregations mainly cause proliferative disadvantages which are selected against, but in some cases it causes a proliferative advantage. These advantages are caused by dosage changes, which change the relative abundances of genes and therefore influence transcript levels. Missegregations can also cause deletions of cancer genes or removal of the only functional copy due to LOH. Additionally, there is a negative effect when genes where both copies are required for correct function, called haplo-insufficient genes, are deleted . Missegregations cause, in this case, clonal heterogeneity which results in clones that differ in terms of growth, drug response and variable recognition by the immune system.

Next to simple copy number aberrations where complete chromosomes are gained and lost, there are also more complex copy number aberrations such as, for example, chromosomal translocations. During chromosomal translocations different regions of two or more chromosomes end up in the same chromosome. Translocations are caused by breaks of the DNA, and are often observed in cells with a defective DNA damage response. Structural CIN is caused by replication stress [186], which can be caused by the amount of available nucleotides, but can also be influenced by cyclin oncogenes [24]. CIN can drive cancer progression and can be seen as a scan for karyotypes with proliferative advantage. The presence of CIN correlates with tumor progression and poor prognosis, in part, because heterogeneous populations of cells with a variety of copy number aberrations might be hard to target by drugs.

Single cell copy number analysis

Detecting DNA copy number genome-wide in single cells using DNA sequencing requires a protocol which samples the genome in a relatively *uniform* fashion. This allows inference of the *absolute* number of DNA copies for regions in single cells. With small bins which are densely populated across the genome, a *high-resolution* signal can be achieved. To determine which allele is lost, the haplotype of each read needs to be established. Protocols which generate longer fragments with high base-calling accuracy will be better at haplotyping. Once the copy number is determined, next steps can be taken to infer phylogenetic relationships between cells and match copy number profiles to other modalities, for example, the transcriptome or epigenome.

Computational copy number analysis can be separated in 3 steps: counting, abundance estimation and segmentation and lastly, phylogenetics.

Counting

By sequencing the DNA of a single cell, the genome present in the cell is sampled. When for two identically sized regions A and B, region A has more fragments detected than region B, it indicates that region A has more copies than region B. During the counting step the number of molecules per genomic bin (equally sized and regularly spaced regions across the genome) is estimated. The size of the genomic bins used is limited by the depth of sequencing and library quality, smaller bins will result in a higher resolution but require more reads which are evenly distributed across the genome. Currently, bin sizes used in single cell DNA sequencing analysis range from 2Mb down to approximately 10kb. Amplification or sequencing biases leads to reads being less evenly distributed which is usually a local effect. Using sufficiently large bin sizes will often average out this effect.

The copy number profile readout can be biased by the *Polymerase Chain Reaction* (PCR) amplification. Molecules with CG percentages[3] around 30% are easier to amplify and sequence, therefore resulting in more reads for these regions. If this effect is not taken into account, it will appear that regions with an optimal GC percentage for amplification have gained copies, and regions with a low or high GC percentage have lost copies. Normalization based on GC is a well described problem and not unique to single cell data. Computational solutions have been developed which try to reduce counts in regions with optimal amplification conditions and boost counts in regions with poor amplification conditions [28, 29, 23].

Another processes which causes over and under counting includes the length of the generated fragments. Regions which generate longer fragments will have fewer reads to sequence. When using *Unique Molecular Identifier* (UMI)s for de-duplication, errors in the UMI sequence can cause spurious counts because it

---

[3] The fraction of C and G nucleotides over the total in a genomic region

will seem that there are more molecules present than in reality available [163]. The quality of the copy number profile is very dependent on the way reads are mapped to the genome and therefore, improves upon removal of problematic mapping regions from the analysis [29].

The counting process results in a count matrix of cells by genomic locations. This matrix is an input to the segmentation and copy number estimation algorithms.

Copy number estimation and segmentation

Before the count matrix is biologically interpretable, copy number estimation and segmentation need to be performed. Copy number estimation and segmentation algorithms transform a count matrix into sets of segmented (continuous) parts of the genome which are linearly connected in each cell and a corresponding estimation of the number of copies for each segment present in each cell.

The segmentation algorithm tries to detect sudden changes in copy number. Sometimes read alignment information, such as split alignments, are used to identify DNA break point locations. Circular Binary Segmentation is often used [127] to segment copy number profiles. However, there are also alternatives such as Hidden Markov Models, which have the advantage to directly estimate a copy number value [14]. In addition to detecting a difference in copy number between segments, it is also challenging to estimate the integer copy number for each segment. A doubling from 1 to 2 copies, for example, can be interpreted as a change from 2 to 4 copies, because both cases will result in a doubling of signal between the segments. Some algorithms take available *Fluorescence-activated cell sorting* (FACS) ploidy information into account to resolve this issue [69]. Taking into account the haplotypes and read depth at the same time allows to more accurately resolve single cell copy number profiles [192]. *B-Allele Frequency* (BAF) frequencies can also be incorporated for detecting which allele is lost and for detecting LOH.

Copy number phylogenetics

In addition to estimating the copy number profile of a single cell, it is of interest to estimate the copy number profile of a population of single cells and infer the phylogenetic relationships between the cells. This requires an algorithm which performs segmentation on a population of cells, and is currently an open problem. The problem is hard for a number of reasons: segment boundaries have to be determined for the ensemble of cells and integration of (raw) copy number with haplotype information from multiple cells [182]. In shallow single cell DNA sequencing data, the location of a segment boundary cannot be determined exactly. An algorithm where segmentation and phylogenetic reconstruction are integrated is essential for performing a sound reconstruction [98]. For phylogenetics, the exact location of a break point can be used to identify lineages and it is therefore of importance to know if a segment is shared between cells in order to discriminate independent events. Some breakpoints are much less likely than others. A chromosome is much more likely to, for example, break on a centromere than on a gene dense region. Complete chromosome losses are even more common. Very common events are likely to happen multiple times in parallel. In summary, the infinite site hypothesis [91] which is often used in phylogenetics does not hold. Inference algorithms that allow for the same event to happen in parallel as well as back-mutations should be used [97, 98].

1.2 Single cell epigenomics

Epigenetics is not a well-defined term, but broadly refers to the genomic features, excluding the DNA sequence, which influence the phenotype of a cell. These epigenomic features constitute the epigenome. The strict definition of epigenetics requires that the epigenetic marks are heritable, but only few of the epigenetic marks discussed in this thesis show heritability over multiple cell divisions [176]. The epigenome regulates a cell state by long-term gene repression or activation and protects the genome from transposons. In this section, the detection and analysis of histone modifications and DNA methylation in single cells are introduced.

1.2.1 Single cell DNA methylation detection

DNA methylation is the addition of a methyl group to a DNA nucleotide. Both cytosine and adenine can be methylated in multiple configurations, but by far the most common and well studied methylated modification is 5-methylcytosine methylation. DNA methylation is known to regulate gene expression. Methylation of promoters make the promoter inaccessible to transcription factors and RNA polymerases, which are required for gene expression. DNA methylation is used, for example, to repress one of the X chromosomes during X-chromosome inactivation and for the long-term repression of transposable elements. While methylation levels and expression on promoter regions are negatively correlated, gene body methylation positively correlates with expression [190].

In human DNA, 5-methyl-cytosine is most commonly found in a *CpG* context, referring to a situation where two cytosines are present on opposing strands. The *p* refers to the phosphate which links any nucleotides together. The presence of two cytosines on the opposing strands allows CpG methylation to be heritable: when a genome is replicated, a new complementary DNA strand is formed which does not have DNA methylation yet. CpGs where only one of the two C's is methylated are called hemi-methylated. The unmethylated cytosine is then methylated by the *DNA Methyltransferase 1* (DNMT1) protein, which maintains the original methylation state over cell divisions [26]. Clusters of CpG sites in the genome are called CpG islands. These CpG islands are present in about 70% of all promoters and their presence is used to divide promoters into two classes [150]. Methylation states of the CpGs in a CpG island are highly correlated. Therefore, the methylation state of CpGs can be imputed based on the methylation state of nearby CpGs [197]. Deamination causes conversion of methylated cytosines into thymines over many generations which causes CpG sites to be relatively under-represented in the genome [196]. CpG methylation is also related to aging, as it increases over a life-time for about 2% of the CpGs and 0.5% of the non-CpGs [173]. Loss of DNA methylation is commonly found in cancer [199].

Multiple techniques have been published that measure DNA methylation using DNA sequencing. Four main classes can be distinguished; there are methods based on enzymatic digestion, which rely on proteins, such as MspJ1 and HpaII, to recognize methylated bases [154]. When using enzymatic digestion, the mapping location of the resulting fragments indicate which cytosine was methylated. The drawback of using a methylation-sensitive restriction enzyme is that there is no negative readout. When no methylation is present, these protocols generate no reads. Hence, the lack of methylation is indistinguishable from allelic dropout. The second class of methods rely on physically separating fragments with methylated residues. The third class relies on base conversion, where methylated or unmethylated bases are converted to other bases, which allows distinguishing methylated from unmethylated bases. For profiling the methylome in single cells, bisulfite sequencing is most commonly used [162,

45, 66], whereby all cytosine bases are converted to thymine except 5-methylcytosine residues. Bisulfite conversion damages DNA, resulting in fragmentation, loss of material and coverage bias. Whole genome bisulfite sequencing does, however, yield very large libraries which are costly to sequence at sufficient depth. Therefore, a reduced representation method, called reduced representation bisulfite sequencing (RRBS), has been developed which covers much less of the genome but enriches for CpG rich regions. This method requires fewer reads per cell in order to reach a similar sequencing depth for the most variable regions [117].
Recent alternatives to bisulfite sequencing are *Tet-assisted pyridine borane sequencing (TAPS)* [107] and *NEBNext® Enzymatic Methyl-seq (EM-seq™)* [174] which target 5-methylcytosine residues instead of unmodified cytosines and are less prone to biases caused by conversion.

The fourth class of methods measures the cytosines directly during sequencing. Currently, this approach has only been shown to work well in large bulk samples on the Oxford Nanopore platform. This method is likely to be the future for methylation detection because no conversion steps are required [158].

Computational challenges in single cell methylome analysis

The general computational workflow of single cell methylome analysis can be separated into 6 steps: demultiplexing, trimming, mapping, methylation calling, differential methylation calling, and a clustering and imputation step. For protocols where multiple samples are sequenced from the same pool of DNA, *demultiplexing* assigns reads to a cell of origin. Single cell bisulfite protocols are notorious for generating chains of fragments, so called concatamers, which are not derived from the same genomic locations. The effect of concatamers is alleviated by *trimming* off the known adapter sequences. Additionally, bases with low base-calling confidence are usually removed from the sequenced reads.

Bisulfite conversion converts most cytosine bases to thymines. When *mapping* the resulting reads to a normal reference genome, the mapping rate is very low, partially due to the high edit distance caused by the conversion of all cytosines. To alleviate this, the trimmed reads are mapped to a reference where all cytosines are converted to thymines. For protocols were not all cytosines are converted, such as TAPs, using a converted genome is not required.

During *methylation calling*, the aligned reads are scanned for evidence of a cytosine being methylated or unmethylated [94]. For most protocols, the strandedness of the read can be determined and used to find which strand contained the methylated base. Additional information is stored, such as the number of reads covering the location, the confidence of the call, the strand and the haplotype/allele and the methylation context. This process results in methylation calls for all covered cytosines in the genome. Some locations will not get a methylation call due to missing data or a low number of covered cytosines. For a single cell DNA sequencing library, the amount of methylation calls is usually low. To reduce this sparsity, methylation calls are often binned in larger genomic bins. To *cluster* single cell methylation data, the Frechèt distance metric is commonly used to calculate distances between cells [175] followed by clustering and complete linkage clustering with weighted Euclidean norm [162].

Differential methylation calling tries to identify regions which are differentially methylated, either between single cells or between cells which have been clustered together (pseudobulk). The simplest way to calculate differentially methylated regions is to use a Fisher exact test or by logistic regression [40]. Not taking into account neighboring bins, is however, a drawback of both methods, as neighboring bins are frequently correlated and can potentially help to boost statistical power. To overcome this drawback, smoothing strategies which incorporate the information of multiple CpGs or bins are regularly used [76].

Imputing sparse single cell methylation data beforehand often results in better clustering. Single cell imputation algorithms try to guess the methylation state of uncovered CpGs. This is done by taking into account the state of nearby CpGs, because nearby CpGs are generally correlated. Additionally, the DNA sequence context of a certain window size (1kb) is also used to identify sequence motifs which are related to methylation and aid in imputation accuracy [6]. Also, higher order features like accessibility, histone modifications and expression can be used to perform more accurate imputation [88, 114].

1.2.2 Histone modifications

The DNA of a cell is tightly packed using approximately 30 million nucleosome protein complexes. These nucleosomes help to protect the DNA and reduce the amount of space required for storage. In addition, nucleosomes have important regulatory functions. A mechanism of regulation is a result of the DNA wrapped around the nucleosomes (core-DNA), which makes it less accessible to DNA binding proteins. This, in part, explains the large difference in nucleosome density between transcription start sites of expressed versus repressed genes. Expressed genes commonly have a nucleosome depleted transcription start site [152]. Approximately 146bp of DNA is wrapped by 1.7 turns around a nucleosome, which enters and exits at the H1 linker histone. DNA which is not wrapped around a nucleosomes is called linker-DNA, and ranges from a length of 10bp to hundreds of bp. Without additional energy input, nucleosomes are positioned by sequence preference with approximately a 10.4bp period, which relates to the twisting of the DNA strand and interfacing with the nucleosome [137]. A/T rich sequences with minor grooves face the nucleosome, while the minor grooves of C/G rich sequences point away from the nucleosome. Presence of DNA methylation also influences nucleosome positioning [48].

Nucleosome complexes usually consist of two copies of the 4 core histone proteins (H2A, H2B, H3, H4) and one linker histone (H1). These histones can be modified post-translationally with various modifications: methylation, acetylation, ADP-ribosylation, ubiquitination, citrullination and phosphorylation [17]. The long protruding tails of histones are especially commonly modified. All modifications together constitute a *histone code* which signal histone modification reader proteins and are sometimes used to recruit proteins to the genome. Some modifications are mutually exclusive due to their overlapping residues: mono-, di- or tri-methylation of the same residue can have a totally different meaning. For example, $H_3K_4me_3$ marks active transcription start sites, while $H_3K_4me_1$ marks wider domains like active enhancers and promoters (Figure 2).

The highest classification of histone modifications separates them into two main classes, repressive and active modifications. The repressive modifications are associated with tightly packed chromatin states, called the heterochromatin. Generally, genes and transposable elements [179] on the heterochromatin are not expressed. The active modifications are associated with a more open chromatin state and active transcription. An overview of some commonly methylated residues of the H3 subunit are shown in table 2.

The most common technique for detecting histone modifications in bulk is *Chromatin Immuno Precipitation* (ChIP)seq. In ChIPseq, antibodies are targeting DNA fragments with nucleosomes containing specific histone modifications. These fragments are pulled down and sequenced. An alternative method,

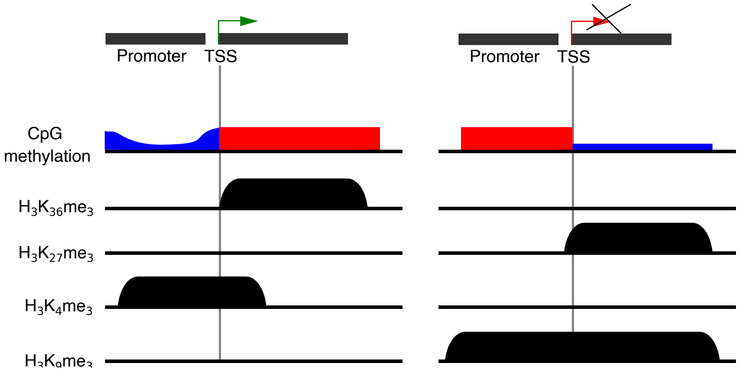


Figure 2: Schematic overview of CpG methylation and histone marks on the promoter and gene body of expressed and repressed genes. CpG methylation of promoters of expressed genes is low in order to be accessible and allow for transcription initiation. The CpG methylation of the gene body of expressed genes is elevated, likely due to increased accessibility caused by transcription. Transcribed genes accumulate $H_3K_{36}me_3$ on their gene body and $H_3K_4me_3$ on their promoter. $H_3K_{27}me_3$ and $H_3K_9me_3$ are primarily present on repressed genes.

modification	Function
H3K4me1	poised epigenetic state [13]
H3K4me3	transcription [83]
H3K9me3	transcriptional silencing and repression, AT rich gene poor regions
H3K27me3	transcription repression and X inactivation, GC rich gene dense regions, DNA damage response
H3K36me3	transcription elongation
H3K79me3	euchromatin, transcription elongation, checkpoint response

Table 2: Overview of some common histone H3 methylation modifications

sensitive enough for low input material, uses *Micrococcal Nuclease* (MNase) tethered to an antibody which guides the MNase to a specific histone modification. The MNase cleaves nearby linker DNA, the resulting fragments are sequenced, and the mapping locations inform where the histone modification was present. Another method suited for low-input material, uses an antibody which is tethered to a protein A-Tn5 transposase fusion protein. In a process called tagmentation, the fusion protein simultaneously cleaves DNA near the histone mark and adds the adapters required for DNA sequencing. Sensitive methods for measuring histones through single cell sequencing are based on using either MNase or Tn5 transposase. See Table 3 for an overview of current methods. All these methods generate reads at locations nearby the histone mark or DNA binding protein of interest. These reads are analyzed to find regions which are differentially modified, and are used to cluster the single cells.

Computational challenges in single cell histone modification analysis

The mundane question which is asked for a single locus is simple: does a nearby nucleosome contain the mark of interest or not? For some cells, a read will cover the locus of interest and therefore, confirms a nucleosome with the modification nearby. For most cells, the question cannot be answered, as there are no reads covering the locus. This either means there was indeed no modification nearby, or the measurement dropped out. Currently, the number of dropouts is very substantial for any of the protocols measuring histone modifications. The main computational challenges of single cell histone modifications detec-

	Method	Year of introduction
Single cell ChIP		
	SC-ChIP [141]	2015
	ht-sc-ChIP-seq [73]	2019
	itChIP-seq [3]	2019
pA-MNase		
	CuT&Run [161]	2018
	uliCUT&RUN [74]	2019
	scChiC-seq [95]	2019
	iscChiC-seq [96]	2021
pA-Tn5		
	ChIL-seq [77]	2019
	CoBATCH [181]	2019
	Cut&Tag [89]	2019
	autoCUT&Tag [86]	2021
	scCut&Tag [16]	2021
	scCUT&Tag [187]	2021

Table 3: Overview of single cell histone profiling methods

tion are therefore similar to single cell genome sequencing and are related to dropouts in the resulting sparse data. In principle, handling this data is quite similar to single cell transcriptome libraries, as there will be some reads at locations with signal (transcripts vs a nearby histone modification of interest). In contrast to transcriptome data, where highly expressed genes have many templates (transcripts) available to amplify and sequence, histone modifications only have few templates available (2 DNA templates for a non-dividing diploid cell). Most histone modifications are deposited in (spatially side by side and dense) clusters. This results in spatially correlated measurements, allowing for the use of binning or smoothing to combat the sparsity. For some histone modifications the location of the domains are known. $H_3K_{36}me_3$ is, for example, found on gene bodies and can therefore be quantified per gene. $H_3K_4me_3$ can be quantified per promoter. Latent Dirichlet Allocation [38] or Latent Semantic Indexing [106] are used for dimensionality reduction of sparse data while simultaneously imputing missing values.

1.3 Combining it all together: single cell multi-omics

Heterogeneity of tissues has been extensively studied using single cell omics. As described in the previous sections, it can be used to find subclones in a tumor, reveal lineage relationships, identify subpopulations of cells on the basis of expression and to identify dynamics, DNA methylation and histone modifications.

Almost all components in a cell interact and affect each other. Hence, there are many questions to be asked about the relationships between various modalities. For example, how do DNA copy number changes influence gene expression? How often does the same copy number aberration occur in the same clone in a tumor? How does methylation affect transcription, and vice versa? A way to answer questions like these is to use single cell multi-omics, where multiple modalities are measured in the same population of cells

Single cell multi-omics datasets can be divided into two major classes [99]. In the first class of datasets [110, 111, 57, 7, 42], multiple modalities are measured from the same cell. The resulting data is usually more sparse than when modalities are measured individually. This sparsity can be caused by competition of the modalities during the protocol or sequencing process. Balancing multiple modalities such that enough information can be extracted from all of them simultaneously has been proven difficult. Additional technical biases can occur when measurements are able to influence one other. For example, if both modalities are derived from genomic DNA, the detection of the modalities in the same region can cause competition and dropout of one of the modalities for that region.

There are many ways to analyze multi-omics data, and the best strategy depends on the specific modalities and questions at hand. Typically, in single cell experiments the aim is to detect subpopulations of cells through, for example, clustering. Clustering can be performed on one of the modalities, as, for example, on the copy number profile per cell. If properly analysed, these clusters reveal subpopulations which are biologically meaningful, such as in the case of copy number profiles, the clones present in a sample. When using the biologically meaningful subpopulations, the remaining modalities can be used to identify differences between the populations [99]. This is especially useful when a modality by itself is more difficult to cluster due to sparsity or novelty, causing unavailability of appropriate distance metrics. In case of a combined protocol with bisulfite or histone modification data, for example, it makes sense to first cluster on the very well characterized modality which is in most cases related to cell types identity, such as transcriptome data. Then, the DNA methylation can be investigated per cell type. It is also possible to cluster using the information contained in all modalities. Such methods are currently in development but will likely work similar to their counterparts used to analyse bulk multi-omics data [160, 139, 8].

In the second class of single cell multi-omics datasets the modalities of interest are measured in different cells. These cells are sampled from the same population of cells. This approach can be beneficial as there is no need to set up an experimental protocol which is able to measure multiple modalities in the same single cell. The drawback is that the modalities require to be aligned or linked to one other. How this is done depends on the types of modalities to be measured, and whether a third modality (for example FACS information) is available which is shared between all cells.

1.4 Thesis outline

This thesis contains three research chapters. The technological foundation of each chapter is measuring multiple modalities in single cells by single cell sequencing.

Chapter 2

Here we study Clonal dynamics in colorectal cancer by evolving a colon cancer organoid model over 100 generations simultaneously monitoring clone size, *Copy Number Variant* (CNV)s, and single nucleotide variants SNVs in individual cells. These integrated measurements reveal the order of events in which chromosomal aberrations occur and allow the identification of aberrations that recur multiple times within the same population. We observe recurrent sequential loss of chromosome 4 after loss of chromosome 18 in multiple unique tumor clones and show this reflects clinical observations.

Chapter 3

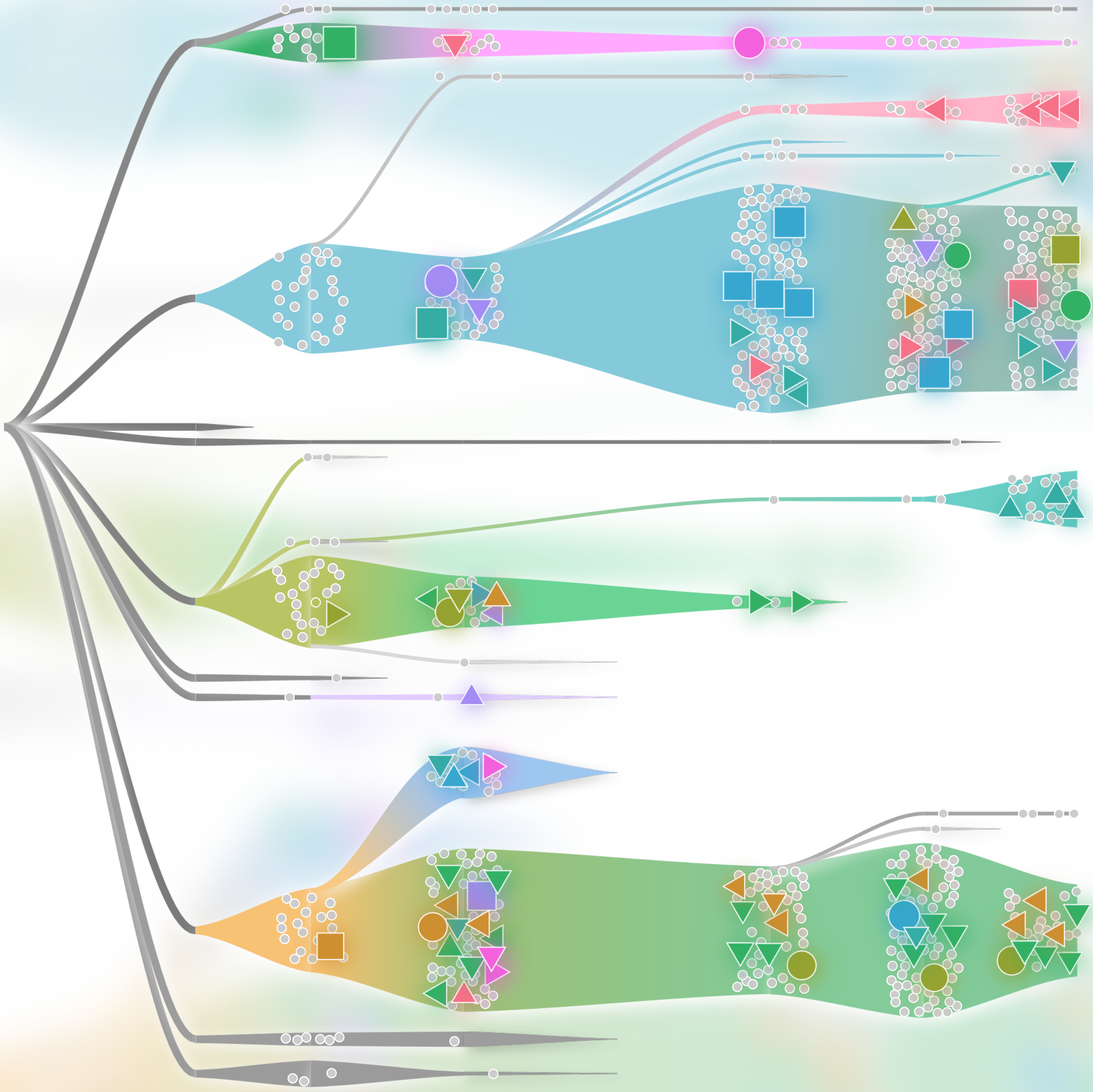
In this chapter, a new technique scSort-ChIC is introduced. ScSort-ChIC can be used to profile histone mark locations in single cells and allows pairing with FACS information. scSort-ChIC is used to map active and repressive histone modifications in *Hematopoietic Stem and Progenitor Cells* (HSPCs), and mature blood cells in the mouse bone marrow. During differentiation, HSPCs acquire distinct active chromatin states that depend on the specific cell fate, mediated by cell type-specifying transcription factors. In contrast, most regions that gain or lose repressive marks during differentiation do so independent of cell fate. Joint profiling of $H_3K_4me_1$ and $H_3K_9me_3$ demonstrates that cell types within the myeloid lineage have distinct active chromatin but share similar myeloid-specific heterochromatin-repressed states. This suggests hierarchical chromatin regulation during hematopoiesis: heterochromatin dynamics define differentiation trajectories and lineages, while euchromatin dynamics establish cell types within lineages.

Chapter 4

In this chapter, a new technique to profile both histone mark locations DNA-methylation and FACS properties from the same single cell is introduced. This combination of measurements has never been performed before. The method is thoroughly validated, and applied to a system where the cell cycle can be closely monitored. The FACS information is used to integrate data from multiple histone marks and compare their behavior during the cell cycle.

Chapter 5

I conclude this thesis in chapter 5 where I will discuss present-day challenges and possible solutions to combat these challenges and a bit of future outlook.



Chapter 2

Integration of multiple lineage measurements from the same cell reconstructs parallel tumor evolution

Accepted in Cell Genomics

Lennart Kester[1,4], Buys Anton de Barbanson[1,2,4], Anna Lyubimova[1], Li-Ting Chen[1,2], Valérie van der Schrier[1], Anna Alemany[1], Dylan Mooijman[1], Josi Peterson-Maduro[1], Jarno Drost[3], Jeroen de Ridder[2,*] and Alexander van Oudenaarden[1,5,*]

[1]Oncode Institute, Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, The Netherlands.
[2]Oncode Institute, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands.
[3]Oncode Institute, Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands.
[4]These authors contributed equally
[5]Lead contact

*Correspondence: J.d.R. (J.deRidder-4@umcutrecht.nl) and
A.v.O. (a.vanoudenaarden@hubrecht.eu)

2.1 Summary

The dynamics of tumorigenesis shares many similarities with Darwinian evolution. Elevated mutation rates result in genetically diverse tumor sub-populations. Selection leads to clonal expansion of the fittest genotypes resulting in tumor outgrowth and dissemination. Here we mimic this process by evolving a colon

cancer organoid model over 100 generations simultaneously monitoring clone size, copy number variants (CNVs), and single nucleotide variants (SNVs) in thousands of individual cells. This offers the unique opportunity to combine multiple measurements to reconstruct evolution. These integrated measurements reveal the order of events in which chromosomal aberrations occur and allow the identification of aberrations that recur multiple times within the same population. For example, we observe recurrent sequential loss of chromosome 4 after loss of chromosome 18 in multiple unique tumor clones closely reflecting clinical observations. organoid evolution models complemented with integrated single-cell sequencing technology provide a powerful platform to study and hopefully control tumor evolution.

2.2 Introduction

Cancer initiation is the result of cells gradually acquiring genetic alterations due to carcinogenic exposure and DNA replication infidelity [63]. Some alterations confer a growth advantage resulting in tumor formation in which DNA repair and genome integrity are increasingly ablated leading to further genetic alterations. As a result, tumor clones arise that exhibit distinct genetic compositions of single nucleotide variants (SNVs), insertions/deletions (indels) and copy number variants (CNVs). This intra tumor heterogeneity (ITH) [71, 85, 125, 155], plays a key role in cancer development. Tumor clones may for instance have varying proliferative and metastatic potentials. Furthermore, ITH plays a key role in therapy resistance and the frequent lethal outcome of cancer, since some tumor clones may have intrinsic resistance to therapy [115, 115, 116]. For this reason, substantial efforts have been made to characterize ITH by mapping the clonal evolution in tumors using whole genome sequencing (WGS) of the tumor bulk. However, without temporal resolution this approach only provides a limited view on ITH. It is for example difficult to infer the order in which genetic lesions have occurred and the number of subclones that can be identified is limited [1, 55, 142, 172]. While regional sequencing may alleviate some of these issues, it only gives a snapshot of this heterogeneity and does not reveal the evolution of the cancer [164].

To characterize ITH at high clonal and temporal resolution, genetic alterations in single-cells and multiple time-points need to be obtained. This is enabled by single-cell DNA sequencing, which was first introduced in 2011 [121] and has since been used to investigate tumor heterogeneity and tumor evolution in many studies [25, 39, 53, 70, 121, 140, 184, 188]. However, most of these studies construct clonal evolution trees exclusively based on either SNVs ([140, 188] or CNVs [25, 39, 53, 121], and thus never exploit the combined information contained in both types of genetic alterations. For instance, recurrent chromosome amplifications or deletions within the same sample cannot be distinguished without knowledge of both the CNVs and the SNVs from the same single cell. Furthermore, the order of events in which CNVs are established can only be determined with high certainty by combining CNV data with SNV data. Constructing trees based on multiple independent lineage tracing strategies moreover enables internal validation by evaluating consistency across multiple independent lineage markers. Most single-cell efforts to date that characterize clonal evolution trees are limited in the number of cells and/or the number of SNVs that are interrogated and do not include multiple time-points. A few existing studies do combine SNVs and CNVs measurements from the same single cells in tumors [70, 184]. However, these studies rely on a limited number of SNVs and suffer from low resolution CNV data, resulting in shallow trees that lack the resolution to acquire a complete picture of the clonal evolution of these tumors.

In this study we combine high resolution CNV data with high quality SNVs from thousands of single cells in order to improve the ability to delineate accurate clonal evolution trees. To increase the accuracy of the clonal evolution trees even further we added a viral barcode-based lineage tracing strategy in addition to single-cell DNA sequencing-based detection of SNVs and CNVs. This provides three (CNVs, SNVs and the lineage barcodes) complementary levels of lineage tracing which can be integrated to acquire a complete view of tumor evolution and which allows internal validation of the resulting clonal evolution trees. We moreover introduce a temporal axis to the data by taking multiple time points during clonal evolution.

We utilize a colon carcinoma organoid model, which has the ability to acquire many CNVs, and allows the introduction of viral lineage tracing barcodes. Colon carcinoma is frequently initiated by mutations in Wnt, Epidermal Growth Factor Receptor (EGFR), P53 and Transforming Growth Factor (TGF)-β signaling pathways. Furthermore, colon carcinoma is often associated with chromosomal instability (CIN) resulting in widespread CNVs [62]. Recent studies have shown that the formation of colon carcinoma can be accurately mimicked *in vitro* using an organoid model [60]. This organoid model uses CRISPR-Cas9 to induce sequential mutations in *APC*, *TP53*, *KRAS* and *SMAD4*. *APC*$^{-/-}$ *TP53*$^{-/-}$ *KRAS*$^{\mathrm{G12D}}$ *SMAD4*$^{-/-}$ (APKS) organoids morphologically and phenotypically resemble carcinoma stage colorectal tumors and have pronounced CIN. The CIN results in genetically heterogeneous cultures mimicking the ITH observed in clinical samples of colorectal carcinoma (CRC) [31]. Copy number aberrations common to the APKS organoids and CRC patients include chromosome 4, 18 and 8. [156, 123, 36, 51],.

Chromosome 4 and 18 deletions are common in tetraploid colon cancer tumors [56]. These properties make the organoid model an interesting system to study ITH and tumor evolution.

After introduction of the viral barcodes into the organoid model at the beginning of the experiment, the organoids undergo a 26-week period of *in vitro* evolution, during which single-cell DNA sequencing is performed at multiple time points in order to detect CNVs, SNVs and viral lineage barcodes. In parallel, the relative amount of each tumor clone was analyzed weekly through bulk sequencing of the lineage barcode. This combination of lineage measurements allows the construction of unprecedentedly detailed clonal evolution trees. Furthermore, due to the combination of three markers (CNVs, SNVs and the viral lineage barcodes) internal validation of the constructed trees can be performed, as trees constructed based on two markers can be validated by the third. The resulting clonal evolution trees reveal the order of events in which chromosomal deletions and amplifications occur and allow the identification of chromosomal aberrations that occur multiple times independently within the same cell population. Based on our trees we could for instance establish that chromosome 4 and chromosome 18 are sequentially lost in multiple unique tumor clones. Such findings can have important clinical implications as we find that this combination of chromosome 4 and chromosome 18 deletions not only provides a strong proliferative advantage in our experiments, but also results in significantly worse recurrence free survival compared to patients with only a single chromosome 18 deletion.

2.3 Results

Lineage tracing in colon carcinoma organoids reveals clonal dynamics

To track the clonal evolution of colon carcinoma we used early passage human-derived APKS colon organoid cultures in triplicates (hereafter referred to as Replicate 1, Replicate 2 and Replicate 3). After establishment of the organoid line, a viral lineage library with around 60,000 unique barcodes was introduced. The organoids subsequently underwent a 26-week (25 passages) *in vitro* evolution period (**Figure 1**). The relative abundance of each of the viral barcodes was analyzed weekly through bulk sequencing of the lineage barcode. By assessing the relative barcode abundance in all three replicates we observed rapid expansion of a relatively small number of clones (**Figure S1**). An important question is whether the observed dynamics could be explained by neutral drift in the culture. To exclude this possibility, we performed stochastic simulation of the organoid culture (**Methods**), taking into account the proliferation rate of the organoids, the number of individual cells at the start of the experiment (start population size) and the number of cells that is left after passaging of the organoids (bottle neck size). We observed that the decrease in entropy in the actual experiment is significantly faster than in the simulations indicating that these clonal dynamics could not be explained by neutral drift in the culture (**Figure S2**), indicating there is a selection process underlying the clonal dynamics. Furthermore, we observe expansion of the same clones (as determined through CNVs) in multiple replicates, providing additional support for a strong selective pressure on the organoids.

High resolution CNV detection allows identification of many unique CNV states

In parallel to the bulk analysis of the viral lineage barcodes, single cells were harvested at regular intervals and processed for single-cell DNA sequencing using an NLA-III restriction enzyme-based technique [93, 119], (**Figure 2A**). In short, adapters containing Unique Molecule Identifiers (UMI), allowing quantification of the absolute number of unique molecules in each single cell, are ligated to the NLA-III cut sites and the molecules are amplified using In Vitro Transcription (IVT) prior to sequencing. After binning the mapped reads using

500 kb bins and filtering of cells with too few reads or fragmented genomes, a total of 1641 cells with a mean number of 326,000 unique molecules per cell remained (**Figure 2B**). Copy number profiles were normalized by dividing by the median and multiplying by 2. Dimensionality reduction using principal component analysis on the median normalized matrix shows that the cells cluster by replicate and time point, while for early time points the cells of the various replicates are more similar (**Figure S7**). Recurrent breakpoints between regions with different copy number were detected by hierarchical clustering followed by circular binary segmentation (**Methods**). The high resolution and low noise (**Figure S3**) of the NLA-III restriction enzyme-based technique allows accurate quantification of the CNV profile for each single cell. For instance, we observed multiple unique CNVs affecting chromosome 18, which could not have been detected through bulk WGS (**Figure 2D**; for example: copy number states 2, 6 and 11).

In parallel to single-cell NLA-III sequencing we performed standard bulk WGS at the start and at the end of the *in vitro* evolution period. In this bulk data we observed a full deletion of chromosome 4 in Replicate 1 at the end of the experiment. However, the B-allele frequency (BAF) revealed that both alleles were still present albeit in unequal amounts (**Figure S4**). This indicated that a fraction of cells had lost one allele of chromosome 4, while the rest of the cells had lost the other allele. To confirm this in the single cells we first acquired the diplotype of chromosome 4, based on another organoid line derived from the same donor which had completely lost one of the alleles of chromosome 4. This diplotype was then used to assess which allele (if any) of chromosome 4 was lost in each of the single cells. Indeed, in Replicate 1 we observed 314 single cells with a loss of chromosome 4 allele A and 96 cells with a loss of chromosome 4 allele B. The diplotype for chromosome 18 could also be acquired. Here we observed that all the unique deletions on chromosome 18 concern the same allele. The observed single cell BAFs of chromosome 4 and 18 are tri-modal with peaks around 0, 0.5 and 1, indicating the cells were diploid and not tetraploid (**Figure S4B**). These observations demonstrate that the combination of single-cell NLA-III sequencing and WGS allows allele specific CNV detection in single cells. Most of the deletions and amplifications observed in the organoid culture are also frequently observed in patients in CRC, emphasizing the relevance of the organoid model for studying colorectal cancer (**Figure 2C**).

In total we identify 25 unique CNVs across 1641 single cells, with 52 unique CNV states (a CNV state is a genome wide CNV profile that is shared by at least three single cells; **Methods**), ranging in size from 434 cells to 3 cells (**Figure 2D**). In Replicate 1 we observe a massive expansion of cells with a chromosome 4 and a chromosome 18 deletion, while in the second (Replicate 2) and third (Replicate 3) experiment we observe expansion of cells with a chromosome 8p deletion.

Construction of High-Resolution Clonal Evolution trees

To establish clonal evolution trees, a directed edit distance graph was created from the CNV states. Since the same CNV state can be present at multiple time points, each time point was added as a separate node in the graph. This enables enforcing temporal consistency (i.e. earlier time points could not be derived from later time points) in the tree construction (**Methods**). A spanning arborescence was extracted from the directed CNV edit distance graph using Edmonds algorithm [61]. The clonal evolution trees were visualized using ToverBoom (**Figure 3C-E, Methods**). The resulting clonal evolution trees indicate the most likely evolutionary trajectories along which the tumor has evolved. The tree for Replicate 1, for instance, indicates that CNV State 3 is a descendant of CNV State 2, which is logical considering that CNV State 2 has a chromosome 18 loss and CNV State 3 has a chromosome 18 and a chromosome 4 loss (**Figure 2D**). In conclusion, the high resolution CNV calling allows construction of detailed clonal evolution trees with a temporal component.

Integration of CNV states with an independent lineage marker is required for confirmation of clonal evolution trees

Although the clonal evolution trees indicate the most likely evolutionary trajectory, it cannot be excluded that two seemingly related CNV states arose independently, in particular given that copy number changes occur frequently in this genetic background. To disambiguate the relation between two CNV states we can leverage the information provided by the viral lineage barcode. This is schematically represented in **Figure 3A-B**. For a new CNV state to be introduced during the experiment, both the cells in the new CNV state and its parental cells must be marked by the same viral lineage barcode. For example, the new CNV state *k* arose from the parental CNV state *l* because both states share viral lineage barcode 1 (**Figure 3A**). On the other hand, CNV states observed in cells that do not share a viral lineage barcode with their putative parent most likely arose prior to the introduction of the lineage markers (illustrated by CNV state *m* in **Figure 3A**). Similarly, a CNV state that contains cells with multiple lineage markers most likely also arose prior to barcode introduction (CNV state *l* in **Figure 3A**).

A clone in a certain CNV state harboring multiple lineage markers, which are also present in their inferred parental CNV state (for example CNV state *y* and CNV state z in **Figure 3B**) are particularly interesting. In this example, CNV state z and CNV state *y* both harbor viral lineage barcodes 1 and 2, indicating that these CNV states are closely related and arose after lineage marker introduction. Since the viral lineage barcodes mark unique lineages, this implies that the loss of chromosome A occurred twice independently, once in a cell with lineage barcode 1 and once in a cell with lineage barcode 2. An alternative explanation for the observation of CNV states sharing lineage barcodes is that these CNV states were already present at the start of the experiment and that the same lineage barcode was introduced multiple times into cells with these CNV states. However, this is very unlikely due to the large number of barcodes present in the viral lineage library (based on simulations the probability of two cells in the starting culture receiving the same viral barcode is smaller than 0.0001).

To detect the viral lineage marker in single cells, allowing us to disambiguate and validate clonal evolution trees, we employed an experimental strategy that enriches for reads containing the viral lineage barcode (**Methods, Figure S6**). This allowed us to detect the lineage barcode for 293 of the sequenced single cells. The lineage barcode information can be superimposed on the clonal evolution trees, which allows us to distinguish between the cases described previously (**Figure 3A-B**). Indeed, we observe shared viral lineage barcodes between several CNV states that are descending from each other according to the CNV based clonal evolution trees. For instance, CNV state 2 (chromosome 18 deletion) and CNV state 3 (chromosome 18 and 4 deletion) share a viral lineage barcode BC1 (**Figure 3C**). This confirms that during the course of the *in vitro* evolution a single cell with BC1 belonging to CNV state 2 lost the A allele of chromosome 4 thereby founding CNV state 3. A shared viral lineage barcode BC2 was also observed between CNV states 3, 5, 16, 20 and 21, all of which share a chromosome 18 deletion. This indicates that, even though this particular viral lineage barcode is not observed in CNV state 2, it must have been present and was most likely not observed due to sampling.

Other examples of new CNV states arising during the *in vitro* evolution period include CNV state 7, 17, 22 and 27 from CNV state 1 in Replicate 2 (**Figure 3D**) and CNV state 9 from CNV state 2 in Replicate 3 (**Figure 3E**). Interestingly, in Replicate 2, the CNV based clonal evolution tree suggests that CNV state 17 and 22 arose independently from CNV state 1. Based on the viral lineage barcodes we can conclude that this is false and that CNV states 17 and 22 in fact are descending from CNV state 1. This illustrates the importance of integrating multiple lineage measurements to achieve an accurate picture of tumor evolution.

Besides observing multiple CNV states sharing the same viral lineage barcode, we also observe multiple viral lineage barcodes within the same CNV state (e.g. CNV states 1, 2, 4 and 6). The most likely explanation for this is that these CNV states were already present at the moment of viral lineage barcode introduction (**Figure 3A**). The observation that CNV states 1, 2, 4 and 6 were already present at the start of the experiment is confirmed by the fact that these states are all present in multiple replicates. Furthermore, we already observe a subclonal chromosome 18 loss (CNV state 2) in the bulk WGS from samples taken at the start of the experiment (**Figure S4D**).

Somatic Single Nucleotide Variants provide an additional layer of information to increase tree resolution

Somatic single Nucleotide Variants (sSNVs) can be used as lineage markers as they are inherited from one cell to its progeny. Shared sSNVs thus indicate a shared common ancestor and sSNVs can be used to disambiguate phylogenetic relationships between previously identified CNV states. While copy number alterations are likely to have a fitness effect, most sSNVs are passenger mutations without any effect on the fitness of the cells, thus providing a lineage marker which is less affected by selection. Moreover, unlike viral lineage markers, sSNVs accumulate throughout time, and therefore provide lineage marking of clones which initiate during the evolution experiment. For these reasons, in addition to the viral barcodes and copy number profiles, we assess sSNVs within the cells of all three replicates. This additional layer of information allows us to identify additional heterogeneity within the population of cells with the same copy number state, verify edges of the inferred lineage trees based on the copy numbers and identify copy number aberration events which occurred multiple times.

sSNVs called from single cell sequencing data suffer from high numbers of false positive calls. To identify reliable sSNVs, we therefore trained a random forest (RF) classifier on the somatic mutations that could be verified in the bulk-library and used the trained classifier to identify reliable sSNVs (**Methods**). Every variant that passed the RF filter is phased to at least one heterozygous germline variant or discarded otherwise. Positive variant calls are identified by presence of the alternative allele among all sequence reads for the position within a cell. Negative variant calls are identified by presence of the reference allele in phase with the germline variant found to be linked with the alternative allele [30, 78]. This procedure allowed us to extract 106 high-quality sSNVs from all cells used for tree inference.

sSNVs can be overlaid on the lineage trees inferred from the copy number calls. **Figure 4A-B** show this for two example sSNVs across all three replicates. The sSNV can be present (red markers), absent (the reference allele is detected, blue markers) or undetermined (insufficient coverage to be detected, grey markers). This example shows that subclones of the $\Delta18$ clone always carry the variant, while the $\Delta8$ clone including subclones do not. This confirms that there is strong association of these two SNVs to clones with a similar copy number profile.

By clustering based on all detected sSNVs the cells separate in two main groups. The first group of cells is characterized by chromosome 18 loss, while the second predominantly carries a chromosome 8p deletion (**Figure 4C**). Most variants are detected in multiple replicates, which indicates that the variant was likely present before the replicates were separated.

Integrating sSNVs with CNVs in single cells suggests parallel evolution of copy number states

In addition to the strong co-segregation of copy number state and somatic variants (**Figure 4C**), we also observe more complex relations between the two lineage markers. We define three classes of sSNVs at the branching point of two copy number states (**Figure 5**).

In the first class, a copy number aberration occurs after the sSNV. This would be consistent with a situation in which at an early time point a clone is marked by an sSNV, while after the CNV-induced bifurcation the newly derived clone only contains cells carrying the alternative allele (**Figure 5A, top**). We find examples of this first class on Chr8 and Chr18 for Replica 2 and 1, respectively (**Figure 5A, bottom**).

In the second class, an sSNV is introduced after a CNV is created. Here, at early time points the clone does not contain any cells with the sSNV. After branching into a new copy number state, the sSNV is exclusively observed for the cells in the new CNV state, indicating it must have been introduced after the acquisition of the CNV (**Figure 5B, top**). This class of sSNVs can be used to verify edges in the copy number tree similar to the viral lineage markers. Examples of the second class are shown for Replicate 1 occurring within chr18-loss subclones (**Figure 5B, bottom**).

In the third, most interesting, class the same CNV occurs twice independently. This situation would be consistent with a clone in a single CNV state that contains cells both with and without an sSNV at an early time point. If after the introduction of a CNV the new clone also contains cells with and without the sSNV, this must mean that the copy number aberration must have occurred at least twice; once in the clone with the sSNV and once in the clone without the sSNV (**Figure 5C, top**). An example for such a parallel evolution event can be found in multiple variants that show both the reference and mutated alleles in the Δ18 state and the Δ18Δ4a state (**Figure 5C, bottom**). The same holds for the Δ18 to Δ18Δ4b state (**Figure 5C, bottom**).

Accurately identifying the first two classes is challenging, because it is always possible that the presence or absence of a particular sSNV is not detected because of drop-outs in the single-cell data. However, distinguishing the third class from the first two classes is less vulnerable to sampling errors. If both the presence and absence of an sSNV are detected before and after a CNV is initiated, it rules out the first and second scenario, in particular when this is supported by several sSNVs. Taken together, an integrated analysis of how sSNVs segregated between copy number states suggests that these copy number states can arise multiple times independently.

Loss of chromosome 18 followed by loss of chromosome 4 worsens survival probability

The most highly abundant and fastest growing clone across the 3 replicates was characterized by a combination of a loss of chromosome 18 and a loss of chromosome 4. Although we did observe cells with a chromosome 18 loss only, we did not observe cells with a chromosome 4 loss only, suggesting that the loss of chromosome 4 only results in a proliferative advantage in the presence of a chromosome 18 loss. To see if there is any evidence that supports this hypothesis we turned to the Memorial Sloan Kettering Colorectal Cancer (MSKCC) database [189]. The MSKCC data show that in colorectal cancer patients chromosomes 18 and 4 are frequently lost (**Figure 2C**).

However, tumors with a lower copy number ratio for chromosome 4 than for chromosome 18 occur less frequently in the same patient than can be expected based on the frequencies of chromosome 18 and chromosome 4 copy numbers (based on a permutation strategy, **Methods, Figure 6A**). At the same time, there is an enrichment for tumors wherein the copy number ratio for chromosome 18 is lower than for chromosome 4, indicating that most often a chromosome 18 deletion occurs prior to a chromosome 4 deletion. Strikingly, this is in line with the order of events we observe in the APKS organoid cultures.

We also find that patients with a combination of a chromosome 18 and a chromosome 4 deletion have higher mortality than patients with a chromosome 18 deletion without a chromosome 4 deletion (**Figure 6B**). This suggests that a chromosome 4 deletion in the context of a prior chromosome 18 deletion results in a deadlier tumor than either deletion on its own. To investigate this in a more systematic manner we investigated all possible combinations of two chromosomal deletions and/or amplifications. We define a ‘priming event’, which is the first aberration and a ‘conditional event’ which is the second aberration in the context of a particular priming event. We then compared the absolute correlation between the copy ratios for any given pair of priming and conditional events to the hazard ratio of the conditional event over the priming event alone (**Figure 6C**). This analysis shows that a loss of chromosome 18 as priming event followed by a loss of chromosome 4 as conditional event has the highest hazard ratio of all possible conditional events.

These analyses highlight the relevance of our organoid system as a model for colorectal cancer that enables insight into clonal heterogeneity and ordering of mutational events with clinical relevance. Additionally, based on the observations in the organoids we find that a chromosome 4 deletion conditional on a chromosome 18 deletion results in a deadlier tumor. To our knowledge, this is the first example of a conditional chromosomal aberration resulting in higher mortality than the corresponding single aberrations.

2.4 Discussion

Delineating the clonal evolution trajectory through which a tumor is formed is pivotal to the understanding of tumor biology. Since every cell inside a tumor is unique, this requires an approach with single cell resolution. Here, we use single cell WGS in combination with viral lineage tracing to acquire CNV states, SNV states and viral lineage barcodes for 1641 single cells. Almost all of the CNVs identified in the organoids also frequently occur in colorectal carcinoma samples, indicating that the organoids are a valid and valuable model for colorectal carcinoma. Based on the CNVs in the single cells we could identify 52 unique CNV states in the organoids. From the 52 CNV states we derived highly detailed clonal evolution trees, which could in turn be internally validated based on the viral lineage markers and the SNVs. This internal validation is only possible due to the multiple independent lineage markers simultaneously.

The addition of the viral lineage markers revealed that certain CNVs occurred in multiple independent events in the organoid cultures. For instance, we observed at least 4 independent events in which a copy of chromosome 4 was lost. The frequent loss of chromosome 4 suggests that this event provides the organoids with a proliferative advantage. Indeed, the viral lineage barcodes showed that the clones that lost chromosome 4 expanded during the *in vitro* evolution period.

Clustering of the SNVs identified two main groups, which perfectly overlap with the two main CNV clones in the data, the chromosome 18 loss group and the chromosome 8p loss group. However, more detailed interrogation of the SNVs revealed several SNVs that can only be explained by multiple occurrences of a certain CNV. Again, this confirms the observation that chromosome 4 loss occurred multiple times during the *in vitro* evolution period.

In our data, the loss of chromosome 4 only occurred in the context of a loss of chromosome 18. This implies the order of the chromosomal aberrations is in this case important for progression. Analysis of patient data from the MSKCC colorectal cancer dataset revealed that, in patients, loss of chromosome 4 also very frequently occurs in the context of a chromosome 18 deletion. Furthermore, the combination of a chromosome 18 and a chromosome 4 deletion results in a higher mortality than a chromosome 18 deletion alone. Strikingly, a further systematic exploration of context dependent deletions or amplifications that result in a higher mortality than the initial amplification or deletion alone revealed that only the conditional deletion of chromosome 4 in the context of a deletion of chromosome 18 results in higher mortality.

In conclusion, we showed that using three independent lineage measurements acquired through single cell WGS yields highly structured clonal evolution trees of colorectal carcinoma organoids. The lineage measurements revealed that certain chromosomal aberrations occurred in multiple independent events. Finally, the most frequently occurring chromosomal aberration identified in the organoids results in higher mortality when occurring in patients with colorectal carcinoma.

Acknowledgments

This work was supported by a European Research Council Advanced grant (ERC-AdG 742225-IntScOmics), a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) TOP award (NWO-CW 714.016.001), and a Koningin Wilhelmina Fonds KWF grant (10158). J.de.R was supported by an NWO VIDI grant (639.072.715). J.D. was supported by a KWF (10218) and ERC (ERC-StG 850571) grant. This work is part of the Oncode Institute, which is partly financed by the Dutch Cancer Society. In addition, we thank the Hubrecht Sorting Facility, the Utrecht Sequencing Facility, subsidized by the University Medical Center Utrecht, Hubrecht Institute and Utrecht University and Ruben van Boxtel and Edwin Cuppen for their help and advice on the bulk WGS of the organoid line.

Author contributions

A.v.O and L.K conceived and designed the project. B.de.B and L.K. analyzed the data. A.L., V.v.d.S., D.M., and J.P. assisted with the experiments. L-T.C and A.A. assisted with data analysis. J.D. advised on organoid culturing. J.de.R. and A.v.O. supervised the work. L.K., B.de.B, and J.de.R and A.v.O wrote the manuscript. L.K. and B.de.B contributed equally to this work.

Competing financial interests

The authors declare no competing financial interests. J.de.R is founder of Cyclomics B.V.

Data availability

The accession numbers for the datasets reported in this study are available on SRA: Bio project id: PRJNA645018. The copy number tree inference and plotting code is available at `https://github.com/BuysDB/ToverBoom` the imputation generation code at `https://github.com/zztin/siCloneFitIO` and scripts at `https://github.com/BuysDB/TumorEvolutionReconstruction`.

2.5 Methods

Viral library construction

The viral construct was created using the pCDH lentivector CD811A-1 (System Bioscience) in which a GFP was inserted under control of the PGK promotor and a puromycin resistance cassette was inserted under control of the Eef1a promotor. NsiI and AscI restriction sites were inserted in the 5' UTR of the GFP gene using inverse PCR. The barcode insert was created using a 80bp primer containing the barcode (consisting of 4 stretches of 5 random nucleotides interspersed by A's) flanked by M13 forward and reverse sequences and restriction sites for Nsi1 and Asc1 and made double stranded using a complementary primer and Klenow fragment (NEB). The insert was subsequently digested with NsiI-HF and Asc1-HF (NEB), to create the right overhangs for ligation into the plasmid. Plasmid was linearized using NsiI-HF and AscI-HF and barcode insert was ligated using T4 DNA Ligase (NEB). Ligated plasmid was transformed into Stable Competent E. Coli cells (C3040 NEB) and 30.000 colonies were harvested from which plasmids were extracted.

Viral library complexity assessment

8 replicates of 1 ug of viral library were amplified using NEBNext High fidelity PCR mix (NEB) for 10 cycles with barcoded PCR forward primer 1 and PCR reverse primer 1 (**Table S1**). Illumina sequencing libraries were generated through 5 additional cycles of PCR with Illumina Truseq small RNA library PCR primers. The viral library was sequenced on a NextSeq500 using 2x75bp paired end sequencing. Barcode sequences were merged if they were within

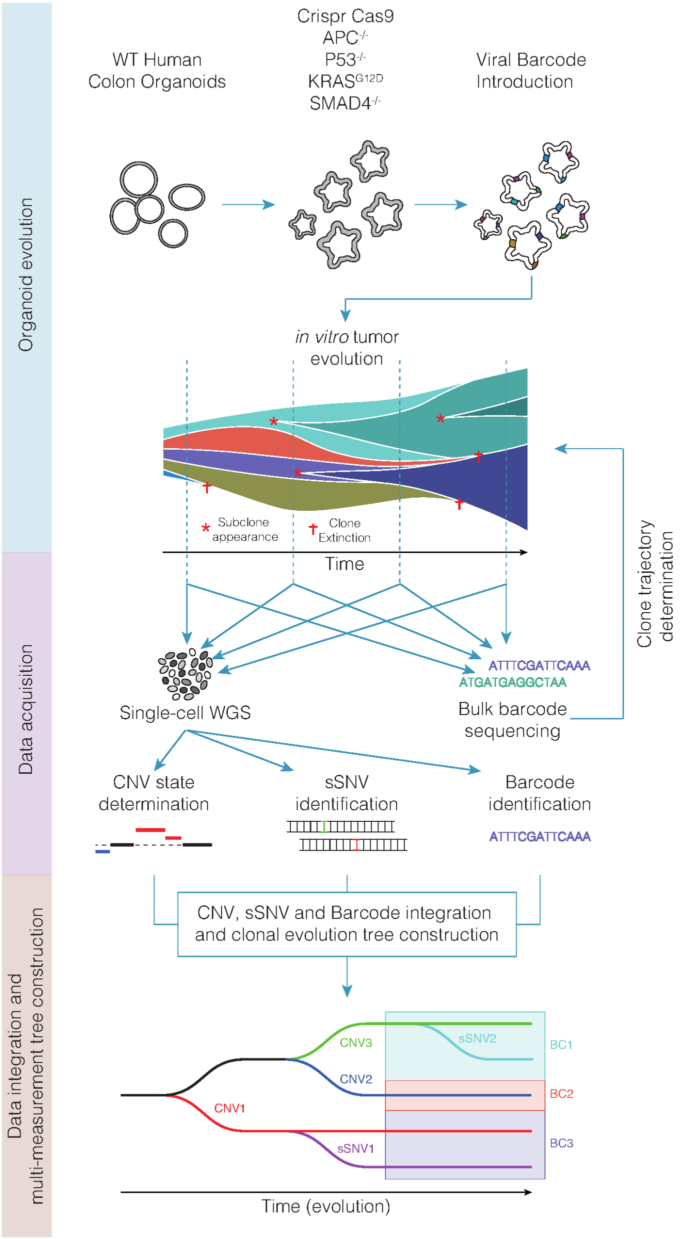


FIGURE 1. Experimental setup. Wild-type human colorectal organoids were transformed using a Crispr-Cas9 based strategy [60]. Transformed organoids were transduced with a lentiviral library introducing a lineage barcode. organoids underwent a 25 week *in vitro* evolution period during which single cell WGS was performed at regular intervals and culture complexity was assessed weekly. From the single cell WGS, copy number state, sSNV state and lineage barcodes were acquired, allowing the construction and validation of highly detailed clonal evolution trees.

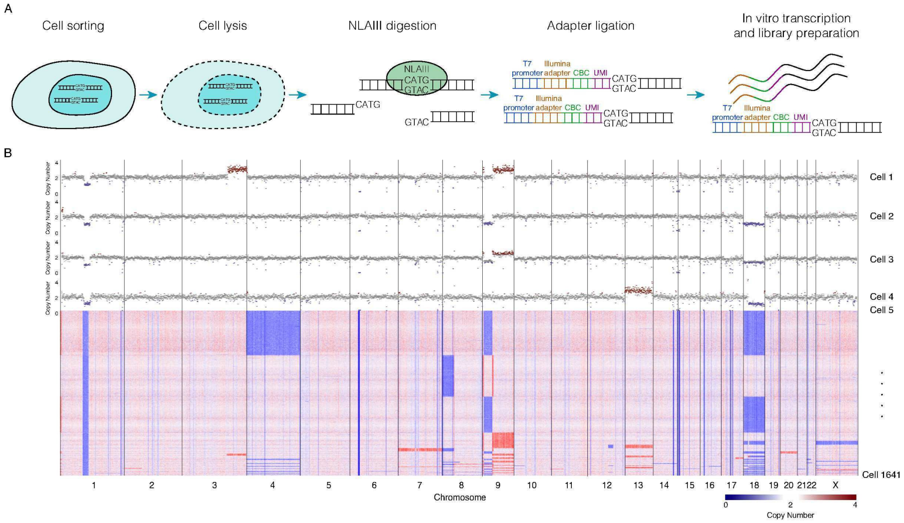


FIGURE 2. CNV state landscape during clonal evolution. (A) schematic overview of single cell WGS strategy. (B) CNV profile of the 1641 single cells from which single cell WGS data was acquired.

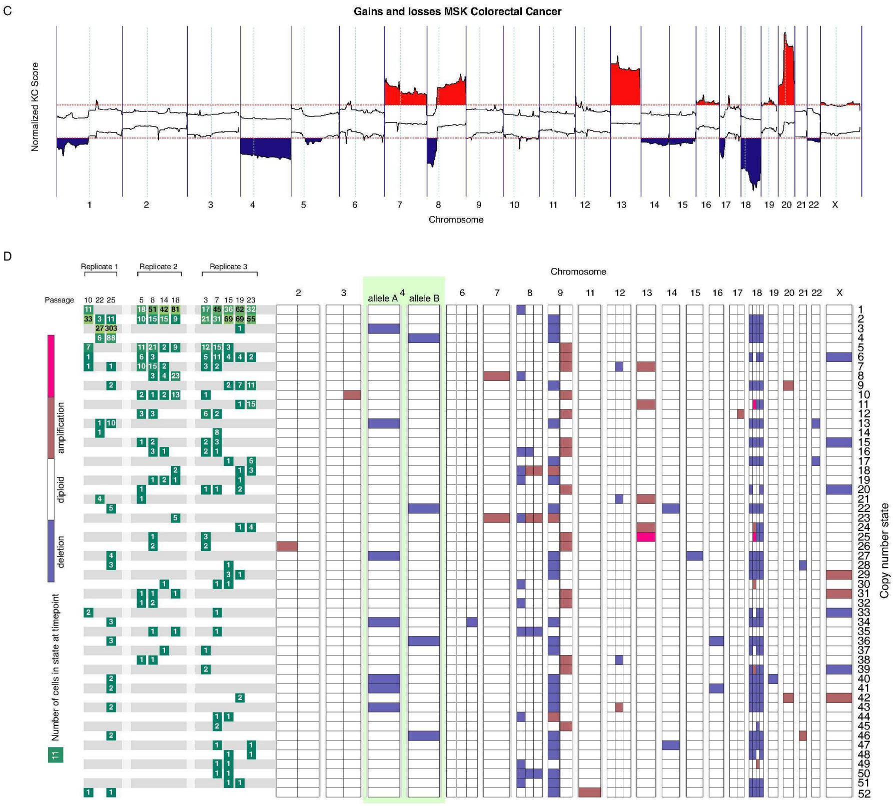


FIGURE 2. CNV state landscape during clonal evolution. (C) Frequently occurring chromosomal aberrations in the Memorial Sloan Kettering Colorectal Cancer dataset. (D) Overview of the 52 unique CNV states identified in the organoids. Each line represents a CNV state, the left part of the figure shows the abundance of the CNV states in the 3 replicates at the different time points. The right part of the figure shows the deletions and amplification that were detected in each of the CNV states. Chromosome 4 has been split into allele A and allele B.

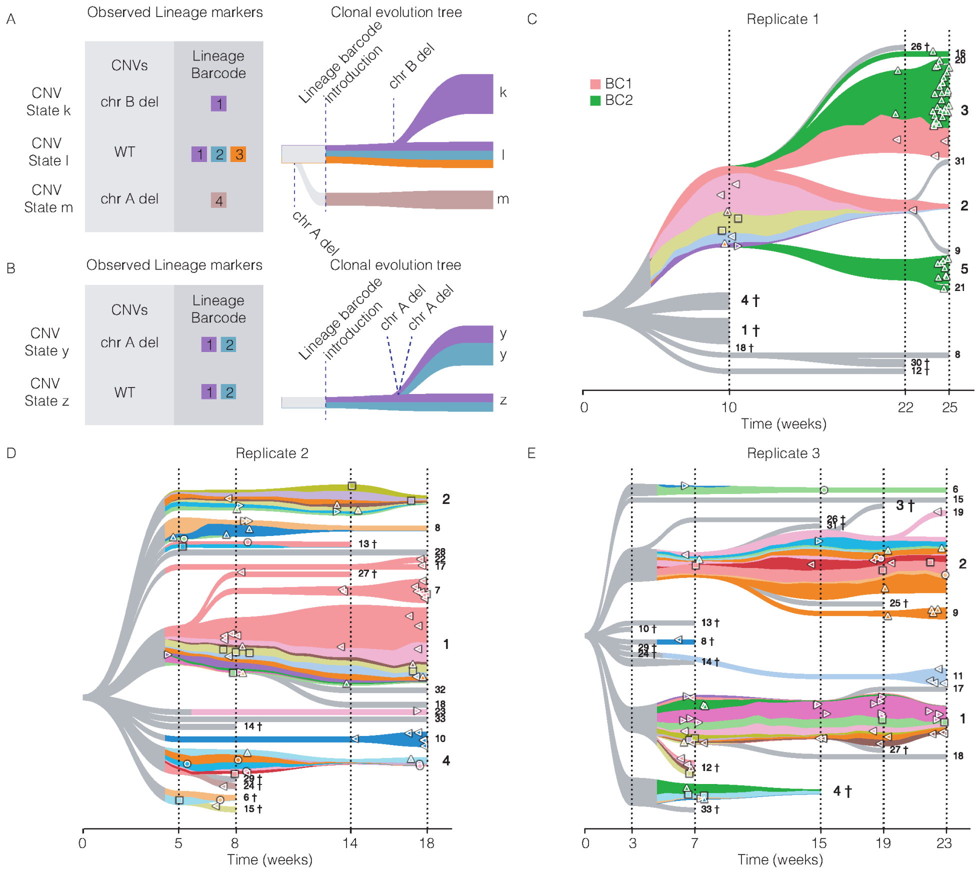


FIGURE 3. Viral lineage barcodes projected on CNV based clonal evolution trees. (A-B) Examples of CNV events that can be explained by the combination of CNV state and viral lineage barcode information. (C-E) Viral lineage barcodes projected on CNV based clonal evolution trees for replicates 1 through 3. Numbers on the right side of each tree indicate the CNV state. + indicates CNV state has gone extinct.

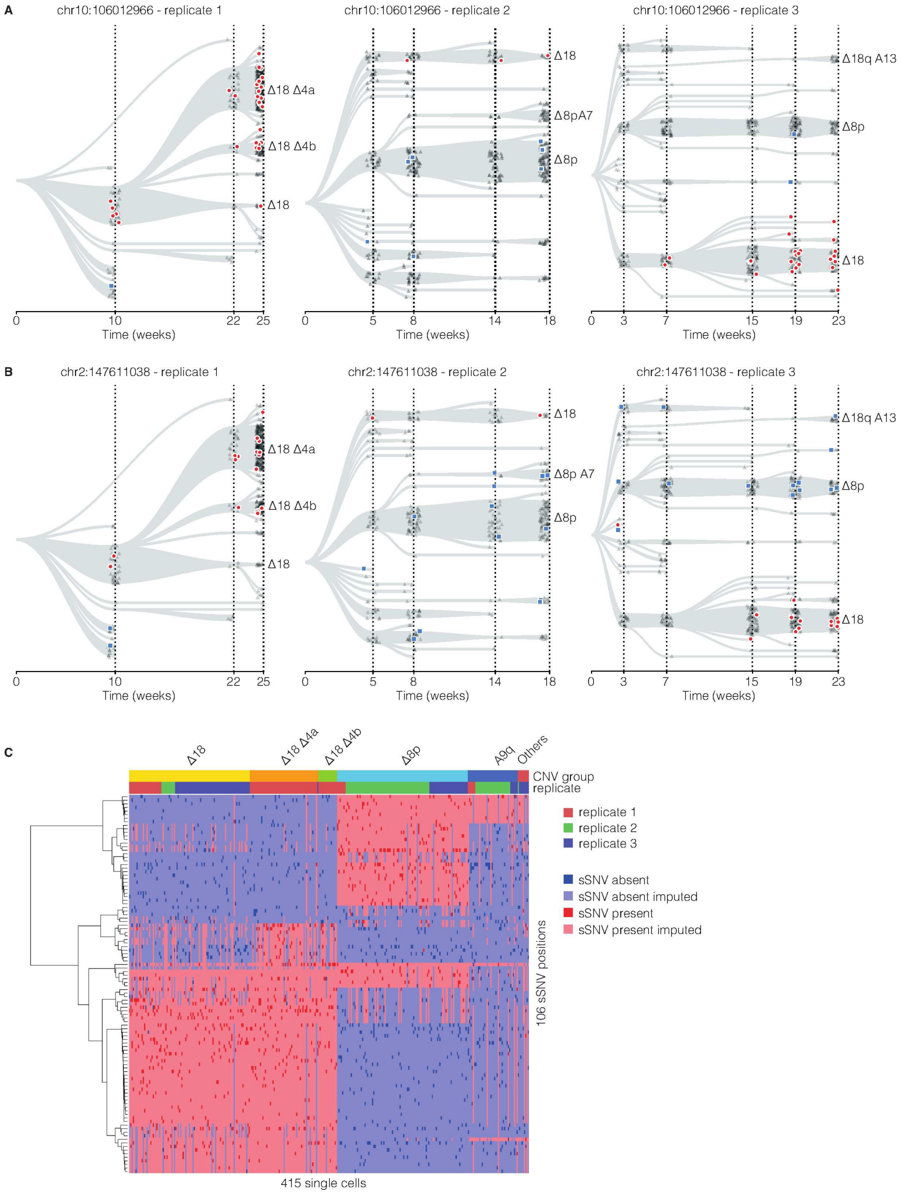


FIGURE 4. sSNVs projected on CNV based clonal evolution trees. (A-B) The presence of one sSNV projected onto CNV trees of 3 replicates. Each marker indicates a single cell and its color and shape shows whether the alternative allele is detected (red, circle) or the reference allele (blue, square) is detected or the sSNV is not covered (grey, triangle). (C) Single cell sSNV matrix, cells are grouped based on their copy number state.

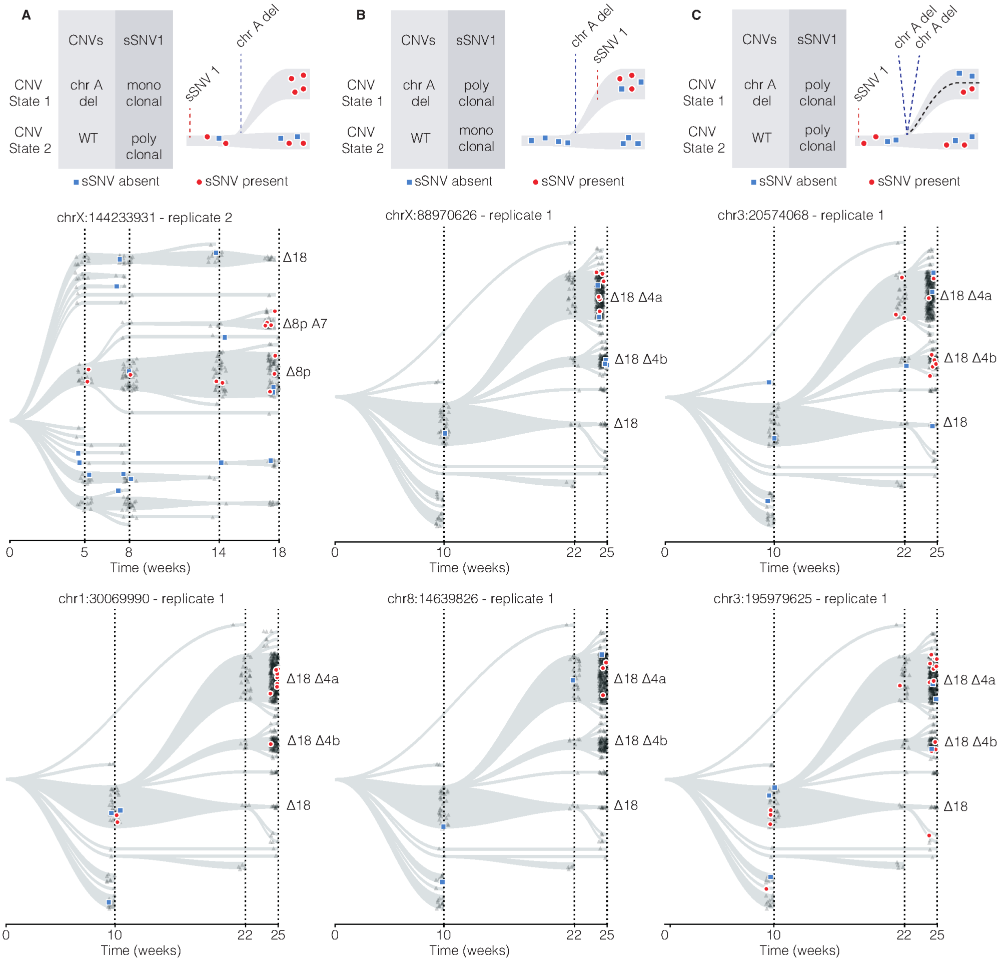


FIGURE 5. Three classes of sSNVs at the branching point of two copy number states. Schematic representations of class 1-3 on the top row. The two rows below show identified instances of each class. Each marker indicates a single cell. Its color and shape shows whether the alternative allele (red, circle) or the reference allele (blue, square) is detected or the sSNV is not covered (grey, triangle). (A) In class 1 the copy number aberration occurs after the sSNV appearance. (B) In class 2 the copy number aberration occurs before the sSNV appearance. (C) Class 3 is an example of a parallel evolution event, where a sSNV is followed by two independent copy number aberrations.

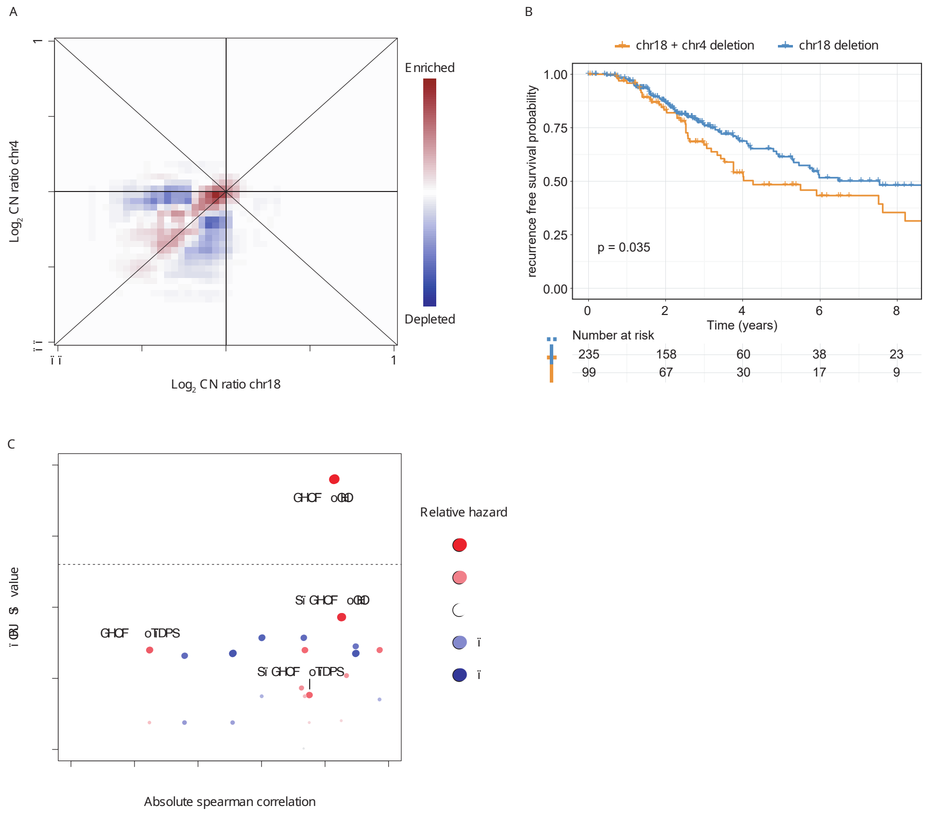


FIGURE 6. Loss of chromosome 18 followed by loss of chromosome 4 worsens survival probability. (A): Relation between chromosome 18 loss and chromosome 4 loss within the MSKCC patient cohort. Red and blue shading indicate enrichment and depletion compared to a random background distribution. (B) Kaplan-meier curve comparing survival between patient with a chromosome 18 and a chromosome 4 loss compared to patient with a chromosome 18 loss without a chromosome 4 loss. (C) Absolute correlation between chromosomal alterations compared to the p-value for the difference in hazard for the priming event alone compared to the priming event and the conditional event. Shading and size of the points indicate direction and magnitude of the difference in hazard.

hamming distance 2 from each other, merging them into the most abundant of the two, while taking into account sequencing quality.

DNA extraction, barcode amplification and barcode sequencing

For each organoid line DNA was harvested weekly. Cell lysis was performed overnight at 50C using 0.05 units of Qiagen Protease in 10 mM tris pH 7 in a total volume of 1 ml. All samples were split into two and DNA was extracted using phenol/chloroform extraction followed by AMPure DNA bead clean-up (Beckman). Viral barcodes were amplified using a two-step PCR strategy [102]. All PCRs were done in 96 well plates, using the 96 barcoded forward primers (1 primer per well, **Table S1**) in combination with a mix of 5 reverse primers. The 5 reverse primers are identical except for a small (0, 1, 2, 3 or 4 base insertion), which ensures high complexity of the libraries, required for sequencing. First, for both replicates of a sample 5 cycles of PCR were performed on 500ng of genomic DNA using NEBnext High Fidelity PCR master mix (NEB) and barcoded primers containing a Unique Molecule Identifier (UMI) (**Table S1**). After PCR, excess primers were digested using ExoSap (Agilent) to prevent UMI replacement during later stages of amplification. After ExoSap treatment PCR reactions were cleaned up using AMPure beads and another 25 cycle PCR was performed using Illumina Truseq small RNA library PCR primers. Libraries were sequenced on Illumina NextSeq 500 using 2x75 bp paired end sequencing. DNA reads were mapped to an artificial reference genome containing 30.190 viral genomes, each with their own unique barcode. Only reads that mapped uniquely to a single viral barcode were considered for further analysis. Library PCR duplicates (based on UMI sharing) were removed. To estimate barcode frequency for each individual time point we used the approach described in [102], which uses a Bayesian model to infer the frequency of the barcode in the original culture through the number of reads sequenced in the two replicates from that timepoint.

Whole genome sequencing and bulk variant calling

At passage 4 and passage 21 WGS was performed on the APKS organoids. At passage 4 a mix of DNA from the three replicates was used, while at passage 21 each replicate was sequenced individually. For this DNA was isolated from cells that were left over after passaging the culture. Library preparation and whole genome sequencing was performed at Macrogen using Illumina TruSeq DNA PCR free library preparation and sequenced on a HiSeq 10X with 2 x 150 bp paired end sequencing. Reads were aligned to GRCh38 using Burrows Wheeler Aligner v0.7.14 mapping tool with settings 'bwa mem –M' [105]. Duplicate reads were marked using Sambamba (version 0.6.6) dedup. Base Quality Score Recalibration was done using GATKBaseRecalibrator v3.7 [11]. Somatic variants were detected using Mutect 2.2 [44].

Single Cell Whole genome Sequencing

Cells were sorted into 384-well plates with 5 ul of mineral oil (Sigma-Aldrich). After sorting, cells can be stored at -20C. 500 nl of lysis mix (0.0005 u Qiagen Protease in NEB Buffer 4) was added to each well and lysis was performed at 55C overnight followed by heat inactivation for 20 minutes at 75C and for 5 minutes at 80C. 500nl of Restriction Enzyme mix (0.5 u NlaIII in NEB Cutsmart buffer) was added to each well and restriction was performed for 3 hours at 37C followed by heat inactivation for 20 minutes at 65C. 100 nl of 1 uM barcoded double stranded NlaIII adapter was added to each well. 1100 ul of Ligation mix (200 u T4 DNA Ligase in 1x T4 DNA Ligase buffer supplemented with 3 mM ATP) was added to each well and ligation was performed overnight at 16C. After ligation, single cells were pooled and library preparation was performed as described in Muraro et al. [120]. Libraries were sequenced on an Illumina Nextseq500 with 2 x 75 bp paired end sequencing or on a HiSeq 10X with 2 x 150 bp paired end sequencing.

Single cell whole genome data processing

Sequencing data were analyzed through custom *snakemake* workflows (*Python* v3.6), which are available at
`https://github.com/BuysDB/SingleCellMultiOmics/tree/master/singlecellmultiomics/snakemake_workflows/nlaIII`

The UMI and cell barcode were extracted and trimmed from read 1 of the read pair and the 6bp random hexamer was trimmed from read 2. From the resulting trimmed reads, only those starting with the NlaIII recognition sequence CATG were kept. Additionally, adapters were trimmed using cutadapt [112]. The trimmed reads were mapped to hg38 using BWA 0.7.16a-r118. Next, the mapping location and strand of the NlaIII recognition sequence in combination with the UMI sequence and cell barcode was used as a unique molecular identifier. This step associates reads to unique molecules in order to deduplicate reads to reduce amplification biases and is used to extract a consensus sequence for each molecule. The consensus base calls are used to genotype germline and somatic SNVs

In order to remove non-uniquely mapping reads, the reference genome was digested in-silico using the NlaIII cut site. For each NlaIII cut site the two flanking fragments were determined for sequences up to 69 bases in length. These fragments were mapped back to the hg38 reference. For each site multi-mapping fragments were recorded. Only molecules mapping to uniquely mappable sites according to the in-silico digestion were kept for copy number analysis. For each cell, molecules were binned in 500kb bins. Bins with fewer than 3000 unique cut sites are considered to have poor mappability and were excluded from the analysis. Due to unavailability of wild-type WGS single cell libraries the copy number profiles could not be normalized against a reference profile. Instead count data was median normalized for each cell and multiplied by 2, resulting in a median copy number of 2 for every cell. Next, we carried out GC bias correction by performing a LOESS regression for the copy number profile of each cell. The corrected values were clipped to a maximum copy number of 4, to mitigate inflated noise at high copy numbers. We find that, even after the rigorous data processing described above, we do not obtain a reliable copy number profile for all cells, these profiles might be caused by cell division or a cell lysis-induced artefact. To filter cells with an unreliable copy number profile we trained a random forest classifier. Training labels were obtained by k-means clustering (k=12) the cells in UMAP 2D space and manually identifying the cluster which predominantly contains cells with unreliable copy number profiles. The final classifier was applied on the total matrix and all cells with a posterior >0.99 for the noisy cluster were discarded. The out-of-bag classification score of the random forest was 0.985.

Copy number segmentation and state definition

Before copy number segmentation, cells were clustered using Ward's hierarchical clustering on the Euclidean distance. The number of clusters were set based on the maximum silhouette score, but to ensure conservative (tight) clusters, overclustering was manually enforced for certain large clusters. For each resulting cluster of cells, the mean copy number was calculated per bin and copy number segments were detected using circular binary segmentation [127] with p=0.05 and 10,000 shuffles. Segment calls with a mean absolute difference of smaller than 0.6 were rejected. For each cell, the median for each segment was calculated and rounded to the nearest integer.

Segments with variance higher than 0.025 across all cells, which in practice turned out to be small genomic segments with hard to resolve copy numbers, were rejected to prevent those small segments from majorly influencing lineage tree inference.

To obtain diplotypes for both chromosome 4 and 18, data from a bulk sample (AP1-P23) derived from the P11N line was leveraged, which contains a complete and clonal loss of both chromosome 4 and 18. For each heterozygous gSNV, the allele with a BAF of 1 is assigned to allele B and the allele with a BAF of 0 to allele A. The A and B allele-frequencies were determined per cell for each segment on chromosome 4 and chromosome 18. Per segment the allele specific copy number was estimated by multiplying the estimated total copy number by the A and B allele frequency.

To define the copy number states, a second round of clustering of the cells was performed based on the integer copy number segmentation. Cells with hamming distance of zero were grouped to form the copy number states. Copy number states were sorted by the number of cells associated with the state, which ranges from 395 cells in copy number state 1 and 2 cells in copy number state 52. Copy number states with fewer than 2 cells were discarded. The segmented copy number calls along with the diplotype specific segmented copy number calls for chromosome 4 and 18 for each cell individually gives rise to the copy number state matrix.

Copy number tree inference

To extract a copy number tree we first infer a directed graph from the single cell copy number state matrix. Every node in the graph represents a single copy number state at one point in time. To incorporate a time axis in the graph, every copy number state is represented by one node for every time point a copy number state has been measured. When a copy number state of a particular clone is missing we interpolate its abundance using linear interpolation. In this directed graph, every edge represents a copy number change and the weight of the edge represents the amount of edits between two nodes. Edges are pruned if they are biologically not plausible, e.g. in case they connect nodes with zero copies to a higher copy number or if they connect nodes in opposite temporal direction. A zero-weight edge is added between temporally adjacent nodes representing the same copy number state. An artificial root node is added to the graph wherein all segments are set to a diploid copy number state. From the resulting graph, an arborescence is extracted by using Edmonds algorithm, resulting in a copy number tree.

Copy number tree plotting

Copy number trees are plotted using a novel visualization, developed for this purpose, called ToverBoom. In Toverboom, each node represents a copy number state and the width of each node represents the relative amount of cells in the copy number state. Each branch represents a transition to another copy number state. The width of the nodes is smoothed using cubic interpolation. Lineage barcodes for each single cell were extracted from the single cell whole genome sequencing libraries. The lineage abundance was extracted from the bulk barcode sequencing libraries. Within one copy number state the relative abundance of every associated lineage barcode is calculated and projected on the lineage tree using a stacked area chart, where the area reflects the relative abundance of the lineage barcode within the associated copy number clone. The copy number tree inference and plotting code is available at `https://github.com/BuysDB/ToverBoom`.

Somatic single nucleotide variants detection

Variants were jointly called on 7841 cells derived from the three quadruple mutant replicates, and two other replicates (one single mutant and one double mutant which both serve as a normal control). All cells are descendants from the same donor.

Basecalling phred scores of the single cell bam files were recalibrated using GATK base quality score recalibration. All variants detected using the GATK HaplotypeCaller in the (Wildtype/P11N) bulk library and variants detected by Mutect2 in any of the bulk samples were supplied as known variation to be masked during covariate analysis. Candidate sSNVs were jointly called using BCFtools 1.9-174 [103] on a bam file containing all cells, and a threshold was set on the QUAL column for a phred score of at least 30.

To remove technical artifacts and germline variation only sSNVs uniquely detected in the quadruple mutant cells were kept, while sSNVs detected in single cells from the normal control samples were dropped. Furthermore, sSNVs detectable in the (Wildtype/P11N) bulk library with more than one read were dropped.

Haplotype phasing was performed using a strategy adapted from Bohrson et al.[30]. Briefly, for each sSNV phased heterozygous single nucleotide germline variants (gSNV) were determined in the (Wildtype/P11N) library. For the sSNVs with at least one phased gSNV, it was determined if phasing between the heterozygous gSNV and the sSNV is concordant in at least 95% of all cells, otherwise the sSNV was discarded. Molecules containing the sSNV are used as evidence indicative of presence of the sSNV. Absence of the sSNV is inferred when a cell has a molecule containing both the phased gSNV allele and the reference allele at the sSNV locus.

The sSNVs were further filtered by a random forest classifier trained on the 198150

sSNVs detected in bulk using Mutect2 as the ground truth. The features consisted of all the columns generated by the GATK variant caller, of which ReadPosRankSum and BaseQRankSum were most informative for classification. These features were appended with the following: the number of reads carrying the alternative base, the mean base quality of the alternative base in the single cell data, the mean number of gSNVs overlapping with reads containing the alternative allele and the mean number of gSNVs overlapping with reads containing the reference allele and the complexity of the reference sequence in a 75bp, 150bp, 300bp, 500bp and 1kb window, encoded by counting the number of unique 5bp and 7bp k-mers. Final classification of the candidate variants was performed using leave one out cross-validation. The classifier used is a sklearn random forest classifier with 100 trees and class balancing weights enabled. Finally, all selected variants were inspected in a genome browser (IGV). A few variants were removed upon manual inspection.

Somatic single nucleotide variant imputation

Genotypes of all quadruple mutant single cells to which a copy number state could be assigned are inferred and imputed using a Bayesian inference algorithm, SiCloneFit [193]. The imputation allows for clustering of the single cell SNV genotypes. Only variants which were present in at least 2 cells and only cells with at least 4 sSNVs were used for the imputation. Expected false negative and false positive rates of the sSNV measurements are set at 0.001 and 0.0001, respectively. SiCLoneFit utilizes Gibbs sampling of the posterior distribution of measured SNVs to infer tumor clones, phylogeny, and genotype in each tumor clone. The missing sSNV measurements are imputed according to their genotype in the assigned clone. The imputed sSNVs were then combined with the measured sSNVs and clustered and plotted (Fig. 4B). The imputation and visualization tools are available at `https://github.com/zztin/siCloneFitIO`. To test the accuracy of the imputation we performed 10 fold cross-validation by leaving out a fold of 10% of known sSNV calls. The estimated accuracy is approximately 0.86.

Neutral drift simulations

To investigate neutral drift, we performed in silico stochastic simulations. To this end, three parameters were defined: the replication rate (rr) (number of cell divisions per hour) of the organoids, the number of cells starting the population (sp) and the number of cells that were retained when passaging the culture (bottle neck size (bns)). The simulation was then executed as follows. Every cell in the starting population is considered a unique clone, every hour every cell belonging to a certain clone has a certain probability to proliferate (rr). When a cell proliferates the number of cells belonging to that clone increases by 1. After 168 hours (1 week) bns cell are randomly selected and allowed to restart the culture. This process then continues for 25 weeks. Finally, the Shannon's entropy of the clones in the culture was analyzed to estimate the clonal dynamics.

Conditional chromosomal aberration analysis

Contigs (entire chromosomes, except chromosome 8 which was split into 8p and 8q) were filtered on having an average absolute log2 copy ratio of > 1.5 in the MSKCC data set. For these contigs the absolute correlations of the log2 copy ratio for all combinations were calculated. For each combination of contigs a Cox regression model was created in which the hazard ratio for tumors harboring both the priming and the conditional event was compared to tumors harboring only the priming event. P-values were corrected using Benjamini Hochberg p-value correction.

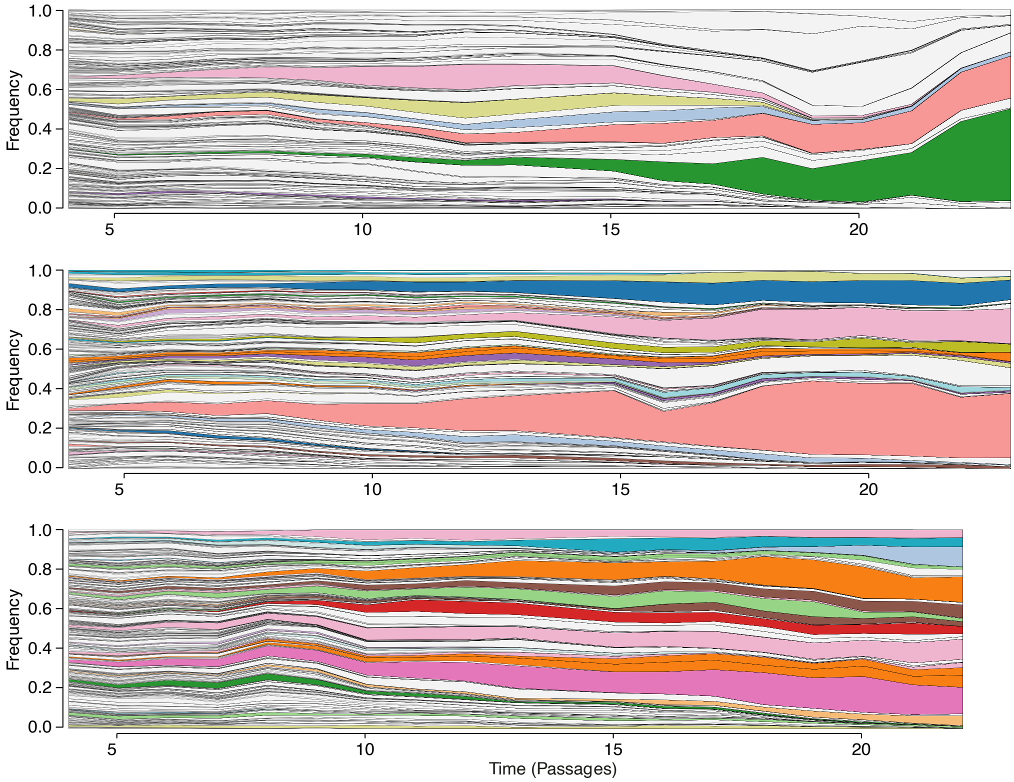


FIGURE S1. Clonal dynamics during *in vitro* evolution. (A-C) Relative frequency of observed viral lineage barcodes in replicates 1 through 3.

2.6 Supplementary figures

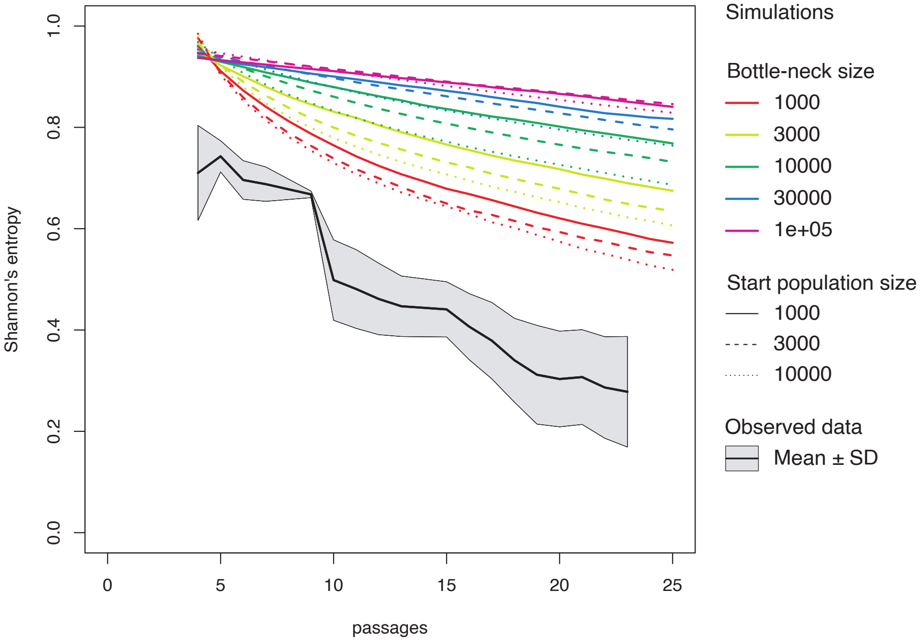


FIGURE S2. Simulation of clonal dynamics. Simulation of the Shannon entropy of the relative viral lineage barcode frequency as a function of time given a certain culture complexity at viral lineage introduction and a certain bottle neck size during the weekly passaging of the organoids (colored lines). Simulations assume there is no selection pressure on the culture and all cells have equal proliferative capacity. We calculated the Shannon's entropy of the clones as a measure for culture complexity. The replication rate was varied between 1/36, 1/48, 1/72 and 1/96 cell divisions per hour but there were no differences in entropy between the different proliferation rates, indicating that proliferation rate does not influence the clonal dynamics if it is assumed that all cells have the same proliferative potential. The entropy for the different simulations showed that the culture complexity is primarily depending on the bottle neck size, where a smaller bottle neck size shows a faster decrease in culture complexity, and to a lesser extend on the starting population size, where smaller starting sizes show a faster decrease in complexity.

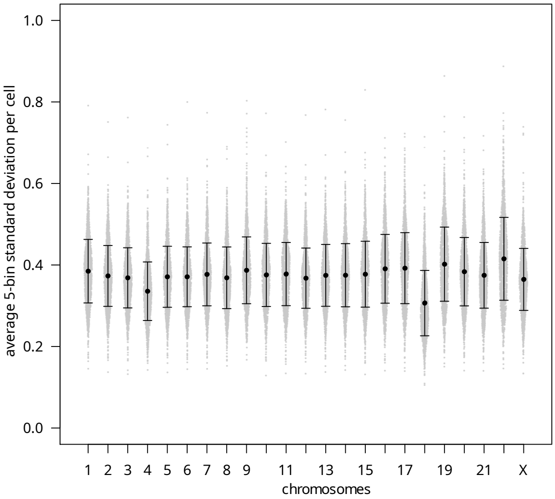


Figure S3. Variation in copy number measurement. Standard deviation of single cell CNV profiles, every dot indicates the average 5-bin standard deviation per chromosome for all single cells.

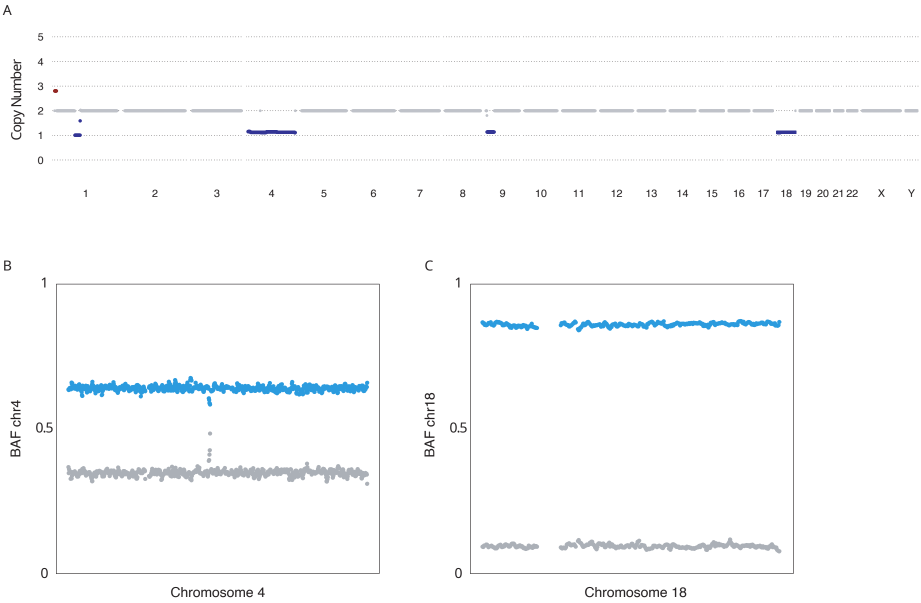


Figure S4. Bulk copy number profile (A): Bulk copy number profiles for replicate 1-3 at 23 passages. (B-C): B-allele frequencies in bins across chromosome 4 (B) and 18 (C), each bin contains 500 SNPs present in the Hapmap SNP project.

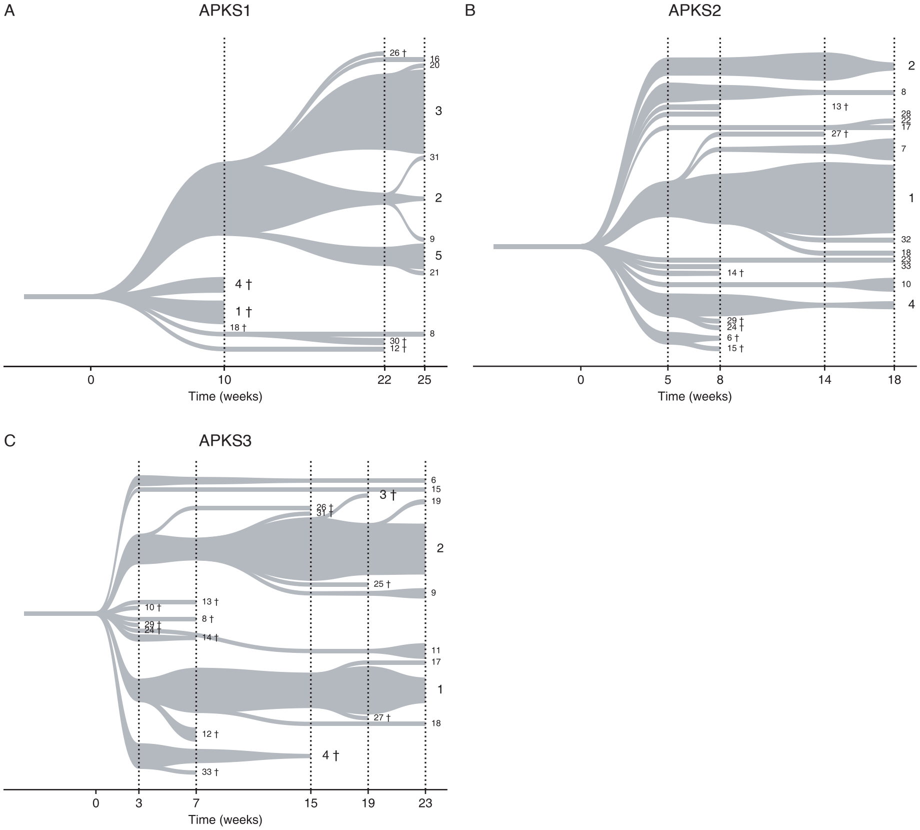


Figure S5. Clonal evolution trees build from CNV states.

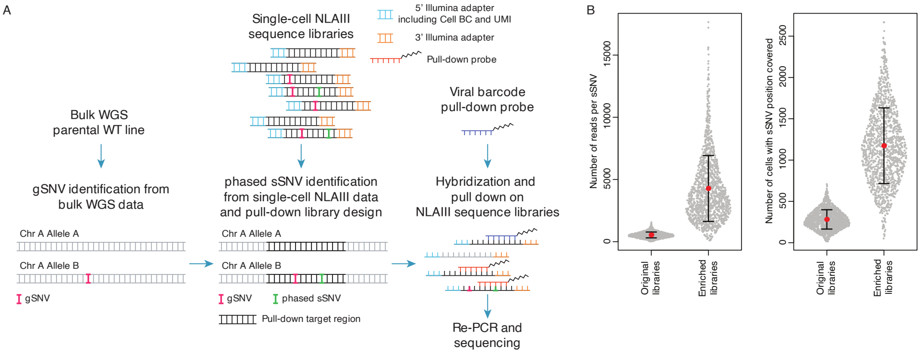

A
Single-cell NLAIII sequence libraries
5' Illumina adapter including Cell BC and UMI
3' Illumina adapter
Pull-down probe
Bulk WGS parental WT line
Viral barcode pull-down probe
gSNV identification from bulk WGS data
phased sSNV identification from single-cell NLAIII data and pull-down library design
Hybridization and pull down on NLAIII sequence libraries
Chr A Allele A
Chr A Allele B
gSNV
phased sSNV
Pull-down target region
Re-PCR and sequencing
B
Number of reads per sSNV
0
5000
10000
15000
Original libraries
Enriched libraries
Number of cells with sSNV position covered
0
500
1000
1500
2000
2500
Original libraries
Enriched libraries

Figure S6. Enrichment of sSNVs positions from the single cell libraries. (A) Enrichment strategy for the sSNV positions. In brief, candidate sSNV positions were identified from a first round of sequencing of the single cell libraries. Anti-sense oligo's to the candidate positions were designed and hybridized to the single cell DNA libraries. Oligo's were pulled down from the mix and enriched libraries were sequenced. (B): Enrichment of candidate sSNV positions after pull-down enrichment.

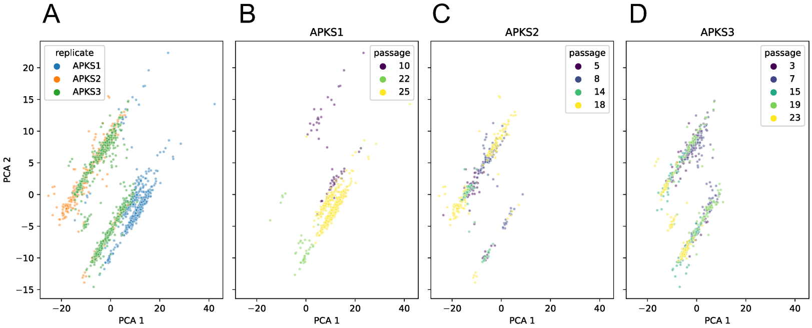
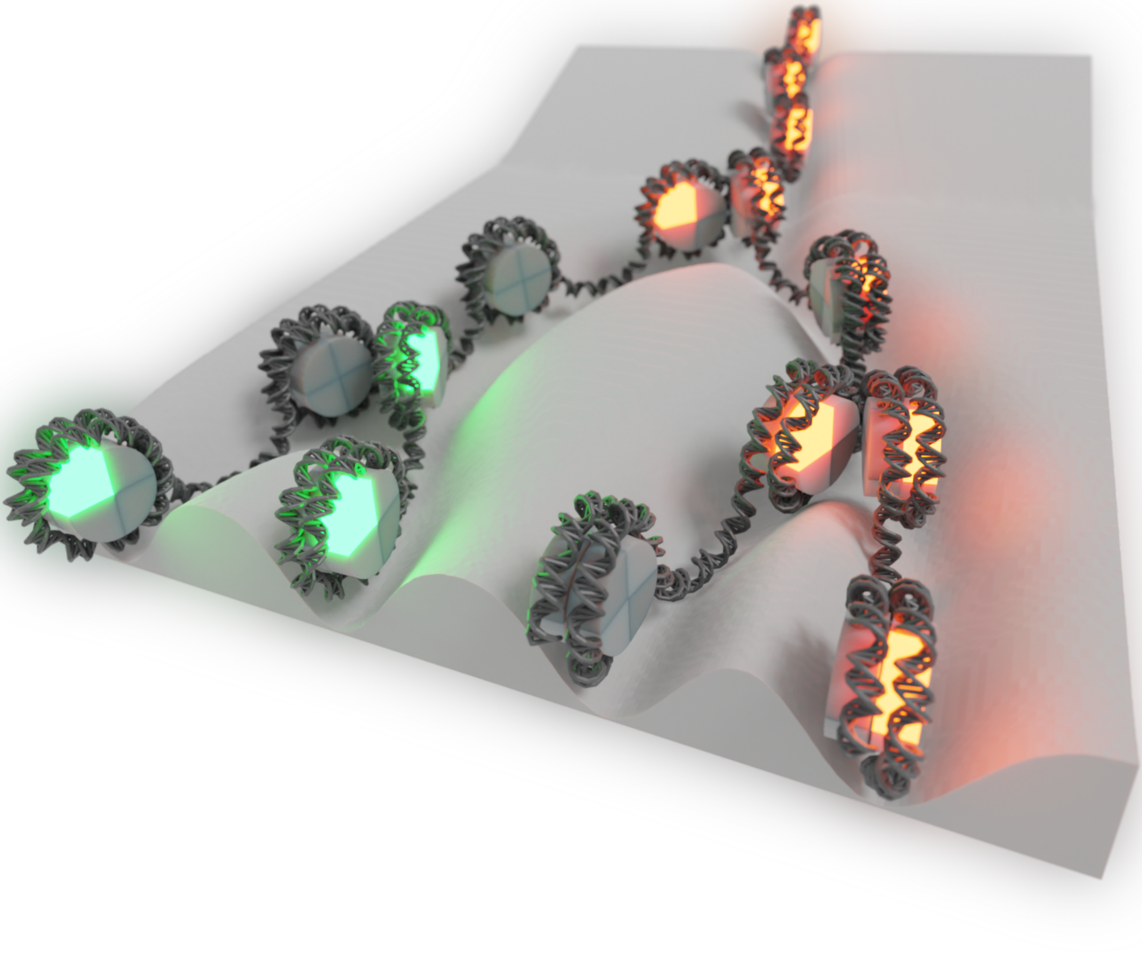


Figure S7. Quality control plots of the single cell copy number data by principal component analysis on the median normalized count matrix with allele specific counts on chromosome 4 and 18. Each dot corresponds to a single cell. (A) Cells are colored based on the replicate the cells belong to. (B-D) Cells of a single replicate are shown and the brightness of each dot indicates the passage (time-point).



Chapter 3

Hierarchical chromatin regulation during blood formation uncovered by single-cell sortChIC

Under review

Peter Zeller*, Jake Yeung*, Buys Anton de Barbanson, Helena Viñas Gaza, Maria Florescu, and Alexander van Oudenaarden

Oncode Institute, Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) and University Medical Center Utrecht, 3584 CT, Utrecht, The Netherlands

*These authors contributed equally Correspondence: a.vanoudenaarden@hubrecht.eu (A.v.O.)

3.1 Summary

Post-translational histone modifications modulate chromatin packing to regulate gene expression. How chromatin states, at euchromatic and heterochromatic regions, underlie cell fate decisions in single cells is relatively unexplored. We develop sort assisted single-cell chromatin immunocleavage (sortChIC) and map active ($H_3K_4me_1$ and $H_3K_4me_3$) and repressive ($H_3K_{27}me_3$ and $H_3K_9me_3$) histone modifications in HSPCs, and mature blood cells in the mouse bone marrow. During differentiation, HSPCs acquire distinct active chromatin states that depend on the specific cell fate, mediated by cell type-specifying transcription factors. By contrast, most regions that gain or lose repressive marks during differentiation do so independent of cell fate. Joint profiling of $H_3K_4me_1$ and $H_3K_9me_3$ demonstrates that cell types within the myeloid lineage have distinct active chromatin but share similar myeloid-specific het-

erochromatin repressed states. This suggests hierarchical chromatin regulation during hematopoiesis: heterochromatin dynamics define differentiation trajectories and lineages, while euchromatin dynamics establish cell types within lineages.

3.2 Introduction

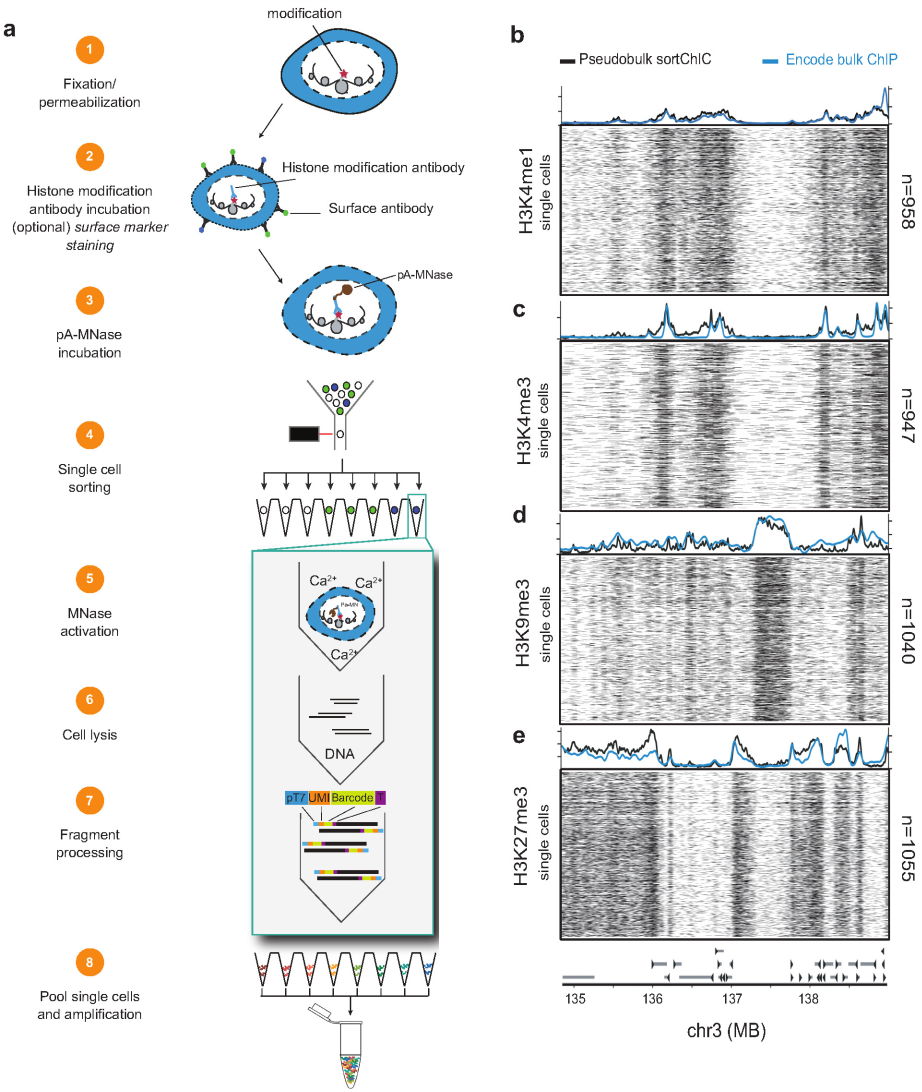
Hematopoietic stem cells (HSCs) reside in the *Bone Marrow* (BM) to replenish blood cells while maintaining a balance of diverse blood cell types [129, 165]. During differentiation, HSCs progressively restrict their potential to fewer lineages to yield mature blood cells [128]. These cell fate decisions are accompanied by gene expression dynamics, which have recently been dissected through single-cell *Messenger Ribonucleic Acid* (mRNA) sequencing technologies [72, 12, 130].

The regulation of gene expression relies, in part, on post-translational modifications of histones that modulate the packing of chromatin [5, 22]. Chromatin dynamics during hematopoiesis have so far been analyzed in detail for open chromatin regions in single cells [34, 136] and active chromatin marks in sorted blood cell types [101]. Although the role of repressive chromatin has been characterized in embryonic stem cell cultures [118, 185, 32, 132] and during early development [122, 64, 67], the dynamics of repressive chromatin states during hematopoiesis has been relatively unexplored.

Two repressive chromatin states play a major role in gene regulation: a polycomb-repressed state, marked by $H_3K_{27}me_3$ at gene-rich, GC-rich regions [21, 131], and a condensed heterochromatin state mainly found in gene poor AT-rich regions, marked by $H_3K_9me_3$ [122]. Conventional techniques to detect these histone modifications involve chromatin immunoprecipitation (ChIP), which relies on physical pull-down of histone-DNA complexes. This pull-down can hinder sensitive detection in single cells, although microfluidic or combinatorial barcoding extensions of ChIP-seq has improved the assay to single-cell resolution [141, 73, 3]. Alternatives to ChIP [151] circumvent this pulldown by using antibody tethering of either *protein-A-micrococcal nuclease* (pA-MNase) or protein-A-Tn5 transposase [89, 77], improving signal-to-noise by cutting only at specific sites of the genome. Although these tethering-based strategies have enabled large-scale profiling of histone modifications in single cells [187, 16, 86], they generally do not enrich for different cell types, making it difficult to dissect chromatin regulation in rare cell types, such as hematopoietic stem and progenitor cells in the bone marrow.

Here we develop sortChIC, which combines single-cell histone modification profiling with cell enrichment, and apply it to map histone modifications in hematopoietic stem cells/early progenitors (HSPCs) and mature blood cell types in the mouse bone marrow. We characterize active (H3K4me1 and H3K4me3) and repressive chromatin (H3K27me3 and H3K9me3). Active chromatin in HSPCs primes for different blood cell fates, while H3K27me3 repressive chromatin in mature cell types silences genes of alternative fates. Although H3K27me3 and H3K9me3 repressive modifications target distinct regions of the genome, most regions that gain or lose repressive modifications during differentiation do so independent of the specific cell fate. By contrast, active chromatin shows divergent changes during hematopoiesis, where gains and losses at genomic regions depend on the specific cell fate. Transcription factor (TF) motif analysis predicts cell type-specifying TFs that drive different chromatin dynamics to acquire distinct chromatin states. Simultaneous targeting of H3K4me1 and H3K9me3 reveals that cell types within the myeloid lineage have distinct active chromatin states, while sharing similar lineage-specific heterochromatin. Our resource reveals a single-cell view of chromatin state dynamics during hematopoiesis in both euchromatic and heterochromatic regions. We propose a hierarchical differentiation program of chromatin regulation in hematopoiesis, by which heterochromatin states define a differentiation trajectory and lineages, while euchromatin states establish cell types.

Figure 1



(Caption on next page.)

Fig. 1: sortChIC maps histone modifications in single cell. (a) Schematic of the sortChIC method. Fixed and permeabilized cells are stained with an antibody targeting a histone modification. Inactive protein A-micrococcal nuclease (pA-MNase) is added, tethering MNase to the histone modification antibody. Single cells are FACS sorted. MNase is activated to induce specific cuts in the genome. Unique molecular identifiers (UMI) and cell-specific barcodes are ligated to the cut fragments. Barcoded fragments are pooled, amplified, and sequenced. (b-e) Location of cuts in $H_3K_4me_1$ (b), $H_3K_4me_3$ (c), $H_3K_9me_3$ (d), and $H_3K_{27}me_3$ (e) in individual K562 cells along a 4 MB region of chromosome 3. Black traces represent the sortChIC signal averaged over all individual cells, blue traces represent ENCODE ChIP-seq profiles.

3.3 Results

SortChIC maps histone modifications in single cells.

To detect histone modifications in single cells, we first fix cells in ethanol to preserve surface antigens for FACS sorting and incubate with an antibody against a particular histone modification. We then add protein A-MNase (pA-MN) which binds to the antibody at specific regions of the genome (Fig. 1a). During this incubation, MNase is kept inactive (i.e., no Ca^{2+}). After washing away unbound antibody, single cells in the G1 phase of the cell cycle are sorted into 384-well plates (Supplementary Fig. 1a). Next, MNase is activated by adding calcium, allowing MNase to digest internucleosomal regions of the DNA that are in proximity to the antibody. Without the need for purification steps, nucleosomes are stripped off the DNA, and the genomic fragments are ligated to barcoded adapters containing a unique molecular identifier (UMI) and cell-specific barcode. The genomic fragments are amplified by *in vitro* transcription and PCR, and sequenced (Methods).

To test if sortChIC is sensitive enough to detect histone modifications in single cells, we apply it on the well-characterized human leukemia cell line K562. In these cells, we map four histone modifications that represent major chromatin states regulating gene expression (Fig. 1b-e). For modifications associated with gene activation, we profile $H_3K_4me_1$ (Fig. 1b), found at active enhancers and promoters and $H_3K_4me_3$ (Fig. 1c), found at the promoters of active genes [79]. For modifications associated with repression, we profile H_3-

K_9me_3 (Fig. 1d) and $H_3K_{27}me_3$ (Fig. 1e), found in gene-poor and gene-rich regions, respectively [20].

For each of the four histone modifications, we process 1128 K562 cells in the G1 phase of the cell cycle to ensure a single genome copy per cell. Using the position of the MNase cut site and unique molecular identifies (UMIs), we map unique MNase cut sites genome-wide (Methods). We use a combination of total unique cuts recovered and fraction of cuts in the MNase-preferred AT context to remove low quality cells (Supplementary Fig. 1b, Methods). Overall, 4176 / 4608 of the cells met our criteria, with a mean of 15000 unique cuts per cell (Supplementary Fig. 1b) with a median of 80% of the sequenced reads falling in peaks identified by adding the sortChIC signal over all cells (Supplementary Fig. 1c, Methods).

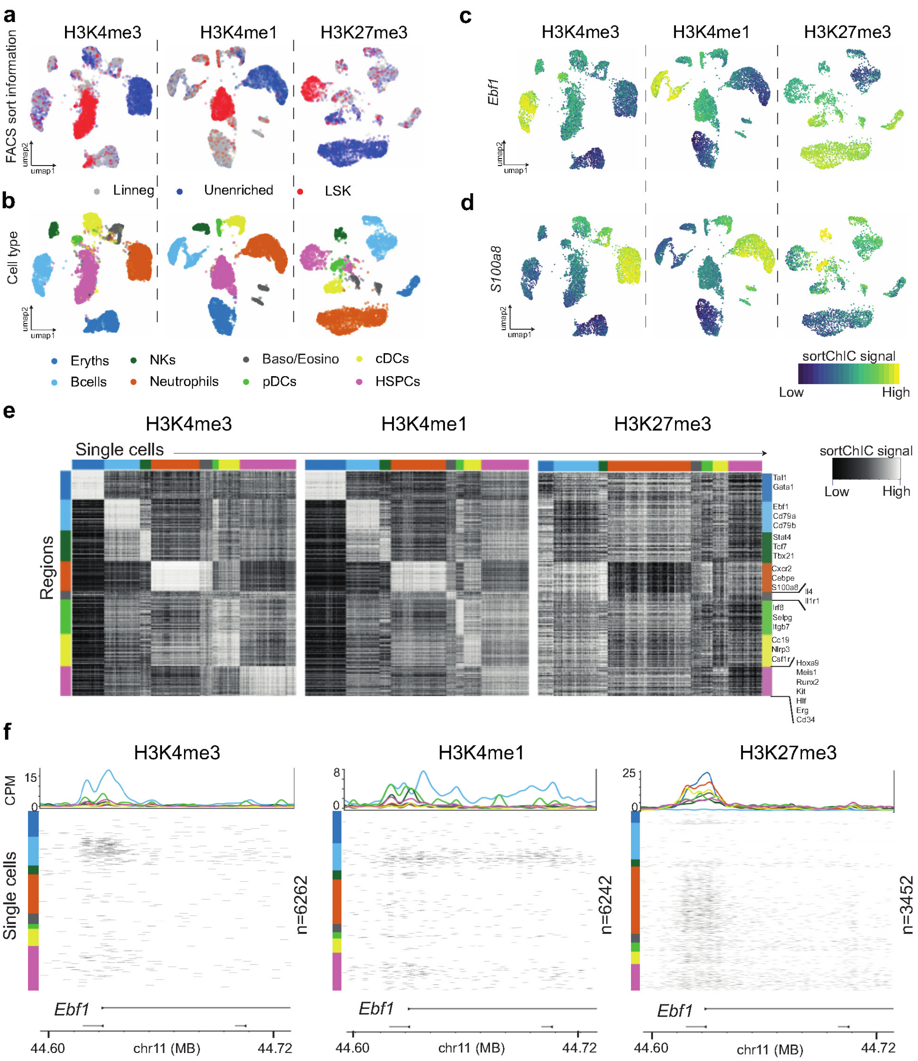
We compare pseudobulk sortChIC profiles with publicly available bulk ChIP-seq results [50], showing high correlation (Pearson correlation > 0.8) between sortChIC pseudobulk and the ChIP-seq signal for each of their respective marks (Supplementary Fig. 1d-e). Single-cell tracks underneath each average track (Fig. 1b) illustrate the high reproducibility of the signal between cells. Of note, the $H_3K_9me_3$ histone modification profiles obtained from sortChIC represent the heterochromatin state without the need for input normalization (Supplementary Fig. 1f), a procedure that is often needed in classical ChIP experiments [170]. Lastly, we compared the sensitivity and specificity of sortChIC with existing high throughput single cell chromatin methods. We compared our profiling of $H_3K_{27}me_3$ by determining the number of unique fragments and the fraction of fragments falling into peaks per cell, then comparing it with H_3K_{27}-me_3 profiling from scChIP-seq and Tn5-based methods (Supplementary Fig. 1g). In both aspects, sortChIC performs equally or better than scChIP-seq and Tn5-based methods. Overall, sortChIC accurately reveals active and repressive chromatin landscapes in single cells.

Active chromatin in HSPCs primes for different blood cell fates, while $H_3K_{27}me_3$ repressive chromatin in differentiated cell types silences genes of alternative fates.

We next map active and repressive chromatin changes during blood formation. We combine sortChIC with cell surface marker staining against lineage markers, Sca-1, and c-Kit to sort abundant and rare cell types from the mouse bone marrow in parallel and map histone modifications associated with different chromatin states (Supplementary Fig. 2a). Dimensionality reduction based on a multinomial model (Methods) and then visualizing this latent space with Uniform Manifold Approximation and Projection (UMAP) reveals distinct clusters that contain LSKs ($Lin^{-}Sca1^{+}cKit^{+}$ sorted cells), unenriched cell types, and mixtures of lineage negative (Lin^{-}) and unenriched cell types (Fig. 2a, Supplementary Fig. 2b). We use the $H_3K_4me_3$ signal in cluster-specific promotor regions (transcription start site (TSS) +/- 5 kb) to determine marker genes for eight blood cell types (Fig. 2b, Methods). These regions are associated with known cell type-specific genes such as the B cell-specific transcription factor, *Ebf1* (Fig. 2c), and the neutrophil-specific gene, *S100a8* (Fig. 2d). For $H_3K_4me_1$ and $H_3K_4me_3$, these regions are marked in their respective cell types, while for $H_3K_{27}me_3$ these regions show specific depletion in their respective cell types (Fig. 2e). Using a publicly available dataset [4], we analyze the mRNA abundances associated with our cell type-specific regions across blood cell types and confirm that these sets of genes are cell type-specific (Supplementary Fig. 2c). Our sortChIC data produces high resolution maps of histone modifications in single cells. For example, the TSS of a B cell-specific transcription factor, *Ebf1*, shows B-cell specific signal in $H_3K_4me_1$ and $H_3K_4me_3$. For $H_3K_{27}me_3$, *Ebf1* is upregulated in non-B cells and depleted in B cells (Fig. 2f and Supplementary Fig. 2d-f). Interestingly, we find that hematopoietic stem and early progenitor cells (HSPCs) already have $H_3K_4me_3$ and $H_3K_4me_1$ marks at the *Ebf1* promoter and gene body, respectively, suggesting HSPCs may already have active marks at intermediate levels relative to differentiated cell types.

We extend the *Ebf1* observation to all TSSs in our cell type-specific gene sets. To quantify the changes as HSPCs differentiate into different cell types, we compared fold changes between differentiated cell types relative to HSPCs across each of our eight sets of cell type-specific genes derived from H_3K_4-me_3 (Supplementary Fig. 2d-f, Methods). When compared to the HSPCs, we find that changes in active chromatin levels are up- or down-regulated depending on the cell fate. For example, at B cell-specific genes, active chromatin levels increase from HSPCs to B cells and to pDCs, but decrease in basophils/eosinophils, neutrophils, and erythroblasts (Supplementary Fig. 2d, e). This divergent pattern occurs in all differentiated cell type-specific gene sets, suggesting that cell type-specific regions in HSPCs already have an intermediate level of active chromatin marks, which are then modulated up or down depending on the cell type into which the HSPCs differentiate.

Figure 2



(Caption on next page.)

Fig. 2: Active and repressive chromatin states in single cells from the mouse bone marrow. (a) UMAPs of $H_3K_4me_3$, $H_3K_4me_1$, and $H_3K_{27}me_3$ single-cell epigenomes from whole bone marrow (unenriched), lineage negative (Lin^-), and $Lin^-Sca1^+cKit^+$ (LSK) sorted populations. (b) UMAPs colored by cell type. Eryths: erythroblasts, NKs: natural killer cells, Baso/Eosino: basophils/eosinophils, pDCs: plasmacytoid dendritic cells, cDCs: common dendritic cells, HSPCs: hematopoietic stem cells and early progenitor cells, (c) UMAP summary colored by sortChIC signal in a region +/- 5 kb centered at the transcription start site of *Ebf1*, a B cell-specific gene. (d) Same as (c) but for a region around *S100a8*, a neutrophil-specific gene. (e) Heatmap of sortChIC signals for regions around cell type-specific genes showing high levels of active marks ($H_3K_4me_1$, $H_3K_4me_3$) in their respective cell type, and correspondingly low levels in the repressive mark ($H_3K_{27}me_3$). (f) Example of active and repressive chromatin states near the transcription start site of a B cell specific transcription factor *Ebf1*. $H_3K_4me_3$ and $H_3K_4me_1$ show large number of cuts specifically in B cells; $H_3K_{27}me_3$ shows B cell-specific depletion of cuts. Colored line plots (same color code as in b) represent the average sortChIC signal for cells of the same cell. Individual cells are ordered by cell type, color coded on the left.

Repressive $H_3K_{27}me_3$ marks at B cell-specific genes, by contrast, are upregulated in non-B cells compared to HSPCs, while only a subset of them loses $H_3K_{27}me_3$ when differentiating into B-cells (Supplementary Fig. 2f). Across other cell type-specific genes, we observe a similar trend where HSPCs upregulate $H_3K_{27}me_3$ at genes specific for alternative cell fates, likely silencing cell type-inappropriate genes. This upregulation suggests that a main role of $H_3K_{27}me_3$ during hematopoiesis is to silence genes of alternative blood cell fates.

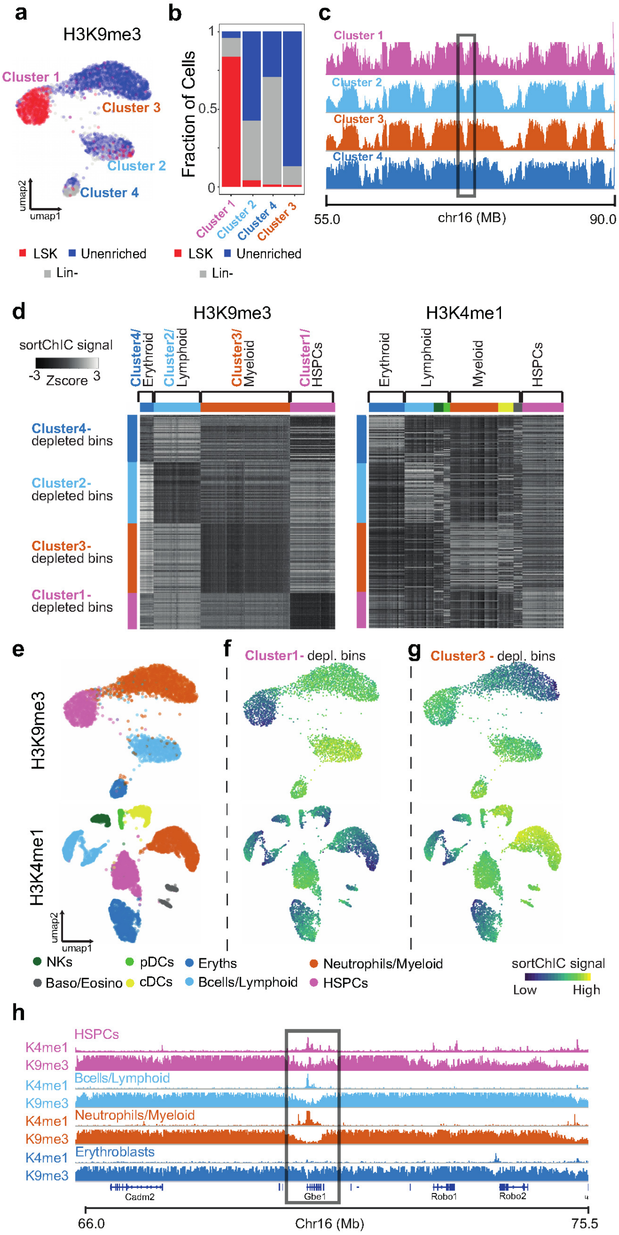
In sum, our analysis at blood cell type-specific genes shows that active chromatin primes HSPCs for different blood cell fates, while $H_3K_{27}me_3$ repressive chromatin during hematopoiesis silences genes of alternative fates.

Dynamic $H_3K_9me_3$ regions reveal clusters enriched for HSPCs, erythroid, myeloid, and lymphoid lineages.

To understand chromatin regulation in gene-poor heterochromatic regions, we map $H_3K_9me_3$ modifications from the same technical batch of HSPCs, lineage negative, and unenriched bone marrow cells as was used for the $H_3K_4me_1$, $H_3K_4me_3$, and $H_3K_{27}me_3$ sortChIC experiments. In contrast to the other three marks, where we find eight cell types, $H_3K_9me_3$ analysis reveals four clusters: one cluster containing mostly LSKs, one cluster containing mostly unenriched cells, and two clusters containing a mixture of unenriched and lineage negative cells (Fig. 3a, b). We find large megabase-scale domains marked by $H_3K_9me_3$ in gene-poor regions that are constant across cell types, but also smaller submegabase regions with cluster-specific signal, suggesting that there may be differences reflecting different cell types or lineages, despite having large regions of heterochromatin that are conserved across cell types (Fig. 3c). Differential analysis on 50 kb regions across the genome identified 6085 cluster-specific regions (q-value $< 10^{-9}$, deviance goodness-of-fit test from Poisson regression, Methods) for $H_3K_9me_3$. These cluster-specific $H_3K_9me_3$ regions have a median distance of 62.8 kb to the nearest TSS of a gene, and are closer to TSSs than $H_3K_9me_3$-marked regions in general, which have a median distance of 138 kb to a TSS (Supplementary Fig. 3a). This suggests that some cluster-specific $H_3K_9me_3$ regions may be associated with gene regulation.

Since cluster-specific $H_3K_9me_3$ regions were closer to TSSs than general $H_3K_9me_3$-marked regions, we hypothesize that the $H_3K_4me_1$ mark in these same regions may also show cluster-specific signal. Out of the 6085 cluster-specific $H_3K_9me_3$ regions, we select 150 regions with the largest depletion of the $H_3K_9me_3$ sortChIC signal relative to HSPCs for each of the four clusters, resulting in four sets of cluster-specific regions (Supplementary Fig. 3b). Comparing the $H_3K_4me_1$ signal in each of these four sets of regions shows cell type-specific retention in these regions (Supplementary Fig. 3c), consistent with an anticorrelated relationship with $H_3K_9me_3$. Heatmaps of the $H_3K_9me_3$ and $H_3K_4me_1$ signal at the four sets of regions reveal an anticorrelated structure between $H_3K_9me_3$ versus $H_3K_4me_1$ (Fig. 3d), allowing the prediction of cell types in the $H_3K_9me_3$ data (Supplementary Fig. 3b, c). Erythroid regions are upregulated in erythroblasts in $H_3K_4me_1$; lymphoid regions show strongest upregulation in B cells; myeloid regions show strongest upregulation in neutrophils. We use this anticorrelation at cluster-specific $H_3K_9me_3$ regions to identify cells related to erythroid, lymphoid, and myeloid lineages in $H_3K_9me_3$ (Fig. 3e). We find that regions depleted of $H_3K_9me_3$ in HSPCs show upregulation of $H_3K_4me_1$ in HSPCs (Fig. 3f). For $H_3K_9me_3$-depleted regions in myeloid cells, we find that $H_3K_4me_1$ is upregulated not only in neutrophils, but also in other cell types that share the myeloid lineage, such as common dendritic cells (Fig. 3g). This anticorrelation is exemplified in a genomic region surrounding the *Gbe1* gene, showing repression of $H_3K_4me_1$ signal specifically in erythroblasts accompanied by high levels of $H_3K_9me_3$. In this region, HSPCs, lymphoid, and myeloid cell types show enrichment of $H_3K_4me_1$ accompanied by a marked depletion in $H_3K_9me_3$ (Fig. 3h). At these lineage-specific $H_3K_9me_3$ regions, we also see cell type-specific signal in $H_3K_4me_3$ and in $H_3K_{27}me_3$, although the pattern is weaker than in $H_3K_4me_1$ (Supplementary Fig. 3d). Overall, we find the $H_3K_9me_3$ clusters are related to HSPCs, erythroid, lymphoid, and myeloid lineages.

Figure 3



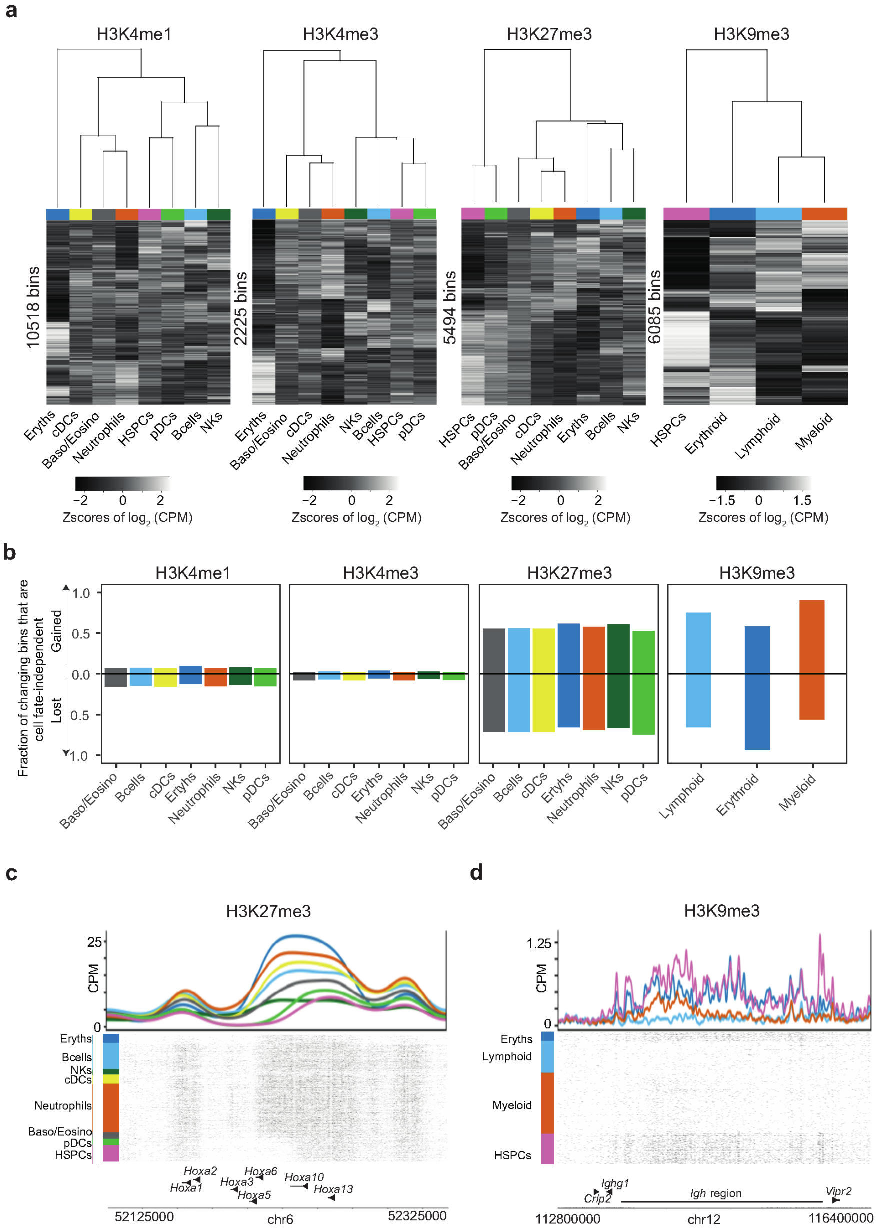
(Caption on next page.)

Fig. 3: heterochromatin state dynamics during hematopoiesis. (a) UMAP of $H_3K_9me_3$ representing single cells from whole bone marrow (unenriched), lineage negative (Lin-), and Lin-Sca1+cKit+ (LSK) sorted cells. (b) Fraction of unenriched, Lin^-, and $Lin^-Sca1^+cKit^+$ cells in each of the four $H_3K_9me_3$ clusters. (c) Region showing the $H_3K_9me_3$ pseudobulk sortChIC signal of the four clusters. (d) Heatmap of 50 kb bins displaying the relative $H_3K_9me_3$ (left) and $H_3K_{27}me_3$ (right) sortChiC signal in erythroblasts, lymphoid, myeloid, and HSPCs. (e) UMAP of $H_3K_4me_1$ and $H_3K_9me_3$ sortChIC data, colored by cell type. (f) Single-cell signal of cluster1-depleted bins (averaged across the 150 bins) showing low $H_3K_9me_3$ and high $H_3K_4me_1$ signal in lymphoid cells. Both $H_3K_9me_3$ (above) and $H_3K_4me_1$ (below) are quantified using the same set of bins. (g) Single-cell signal of cluster3-specific bins showing low $H_3K_9me_3$ and high $H_3K_4me_1$ signal in myeloid cells. (h) Zoom-in of the same genomic region in (b) for $H_3K_9me_3$ and $H_3K_4me_1$ pseudobulk sortChIC signal.

Repressive chromatin dynamics are largely cell fate-independent.

We ask whether global patterns in chromatin dynamics during hematopoiesis differed between repressive and active marks. We apply differential analysis on 50 kb regions for all four marks, resulting in 10518 dynamic bins for $H_3K_4me_1$, 2225 for $H_3K_4me_3$, 5494 for $H_3K_{27}me_3$, and 6085 for $H_3K_9me_3$ (q-value $< 10^{-50}$ for $H_3K_{27}me_3$, $H_3K_4me_1$, and $H_3K_4me_3$; q-value $< 10^{-9}$ for $H_3K_9me_3$, Supplementary Table 1, Methods). For each histone modification, we cluster the pseudobulk signal of each cell type across the bins. Hierarchical clustering reveals global relationships between cell types. In active marks, we find that the largest differences come from erythroblast versus non-erythroblasts (Fig. 4a, left two panels). This erythroblast distinction corroborates with our analysis of fold changes at TSSs of cell type-specific genes, where the erythroblasts shows the largest changes in active chromatin during differentiation (Supplementary Fig. 2d, e). Furthermore, we find that HSPCs often have intermediate levels of $H_3K_4me_1$ and $H_3K_4me_3$ (Fig. 4a, left two panels), suggesting a generally more accessible chromatin state HSPCs.

Figure 4



(Caption on next page.)

Fig. 4: Repressive chromatin dynamics are largely cell fate-independent. (a) Heatmap of $\log_2$ counts per million (CPM) of 50 kilobase bins across pseudobulks. Changing bins that are statistically significant are shown (deviance goodness-of-fit test from Poisson regression, Methods). The rows and columns are ordered by complete-linkage clustering. Above each heatmap is a dendrogram from clustering the columns, showing the relationship between cell types. (b) Barplot of the fraction of changing bins (Methods) that are gained or lost in all non-HSPCs relative to HSPCs. Each cell type shows two bars, one for each direction (either gained or lost). Fraction is calculated by dividing the number of bins that change cell fate-independently by the number of bins that change in that cell type for that direction. (c) Genome browser view of the *Hoxa* region showing a $H_3K_{27}me_3$ domain that is gained during hematopoiesis. (d) Genome view of the immunoglobulin heavy chain (*IgH*) region displaying the loss of a $H_3K_9me_3$ domain in lymphoid and myeloid cells.

Projecting the active mark data onto the two most significant axes of chromatin variation [171], shows that the HSPCs take a central position relative to other cell types, suggesting that changes in active chromatin during hematopoiesis can diverge depending on the specific cell fate (Supplementary Fig. 4a, left two panels).

By contrast, repressive chromatin dynamics, marked by $H_3K_{27}me_3$ and $H_3K_9me_3$, mainly distinguish between HSPCs and differentiated cell types, thereby marking the progress along the differentiation trajectory (Fig. 4a, right two panels). Projecting the repressive mark data reveals an axis connecting HSPCs and other cell types (Sup. Fig. 4a, right two panels). To ask whether regions gain or lose chromatin marks depending on the specific cell fate, we calculate the fraction of changing bins that gain or lose chromatin marks in all non-HSPCs relative to HSPCs (Methods). We find more than half of bins that gain or lose repressive marks between one cell type versus HSPCs are also gaining or losing marks across all other cell fates (Fig. 4b), suggesting that many changes in repressive chromatin during hematopoiesis occur independent of the specific cell fate. By contrast, only 8 percent of bins in active chromatin, on average, show cell type-independent changes. Fold changes between HSPCs and non-HSPCs at changing bins show distinct separation between HSPCs and non-HSPCs in repressive marks, but not in active marks (Supplementary Fig.

4b), corroborating that many changes in repressive chromatin are independent of cell fate. These cell fate-independent changes are exemplified for $H_3K_{27}me_3$ at the *Hoxa* region, which shows low levels of $H_3K_{27}me_3$, and its levels are upregulated in differentiated cell types (Fig. 4c). HSPCs at the *Igh* region show high levels of $H_3K_9me_3$, and its levels are downregulated in myeloid and lymphoid cells, suggesting that this *Igh* region, which encodes the heavy chains of immunoglobulins, are de-repressed during differentiation (Fig. 4d).

We ask whether $H_3K_{27}me_3$ and $H_3K_9me_3$ may regulate distinct processes. We confirm that $H_3K_{27}me_3$ dynamics occur at GC-rich regions close to TSSs while $H_3K_9me_3$ dynamics at AT-rich regions occur further from TSSs (Supplementary Fig. 4c, d), consistent with known sequence-specific contexts of the two repressive marks [20]. GO term analysis of $H_3K_9me_3$ regions unique to HSPCs shows enrichment for immune-related processes such as phagocytosis, complement activation, and B cell receptor signaling (Supplementary Fig. 4e), suggesting that HSPCs use $H_3K_9me_3$ to repress genes that may later need to be used in differentiated blood cells. By contrast, GO term analysis of $H_3K_{27}me_3$ regions unique to HSPCs does not show consistent enrichment for biological processes related to blood development.

Taken together, we find that HSPCs have active chromatin marks at intermediate levels relative to other blood cell types. During differentiation, active marks at these regions can then be up- or down-regulated depending on the specific cell fate. By contrast, most dynamic repressive chromatin regions are gained or lost independent of the specific cell fate.

Transcription factor motifs underlie active and repressive chromatin dynamics in hematopoiesis.

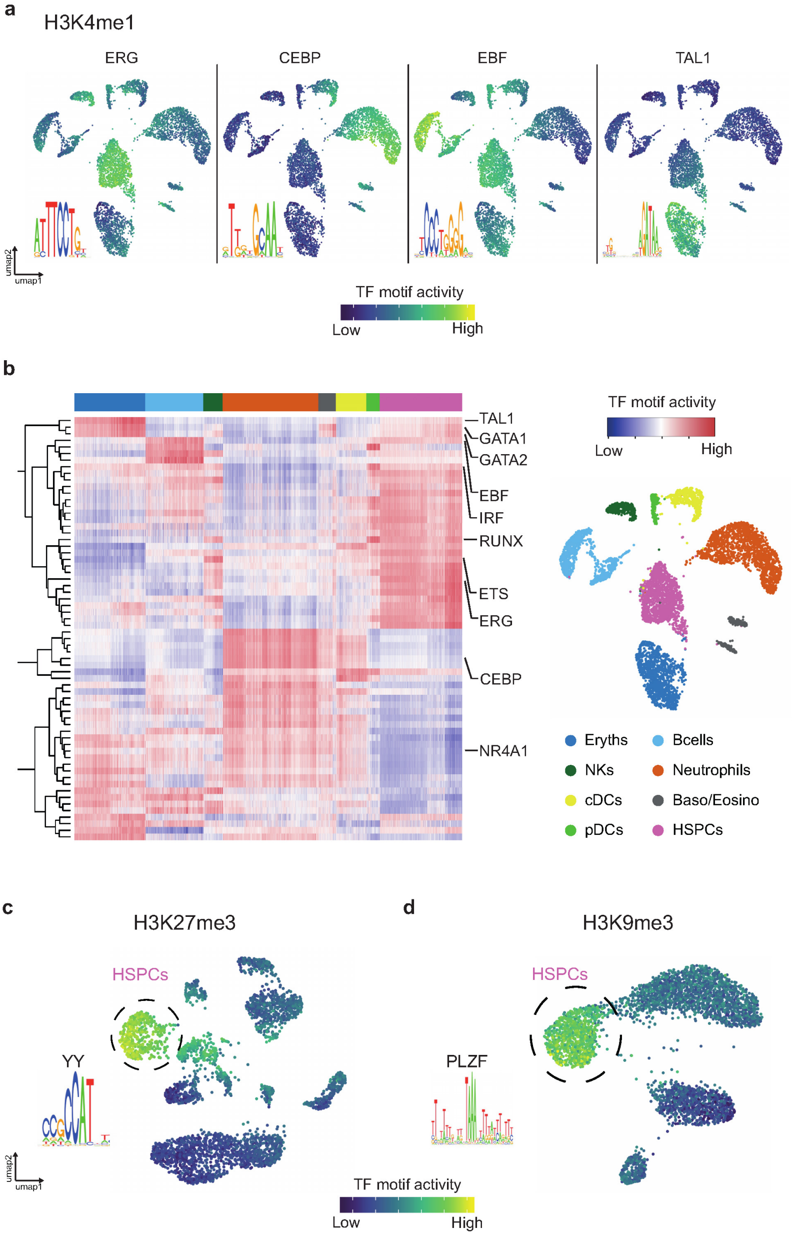
Next, we ask whether regulatory information in the DNA sequences underlying the sortChIC data can explain the cell type-specific distributions in active and repressive chromatin landscapes. We hypothesize that regions with correlated sortChIC signal across cells can be explained in part by transcription factor binding motifs shared across these regions [9, 92] (Sup. Fig. 5a, Methods). Overlaying the predicted single-cell TF motif activities onto the UMAP representation shows the expected blood cell type-specific activity for known regulators. We find the ERG motif active specifically in HSPCs (Fig. 5a, left), consistent with the role of ERG in maintenance of hematopoietic stem cells [92]. CEBP family motif is active in neutrophils (Fig. 5a, mid left), consistent with its proposed function [153, 46]. EBF motif activity is specific to B cells (Fig. 5a, mid right), consistent with its role in specifying B cell differentiation [177]. We find TAL1 to have erythroblast-specific activity (Fig. 5a, right), in agreement with its role in erythropoiesis [75].

We summarize the inferred single-cell TF activities in a heatmap to comprehensively predict transcriptional regulators underlying the cell type-specific distribution of active chromatin (Fig. 5a, b). We predict motifs active in pDCs belonging to the IRF and RUNX family (Fig. 5b, Supplementary Fig. 5b-d). High activity of IRF motifs corroborates with pDC function to secrete type 1 interferon [80, 180]. Runx proteins have been shown to regulate development of dendritic cell progenitors [148] and migration of pDCs from the bone marrow [149]. We find NK cells to have high ETS family motif activity (Fig. 5b, Sup. Fig. 5b, e), consistent with the role *of Ets1* in development of natural killer and innate lymphocyte cells [15, 201]. Finally, we predict transcription factors that have low activity in HSPCs and pDCs but high activity in other cell types, such as the NR4A family (Fig. 5b, Sup. Fig. 5, b and f. *Nr4a1* has been shown to repress gene expression [126] and control hematopoietic stem cell quiescence by suppressing inflammatory signaling [65]. The low activity of several TFs specific in HSPCs and pDCs suggests that the pDCs we identify could be in a more progenitor-like state, consistent with the pseudobulk clustering results in $H_3K_4me_1$, $H_3K_4me_3$ and $H_3K_{27}me_3$ (Fig. 4a).

We apply our TF motif analysis to the two repressive chromatin landscapes to predict motifs that explain HSPC-specific distributions of repressive chromatin. In $H_3K_{27}me_3$, we predict a CCAT motif belonging to the Yin Yang family [90] specifically active in HSPCs (Fig. 5c). Of note, *Yy1* is gene encoding the polycomb group protein and has been shown to regulate hematopoietic stem cell self-renewal [108]. In $H_3K_9me_3$, we predict an AT-rich motif belonging to the transcriptional repressor PLZF specifically active in HSPCs (Fig. 5d). PLZF has been implicated in regulating the cell cycle of hematopoietic stem cells [178].

In sum, our motif analysis explains differences in chromatin levels across cells in terms of TF activities. Our predictions suggest that differentiating blood cells decide which active regions to up- or down-regulate depending on the cell type-specific TFs that associate with different regions. Although repressive chromatin dynamics in both $H_3K_{27}me_3$ and $H_3K_9me_3$ are mainly cell fate-independent, our analysis suggests that distinct TFs regulate the two separate pathways.

Figure 5



(Caption on next page.)

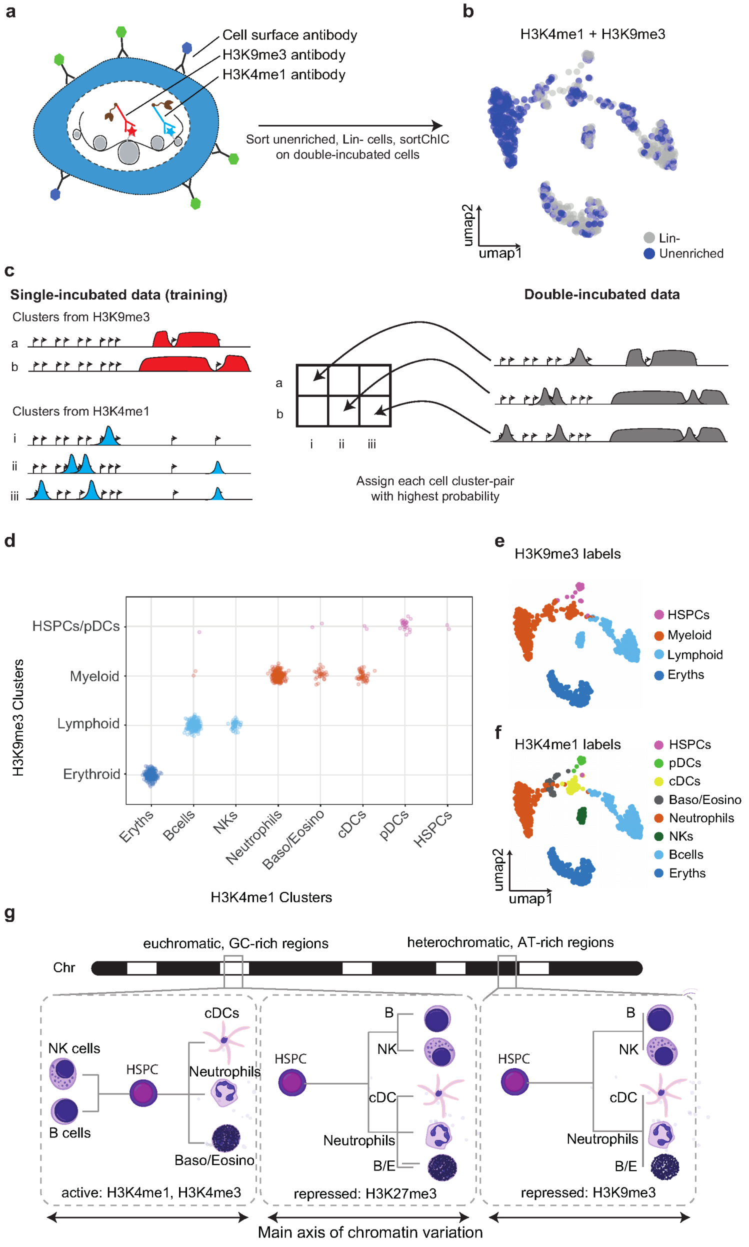
Fig. 5: Transcription factor motifs underlie active and repressive chromatin dynamics in hematopoiesis. (a) Examples of four transcription factor (TF) motifs whose activities are predicted to drive cell type-specific $H_3K_4me_1$ distributions. The ERG motif is predicted to be active in HSPCs, the CEBP motif in neutrophils, the EBF motif in B cells, and the TAL1 motif in erythroblasts. Cell type for each cell cluster is labeled in (b). (b) Heatmap of $H_3K_4me_1$ TF motif activities in single cells. Rows represent motifs. Columns are individual cells whose cell types are annotated by the top color bar. The right panel shows a $H_3K_4me_1$ UMAP colored by cell types, with cell type-to-color legend below. (c) Predicted $H_3K_{27}me_3$ activity of a motif belonging to the Yin Yang (YY) protein family in single cells. Circled cluster is enriched for HSPCs. (d) Predicted $H_3K_9me_3$ activity of PLZF motif in single cells. Circled cluster is enriched for HSPCs.

Distinct cell types can share similar heterochromatin landscapes.

To understand in more detail the relationship between the eight cell types identified by histone marks of gene-rich regions ($H_3K_4me_1$, $H_3K_4me_3$, and H_3-$K_{27}me_3$) to the four clusters identified by $H_3K_9me_3$, we stain cells with both $H_3K_4me_1$ and $H_3K_9me_3$ antibody concurrently [191]. This double-incubation strategy generates cuts that come from both $H_3K_4me_1$ and $H_3K_9me_3$ from the same cell, and uses our sortChIC resource to infer the relationships between the two marks in single cells (Fig. 6a). We sort Lin^- and unenriched cells to profile both abundant and rare cell types. UMAP of the joint landscape reveals clusters that are depleted of mature lineage markers as well as enriched for mature cell types (Fig. 6b). We use clusters from $H_3K_4me_1$ and $H_3K_9me_3$ single-incubated data to develop a model of how the double-incubated data could be generated (Fig. 6c, Methods).

For our model, we focus on regions that show large differences between cell types. We select 811 regions associated with cell type-specific genes (Methods) found in our $H_3K_4me_1$ analysis (Fig. 2e) and 6085 cluster-specific regions (50 kb bins) found in our $H_3K_9me_3$ analysis (Fig. 4a, right panel) as features in our model, making a total of 6896 regions. To verify that our features show cluster-specific differences, we cluster the single-incubated $H_3K_4me_1$ signal across cell types (Supplementary Fig. 6a). We find that neutrophils, basophils/eosinophils, and cDCs cluster together, consistent with their common myeloid lineage [4]. B cells and NK cells also cluster together, consistent with their common lymphoid lineage [181]. Erythroblasts form a distinct branch from other cell types. Finally, pDCs cluster most closely with HSPCs.

Figure 6



(Caption on next page.)

Fig. 6: Distinct cell types can share similar heterochromatin landscapes. (a) Double incubation experiment produces cuts associated with either $H_3K_4me_1$ or $H_3K_9me_3$. ($H_3K_4me_1$+$H_3K_9me_3$) (b) UMAP representation of the $H_3K_4me_1$+$H_3K_9me_3$ landscape in unenriched and lineage-negative cells in the bone marrow. (c) Schematic of how the standard single-incubated data can produce a model of which cluster-pair (one from $H_3K_9me_3$, the other from $H_3K_4me_1$) generates the observed double-incubated data. (d) Output of cluster-pair predictions from $H_3K_4me_1$+$H_3K_9me_3$ double-incubated cells. Cells are colored by their predicted $H_3K_9me_3$ clusters. (e and f) UMAP representation of the $H_3K_4me_1$+$H_3K_9me_3$ landscape, colored by their predicted $H_3K_4me_1$ cluster (e) or $H_3K_9me_3$ cluster (f). (g) Graphical summary of chromatin dynamics as dendrograms showing relationships between HSPCs and differentiated cells. During hematopoiesis, the direction of change in active chromatin depends on the specific cell fate, resulting in global differences that are largest between differentiated cell types from different lineages. By contrast, many regions gain or lose repressive marks during hematopoiesis independent of the specific cell fate, resulting in global differences that are largest between HSPCs and differentiated cell types. Dynamics in active marks and $H_3K_{27}me_3$-marked repressive chromatin reveal cell type information, while dynamics in heterochromatin regions marked by $H_3K_9me_3$ reveal lineage information.

Since *a priori* we do not know which cluster from $H_3K_4me_1$ pairs with which cluster from $H_3K_9me_3$, we generate an *in silico* model of all possible pairings (Fig. 6c, left). For each double-incubated cell, we then perform model selection to choose the pair with the highest probability (Fig. 6c, right, and Supplementary Fig. 6c, d; Methods). This selection reveals that neutrophils, basophils/eosinophils, and cDCs share a common heterochromatin landscape, reflecting their myeloid lineage (Fig. 6d). We find B-cells and NK cell share a lymphoid-specific heterochromatin. Erythroblasts do not share a heterochromatin landscape with any other cell type. Although we did not explicitly sort for HSPCs, a small fraction of cells was assigned to both an HSPC-specific active chromatin and HSPC-specific heterochromatin state, reflecting the rarity of HSPCs in the bone marrow. Surprisingly, we find pDCs associated with the HSPC-enriched $H_3K_9me_3$ landscape, suggesting that these cells that we sorted may have already committed towards a pDC fate through its active chromatin, while its heterochromatin remains undifferentiated.

This joint analysis confirms that distinct cell types in related lineages can share their heterochromatin state (Fig. 6e, f), suggesting a hierarchical model where changes in heterochromatin establish lineages and changes in active chromatin define cell types within lineages.

3.4 Discussion

We profile and analyze active and repressive chromatin states in single cells during blood formation, providing a comprehensive map of chromatin regulation at both euchromatic and heterochromatic regions. We find that repressive chromatin shows distinct dynamics compared with active chromatin, demonstrating that profiling repressive chromatin regulation in single cells reveals novel dynamics not captured by profiling active chromatin. Active chromatin primes HSPCs, and is up- or down-regulated depending on the specific cell fate, mediated by cell type-specific transcription factors. Consequently, active chromatin shows divergent changes for different blood cell fates, resulting in global differences in active chromatin that are larger between mature cell types than between HSPCs and mature cell types (Fig. 6g, left panel). These active chromatin dynamics likely reflect the dynamics in mRNA abundances [101]. By contrast, changes in repressive chromatin during hematopoiesis often occur in the same direction (either gained or lost) regardless of the specific cell fate, resulting in global differences in repressive chromatin that are larger between HSPCs and mature cell types than between mature cell types (Fig. 6g, middle and right panel). Overall, our results show that single-cell repressive chromatin dynamics provide an orthogonal viewpoint to active chromatin and mRNA dynamics during hematopoiesis.

Technologies to profile histone modifications in single cells by sequencing is still in its infancy, but has the potential to unlock the spectrum of chromatin states in the genome of individual cells. The ideal assay strives to have high sensitivity, high throughput, and robustness in both active and repressive chromatin states. Current techniques to map histone modifications in single cells use one of three approaches: ChIP-based, pA-Tn5-based, and pA-MNase-based. ChIP-based strategies utilize microfluidics systems or combinatorial barcoding to overcome the low sensitivity of ChIP [141, 73, 3]. pA-Tn5-based strategies profile histone modifications with very high throughput, but due to the intrinsic affinity of Tn5 to open chromatin regions [89, 77, 187, 16, 86, 181], high specificity can so far only be achieved at the cost of some sensitivity. pA-MNase-based methods profile histone modifications with high sensitivity, and have robust detection of modifications associated with euchromatic regions as well as heterochromatic regions, but has generally less throughput compared with Tn5-based methods [95, 96, 74]. SortChIC is a unique single-cell method that combines cell enrichment to greatly enhance throughput of rare cells, while achieving high sensitivity and robustness to profile active and repressive chromatin states (Supplementary Fig. 1g, Supplementary Table 2), thereby complementing newly available high throughput Tn5-based methods [16].

This comprehensive profiling of rare progenitors and their multiple cell fates enables new systematic analyses, such as quantifying chromatin dynamics that are cell fate-independent during differentiation. This analysis reveals that cell fate-independent changes during differentiation occur frequently for repressive chromatin, while such changes for active chromatin are rare. Our strategy combines rare progenitor cell enrichment with comprehensive differentiated cell type profiling to allow systematic analysis of chromatin dynamics during differentiation into multiple cell fates.

Our single-cell analysis further expands the role of H3K9me3, which has been classically associated with constitutive types of chromatin [122]. We find that H3K9me3 is a dynamic chromatin modification that regulates different lineages in blood and is rewired as HSPCs differentiate into different blood lineages. Although *in vivo* dynamics in H3K9me3 have been recently reported during early development [122, 64, 67], our results extend the role of H3K9me3 dynamics to also regulate homeostatic renewal in adult physiology.

Joint profiling analysis demonstrates that cell types from a common lineage can share a similar heterochromatin landscape. Our results suggest that the distinct chromatin dynamics in active chromatin and heterochromatin can reveal the hierarchical relationships between cell types. We find cDCs, neutrophils, and basophils/eosinophils to share a similar myeloid-specific heterochromatin landscape, suggesting that the H3K9me3 mediated heterochromatin can be relatively stable while other chromatin changes further distinguish between distinct cell types. pDCs have been reported to come from lymphoid precursors [59], although the exact origin has been debated [138]. We find pDCs to be distinct from cDCs at both the active and heterochromatin level, although the heterochromatin of pDCs is also distinct from other lymphoid cell types such as B cells or NK cells. One explanation could be that pDCs diverge early from other lymphoid cell types and do not participate in the heterochromatin rewiring that occurs during lymphopoiesis. Overall, we propose a hierarchical chromatin regulation program during hematopoiesis, in which heterochromatin states define a differentiation trajectory and lineages, while euchromatin states establish cell types.

Acknowledgements: We thank R. van der Linden for assistance during experiments. We thank Amir Giladi for sharing mRNA abundance tables of cell types and critical reading of the manuscript. We thank Marek Bartosovic for sharing method comparison data.This work was supported by a European Research Council Advanced grant (ERC-AdG 742225-IntScOmics), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) TOP award (NWO-CW 714.016.001), the Swiss National Science Foundation, and the Human Frontiers for Science Program. This work is part of the Oncode Institute which is partly financed by the Dutch Cancer Society.

Contributions:

P.Z., J.Y., and A.v.O. designed the project; P.Z. and M.F. developed technique; P.Z. and H.V.G. performed experiments; J.Y. developed and applied the statistical methods. B.A.d.B., M.F., and J.Y. wrote the sortChIC demultiplexing and preprocessing pipeline. J.Y., A.v.O, and P.Z. analyzed the data. P.Z., J.Y. and A.v.O, wrote the manuscript.

Competing interests:

The authors declare no competing interests.

3.5 Methods

Cell culture

K562 cells (ATCC® CCL-243™) were grown in RPMI 1640 Medium GlutaMAX™, supplemented with 5% FCS, Pen-Strep and non-essential amino acids. After harvesting cells were washed 3 times with room temperature PBS before continuing with the sortChIC protocol.

Animal experiments

Experimental procedures were approved by the Dier Experimenten Commissie of the Royal Netherlands Academy of Arts and Sciences and performed according to the guidelines. Primary bone marrow cells were harvested from 3-months-old C57BL/6 mice. Femur and Tibia were extracted, the bones ends were cut away to access the bone marrow which was flushed out using a 22G syringe with HBSS (-Ca, -Mg, -phenol red, Gibco 14175053) supplemented with Pen-Strep and 1% FCS. The bone marrow was dissociated and debris was removed by passing it through a 70 µm cell strainer (Corning, 431751). Cells were washed with 25 ml supplemented HBSS before linage marker staining was performed following the instructions of the EasySep™ Mouse Hematopoietic Progenitor Cell Isolation Kit (Stemcell) at half of the recommended concentration of the biotinylated antibodies. This was followed by 30min incubation at 4 °C with a Streptavidin-PE (Biolegend, 1:5000), anti c-kit-APC (Biolegend, 1:800) and anti sca1-PeCy7 (Biolegend, 1:400). After 2 additional washes with HBBS (+PS, +FCS) cells were prepared following the sortChIC protocol for the 4 different histone modifications.

Pa-MN production

The Pa-MN fusion protein was produced following the methods section in [24]. pK19pA-MN was a gift from Ulrich Laemmli (Addgene plasmid # 86973; http://n2t.net/addgene:86973; RRID: Addgene_86973)

sortChIC-seq experiments

Cell preparation: fixation

All steps were performed on ice. Cells were resuspended in 300 µl PBS per 1 million cells in a 15 ml protein low binding falcon tube and 700 µl ethanol (-20 °C precooled) per 1 million cells are added while vertexing cells at middle speed. Cells were fixed for 1 h at -20 °C. After fixation cells were washed twice in 1 ml wash buffer (47.5 ml H_2O RNAse free, 1 ml 1M HEPES pH 7.5 (Invitrogen), 1.5 ml 5M NaCl, 3.6 µl pure spermidine solution (Sigma Aldrich), 0.05% Tween20, protease inhibitor cocktail (Sigma Aldrich) with 4 µl/ml 0.5 M EDTA). For K562 cells 3 plates were sorted for each modification. For BM we sorted 19, 17, 18, and 17 plates for $H_3K_4me_1$, $H_3K_4me_3$, $H_3K_{27}me_3$, and $H_3K_9me_3$, respectively.

Cell preparation: nuclei

Cells were washed once in 1 ml wash buffer (47.5 ml H2O RNAse free, 1 ml 1M HEPES pH 7.5 (Invitrogen), 1.5 ml 5M NaCl, 3.6ul pure spermidine solution (Sigma Aldrich), 0.05% Saponin, protease inhibitor cocktail (Sigma Aldrich) with 4 µl/ml 0.5 M EDTA). Nuclei were isolated by further Saponin incubation overnight in parallel to the antibody staining. For BM we sorted 9 plates each for $H_3K_4me_1$, $H_3K_4me_3$, and $H_3K_9me_3$.

Antibody staining

Cells were pelleted at 500g for 4 min and resuspended in 200 µl Wash Buffer (+EDTA) per 1 million cells and were aliquoted into 0.5 ml protein low binding tubes containing the primary histone mark antibody (1:200 dilution for $H_3K_4me_1$ and $H_3K_4me_3$ and 1:100 for $H_3K_9me_3$ and $H_3K_{27}me_3$) diluted in 200 µl Wash Buffer (+EDTA). Cells were incubated overnight at 4 °C on a roller, before they were washed once with 500 µl Wash Buffer. In the case of double labeling experiments, cells were incubated with antibodies against $H_3K_4me_1$ and $H_3K_9me_3$ together at the same concentrations as for the single mark experiments. We sorted four plates incubated simultaneously with $H_3K_4me_1$ and $H_3K_9me_3$.

Afterwards cells were resuspended in 500 µl Wash Buffer containing PaMN (3 ng/ml) and Hoechst 34580 (5 µg/ml) and incubated for 1h at 4 °C on a roller.

Finally, cells were washed an additional 2 times with 500 µl Wash Buffer before passing it through a 70 µm cell strainer (Corning, 431751) and sorting G1 cells based on Hoechst staining on an Influx FACS machine into 384 well plates containing 5 µl sterile filtered mineral oil (Sigma Aldrich) per well, always leaving eight wells empty of cells as a negative control. For bone marrow we sorted with the help of the indicated surface marker stainings unenriched (only using G1 gate), lineage negative, and LSK cells into separate parts of 384 well plates. We sorted 28, 26, 18, and 26 plates for $H_3K_4me_1$, $H_3K_4me_3$, $H_3K_{27}me_3$, and $H_3K_9me_3$, respectively. Of these, three plates were sorted with nuclei for unenriched cells only, three for lineage negative only, and three for LSK cells only for $H_3K_4me_1$, $H_3K_4me_3$, and $H_3K_9me_3$. For lineage negative

	Vol. per well (nl)
Klenow large (NEB, M0210L)	2.5
T4 PNK (NEB, M0201L)	2.5
dNTPs 10 mM (Promega, U1515)	6
ATP 100 mM (part of Thermo Fisher Scientific, R0441)	3.5
MgCl2 25 mM (part of Thermo Fisher Scientific, 4398828)	10
PEG8000 50% (Promega, V3011)	7.5
PNK buffer 10X (NEB, B0201S)	35
BSA 20 ng/ml (NEB, B9000S)	1.8
Nuclease-free water (Invitrogen, AM9932)	81.3
Total	150

Supplementary Table 4: end repair mix

only and LSK cells only, cell populations were enriched by FACS before nuclei isolation. Five plates were sorted for both unenriched and lineage negative together. All remaining plates included a mixture of unenriched, lineage negative, and LSK sorted cells. The following small volumes were distributed using a Nanodrop II system (Innovadyme) and plates were spun for 2 min at 4 °C and 2000g after each reagent addition.

Protein A-MN activation

100 nl of Wash Buffer (-protease inhibitor), containing 2 mM $CaCl_2$, were added per well to induce Protein A-MN mediated chromatin digestion that was performed for 30 min in a PCR machine set at 4 °C. Afterwards the reaction was stopped by adding 100 nl of a stop solution containing 40 mM EGTA (chelates Ca^{2+} and stops MN, Thermo, 15425795), 1.5% NP40 and 10 nl 2 mg/ml proteinase K (Invitrogen, AM2548) and incubated for further 20 min at 4 °C. Chromatin is subsequently released and PaMN permanently destroyed by proteinase K digestion at 65 °C for 6h followed by 80 °C for 20 min to heat inactivate proteinase K. Afterwards plates can be stored at -80 °C until further processing.

Library preparation

DNA fragments are blunt ended by adding 150 nl of the following mix per well and incubating for 30 min at 37 °C followed by 20 min at 75 °C for enzyme inactivation.

	Vol. per well (nl)
AmpliTaq 360 (Thermo Fisher Scientific, 4398828)	1
dATPs 100mM (part of Promega, U1335)	1
KCl 1 M (Thermfisher, AM9640G)	25
PEG8000 50%	7.5
BSA 20 ng/ml	0.8
Nuclease-free water	114.8
Total	150

Supplementary Table 5: A-tailing mix

Blunt fragments are subsequently A-Tailed by adding 150 nl per well of the following mix and incubate for 15 min at 72 °C. Through AmpliTaq 360's strong preference to incorporate dATP as a single base overhang even in the presence of other nucleotides, a general dNTP removal is not necessary.
Next fragments are ligated to T-tail containing forked adaptors.
Top strand
GGTGATGCCGGTAATACGACTCACTATAG
GGAGTTCTACAGTCCGACGATCNNN*ACACACTA*T
Bottom strand
*TAGTGTGT*NNNGATCGTCGGACTGTAGAACTC
CCTATAGTGAGTCGTATTACCGGC**GAGCTT**

Sequence features from left to right on the top strand: Bases written in bold form a fork to prevent adaptor dimer- or multimerization. Bases in green represent T7 polymerase binding site for IVT based amplification. Bases in blue are the binding site (RA5) for the TruSeq Small RNA indexing primers (RPIx). The 3 random nucleotides underlined are the unique molecular identifier used for read deduplication and the 8 bases afterwards in italics represent the cell barcode which is different each of the 384 wells. For a full list of adaptors see Supplementary Table 3.

For ligation 50 nl of 5 µM adaptor in 50 mM Tris pH7 is added to each well with a Mosquito HTS (ttp labtech). After centrifugation 150 nl of the following mix are added before plates are incubated for 20 min at 4 °C, followed by 16 h at 16 °C for ligation and 10 min at 65 °C to inactivate ligase.

Ligation products were pooled by centrifugation into oil coated lids of pipettip boxes at 200g for 2 min and the liquid face was transferred into 1.5 ml eependorf tubes and was purified by centrifugation at 13000g for 1 min and

	Volumes per well (nl)
T4 ligase (400K Units/ml, NEB, M0202L)	25
$MgCl_2$ 1 M (ThermoFisher, AM9530G)	3.5
Tris 1 M pH 7.5 (ThermoFisher, 15567027)	10.5
DTT 0.1M (Invitrogen, 15846582)	52.5
ATP 100 mM	3.5
PEG8000 50%	10
BSA 20 ng/ml	1
Nuclease-free water	44
Total	150

Supplementary Table 6: Adaptor ligation mix

transfer into a fresh tube twice. DNA fragments were purified using Ampure XP beads (Beckman Coulter - prediluted 1 in 8 in bead binding buffer – 1 M NaCl, 20% PEG8000, 20 mM TRIS pH=8, 1 mM EDTA) at a bead to sample ratio of 0.8. Beads were washed twice with 1 ml 80% ethanol resuspending the beads during the first wash and resuspended in 8 µl Nuclease-free water and transferred into a fresh 0.5 ml tube. The cleaned DNA is then linear amplified by invitro transcription adding 12 µl of MEGAscript™ T7 Transcription Kit (Fishert Sc, AMB13345) for 12 h at 37 °C. Template DNA is removed by addition of 2 µl TurboDNAse (IVT kit) and incubation for 15 min at 37 °C. The produced RNA is further purified using RNA Clean XP beads (Beckman Coulter) at 0.8 beads to sample ratio, followed by RNA fragmentation for 2 min at 94 °C. After another bead cleanup, 40% (5 µl) of the RNA is primed for reverse transcription by adding 0.5 µl dNTPs (10 mM) and 1 µl randomhexamerRT primer 20 µM (GCCTTGGCACCCGAGAATTCCANNNNNN) and hybridizing it by incubation at 65 °C for 5 min followed by direct cool down on ice. Reverse transcription is performed by further addition of 2 µl first strand buffer (part of Invitrogen, 18064014), 1 µl DTT 0.1M (Invitrogen, 15846582), 0.5 µl RNAseOUT (Invitrogen, LS10777019) and 0.5 µl SuperscriptII (Invitrogen, 18064014) and incubating the mixture at 25 °C for 10 min followed by 1 h at 42 °C. Single stranded DNA is purified through incubation with 0.5 µl RNAseA (Thermo Fisher, EN0531) for 30 min at 37 °C and PCR amplification to add the Illumina smallRNA barcodes and handles by adding 25 µl of NEBNext Ultra II Q5 Master Mix (NEB, M0492L), 11 µl Nuclease free water and 2 µl of RP1 and RPIx primers (10 µM). PCR cycles depended on the abundance of the histone modification assayed (8-10 for $H_3K_9me_3$ and $H_3K_{27}me_3$ 10-12 for $H_3K_4me_1$ and $H_3K_4me_3$). Abundance and quality of the final library are assessed by QUBIT and bioanalyzer.

Data preprocessing

Fastq files were demultiplexed by matching to an 8 nt cell barcode found in read 1 (R1). The 3 nt UMI was placed into the fastq header. To every read pair a MNase cut site is assigned to a genomic location. The cut site is defined as the genomic mapping location of the second base in R1. The ligation motif is defined as the two bases flanking the MNase cut site.

Assignment of read pairs to molecules is performed by pooling all read pairs that share the same UMI, cell barcode, and MNase cut site in a window of 1 kb.

We discarded read pairs if reads have:

- mapping quality scores (MAPQ) below 40,
- alternative hits at a non-alternative locus,
- mapped to separate locations beyond the expected insert size range,
- soft clips,
- more than 2 bases that differed from the reference,
- indels,
- mapping to a blacklist region (`http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/`).

We selected cells with more than 500 total unique cuts for $H_3K_4me_1$ and $H_3K_4me_3$, and more than 1000 total unique cuts for H3K27me and $H_3K_9me_3$. Cells also needed to have more than 50% of their cuts occur in an "AT" context. We also counted cut fragments that map in 50 kb nonoverlapping bins genome-wide, and calculated the fraction of bins that contains exactly zero cuts. Cells with a small fraction of zero cuts relative to other cells are more likely to have unspecific cuts. For each mark, we removed cells with a fraction of zero cuts that was below 2 standard deviations from the mean across all cells.

More details on the preprocessing pipeline can be found in the wiki page pipeline:

https://github.com/BuysDB/SingleCellMultiOmics/wiki.

Calculating reads falling in peaks in sortChIC for K562 cells

For each histone modification, we merged K562 single-cell sortChIC data, and used the resulting pseudobulk as input for *hiddenDomains* [166], with minimum peak length of 1000 bp. We estimated 40574, 58257, 28499, and 28380 peaks for $H_3K_4me_1$, $H_3K_4me_3$, $H_3K_{27}me_3$, and $H_3K_9me_3$, respectively. For each histone modification, we counted the fraction of total reads that fall within each set of peaks.

Dimensionality reduction based on multinomial models

We counted the number of cuts mapped to peaks across cells and applied the Latent Dirichlet allocation (LDA) model [27], which is a matrix factorization method that models discrete counts across predefined regions as a multinomial mixture model. LDA can be thought of as a discrete version of principal component analysis (PCA), replacing the normal likelihood with a multinomial one [35]. LDA models the genomic distribution of cuts from a single cell using a hierarchy of multinomials:

1. $\vec{V}_k$Dirichlet(δ) to specify the distribution over genomic regions for each topic *k* (length G genomic regions).

2. $\vec{U}_i$*Dirichlet* (α) to specify the distribution over topics for a cell *i* (length K topics).

To generate the genomic location of the *j*th read in cell *i*:

1. Choose a topic $z_{i,j}$Multinomial $\left(\vec{U}_j, 1\right)$

2. Choose a genomic region $w_{i,j}$Multinomial $\left(\vec{V}_{z_{i,j}}, 1\right)$

We used the LDA model implemented by the *topicmodels* R package [81], to infer the cell-to-topic matrix (analogous to the scores matrix in PCA) and topic-to-region matrix (analogous to the loadings matrix in PCA) using Gibbs sampling with hyperparameters α =50/K, δ =0.1, where *K* is the number of topics. We used K=30 topics for all of our analyses.

Defining eight sets of blood cell type-specific genes for cell typing

We defined cell type-specific genes for cell type calling by counting reads at +/- 5 kb centered at annotated transcription start sites (TSS). We applied LDA to the resulting count matrix for $H_3K_4me_1$, $H_3K_4me_3$, and $H_3K_{27}me_3$ (we ignored $H_3K_9me_3$ here because $H_3K_9me_3$ marks mostly AT-rich, gene-poor regions). We found eight topics that defined the eight cell types in the data. For each topic, we took the top 150 TSS loadings to make eight sets of genes defining the cell types in the data. To compare this set of TSSs to publicly available scRNA-seq data [72], we took each TSS and assigned it to the corresponding gene.

Defining genomic regions for dimensionality reduction

We initially defined regions based on 50 kb windows genome wide, applying LDA, and using the Louvain method to define clusters to merge single-cell bam files. These merged bam files were then used to call significantly marked regions using *hiddenDomains* [166] with minimum bin size of 1 kb. We merged the regions across clusters and generated a new count matrix using the *hiddenDomains* peaks as features. This new count matrix was used as input for dimensionality reduction.

Batch correction in dimensionality reduction

Initial LDA of the count matrix revealed batch effects in $H_3K_4me_1$ and $H_3K_9me_3$ between cell types of plates that contained only one sorted type (i.e., entire plate was either unenriched, lineage-negative, or LSK cells, referred to as “single-type”) and cell types from plates that contained a mixture of unenriched and non-mature cells (referred to as “balanced”). We corrected batch effects in $H_3K_4me_1$, $H_3K_4me_3$, and $H_3K_9me_3$. Since $H_3K_{27}me_3$ did not have single-type plates, we did not correct batch effects in $H_3K_{27}me_3$. We considered balanced plates as the reference for differences between cell types, and corrected deviations in single-type plates to match the balanced plates. We used the imputed sortChIC-seq signal inferred from LDA as a denoised signal Y for each genomic region g for every cell c:

$$Y_{g,c} = log_2\left(\sum_{k=1}^{K} V_{g,k}U_{k,c}\right)$$

We modeled the cell type-specific batch effect using a linear model for each genomic region. The model infers the effect of a cell *c* belonging to batch *b* and cell type *d*:

$$Y_c\left(b,d\right) = \beta_0 + \beta_1 1_s\left(b\right) + \sum_{j=1}^{J} \beta_{2,j} 1_j\left(d\right) + \beta_{3,j}\left(1_j\left(d\right)\cdot 1_s\left(b\right)\right) + \epsilon$$

Where:

$1_s\left(b\right)$ is an indicator variable equal to 1 if the cell is from a single-type plate (batch *s*), otherwise 0.

$1_j\left(d\right)$is equal to 1 if cell belongs to cell type *j*, otherwise 0.

β_0 is the intercept of the model.

β_1 is the global effect[1] from a cell being from batch *s* (single-type plate).

$\beta_{2,j}$ is the effect from a cell belonging to cell type *j*.

$\beta_{3,j}$ is the interaction effect from a cell belonging to cell type *j* and being from batch *s*.

ϵ is Gaussian noise.

We inferred the effects for each genomic region using $\mathrm{lm}()$ in R with the formula syntax:

$$Y1 + \text{batch} + \text{celltype} + \text{batch}\colon \text{celltype}$$

and estimated the batch-corrected signal:

$$\tilde{Y} = \begin{cases} Y & \text{if cell from complete plate} \\ Y - \beta_1 - \beta_{3,j} & \text{if cell from single-type plate} \end{cases}$$

For cells that belong to complete plates, $1_s\left(b\right) = 0$. Therefore, this batch-correction only corrects signal from cells belonging to single-type plates. In the bone marrow analysis this corresponds to nine plates for $H_3K_4me_1$, H_3K_4-me_3, and $H_3K_9me_3$.

---

[1] (i.e., independent of cell type)

The corrected signal is used to refine the cell-to-topic and topic-to-region matrix by GLM-PCA [171]. We applied SVD to the batch-corrected matrix to use as initializations for U and V, and included batch ID as cell-specific covariates (in glmpca R package: glmpca(), with fam="poi", minibatch="stochastic", optimizer="avagrad", niterations = 500). The batch-corrected U matrix is then visualized with uniform manifold approximation and projection (UMAP).

Differential histone mark levels analysis

To calculate the fold change in histone mark levels at a genomic region between a cell type versus HSPCs, we modeled the discrete counts *Y* across cells as a Poisson regression. We fitted a null model, which is independent of cell type, and a full model, which depends on the cell type and compared their deviances to predict whether a region was "changing" or "dynamic" across cell types.

We used the glm() implementation in R with the formula syntax for the full and null model:

$Fullmodel\colon counts 1 + \text{batch} + \text{celltype} + \text{offset}(\log(\text{totalcounts}))$
$Nullmodel\colon counts 1 + \text{batch} + \text{offset}(\log(\text{totalcounts}))$.

We used *G* as a deviance test statistic:
$G = D_{full} - D_{null},$

where the deviance is two times the log-likelihood, which for Poisson is:
$D = 2\sum_{i=1}^{n}\left\{Y_i log\left(Y_i/\mu_i\right) - \left(Y_i - \mu_i\right)\right\}$

For the full model, the logarithm of the expected value μ is:
$log\left(\mu\right) = \beta_0 + \beta_1 1_s + \sum_{j=1}^{J}\beta_{2,j}1_j,$

While for the null model, it is:
$log\left(\mu\right) = \beta_0 + \beta_1 1_s,$

We fitted the model such that the estimated $\log_2$ fold change of a cell type *j*, $\frac{\hat{\beta}_{2,j}}{log(2)}$, is always relative to HSPCs.

Under the null hypothesis, *G* is chi-squared distributed with degrees of freedom equal to the difference in the number of parameters in the two models. We use this test statistic to estimate a p-value and infer whether a 50kb bin is "changing" or "dynamic" across cell types. For $H_3K_4me_1$, $H_3K_4me_3$, and $H_3K_{27}me_3$, we used a Benjamini-Hochberg adjusted p-value of $q<10^{-50}$. For

$H_3K_9me_3$, where fold changes were generally smaller, we used $q<10^{-9}$. This separate cutoff for $H_3K_9me_3$ allowed comparable number of differential bins for downstream analysis.

Defining bins above background levels for each mark

For each mark, we counted fragments falling in 50 kb bins summed across all cells. We then plotted this vector of summed counts as a histogram in log scale, which shows a bimodal distribution. We manually defined a cutoff for each mark as a background level, and took bins that were above this cutoff. This cutoff resulted in 22067, 12661, 18512, and 19881 bins for $H_3K_4me_1$, $H_3K_4me_3$, $H_3K_{27}me_3$, and $H_3K_9me_3$ respectively.

Calculating bins that change independent of cell type

We defined "changing bins" or "dynamic bins" using the deviance test statistic, detailed above in "Differential histone mark analysis", using a q-value$<10^{-50}$ for $H_3K_4me_1$, $H_3K_4me_3$, and $H_3K_{27}me_3$, and $q<10^{-9}$ for $H_3K_9me_3$. We defined these bins to be changing in a cell fate-independent manner if the estimated cell type effect, $\hat{\beta}_{2,j}$, was either greater than 0 for all cell types (gained relative to HSPCs) or less than 0 for all cell types (lost relative to HSPCs).

Predicting Activities of Transcription Factors in Single Cells

We adapted MARA (Motif Activity Response Analysis) described in [9] to accommodate the sortChIC data. Briefly, we model the log imputed sortChIC-seq signal as a linear combination of TF binding sites and activities of TF motifs:

$$\tilde{Y}_{g,c} = \sum_{m=1}^{M} N_{g,m} A_{m,c} + \epsilon$$

Where $\tilde{Y}_{g,c}$ is the batch-corrected sortChIC-seq signal in genomic region g in cell c; $N_{g,m}$ is the number of TF binding sites in region g for TF motif m; $A_{m,c}$ is the activity of TF motif m in cell c; ϵ is Gaussian noise.

The single-cell motif activity, $A_{m,c}$, is then overlaid onto the UMAP to show cell type-specific activities.

For H3K4me1, we defined genomic regions based on peak calling from *hiddenDomains*. For repressive marks, where domains can be larger, we used 50 kb bins that were significantly changing across cell types as genomic regions.

Creating the TF binding site matrix

We predicted the TF binding site count occurrence under each peak using the mm10 Swiss Regulon database of 680 motifs. We used the Motevo method to predict transcription factor binding sites. Posterior probabilities < 0.1 are rounded down to zero.

Joint H3K4me1 and H3K9me3 analysis by double incubation

To simultaneously infer the H3K4me1 and H3K9me3 cluster from single-cell double-incubated cuts, we focused on regions that were most informative to distinguish between clusters in H3K4me1 and in H3K9me3. For H3K9me3, we used 6085 statistically significant changing bins ($q<10^{-9}$, Poisson regression). For H3K4me1, we used regions near cell type-specific genes that were used to determine cell types from the data (811 regions). Since H3K4me1 had strong signal at both the TSS and gene bodies, we defined regions for each gene from transcription start site (TSS) to either its end site or 50 kb downstream of the TSS, whichever is smaller. We counted cuts mapped to these 6896 regions for H3K4me1, H3K9me3, as well as H3K4me1+H3K9me3 cells.

For a single cell, we assumed that the vector of H3K4me1+H3K9me3 counts $\vec{y}$ was generated by drawing *N* reads from a mixture of two multinomials, one from a cell type *c* from H3K4me1 (parametrized by relative frequencies $\vec{p_c}$) and one from a lineage *l* from H3K9me3 (parametrized by relative frequencies $\vec{q_l}$):

$$\vec{y} \vee c, l, w \sim \textit{Multinomial}\left(w\vec{p_c} + (1-w)\,\vec{q_l}, N\right),$$

where *w* is the fraction of H3K4me1 that was mixed with H3K9me3.

Genomic region probabilities $\vec{p_c}$ and $\vec{q_l}$ were inferred by the single-incubated data by averaging the imputed signal across cell types:

$$q_{l,g} = \frac{1}{D_l \vee \sum_{d=1}^{D_l \vee \sum_{k=1}^{K} V_{g,k} U_{k,d}}}$$

where D_l is the set of cells that belong to lineage l. V and U are estimated from LDA.

The log-likelihood for the $H_3K_4me_1$+$H_3K_9me_3$ counts coming from cluster pair (*c, l*), can be defined as:

$$_{(c,l)} \propto \sum_{g=1}^{G} y_g log\left(w p_{c,g}\left(1-w\right) q_{l,g}\right),$$

where g is a genomic region.

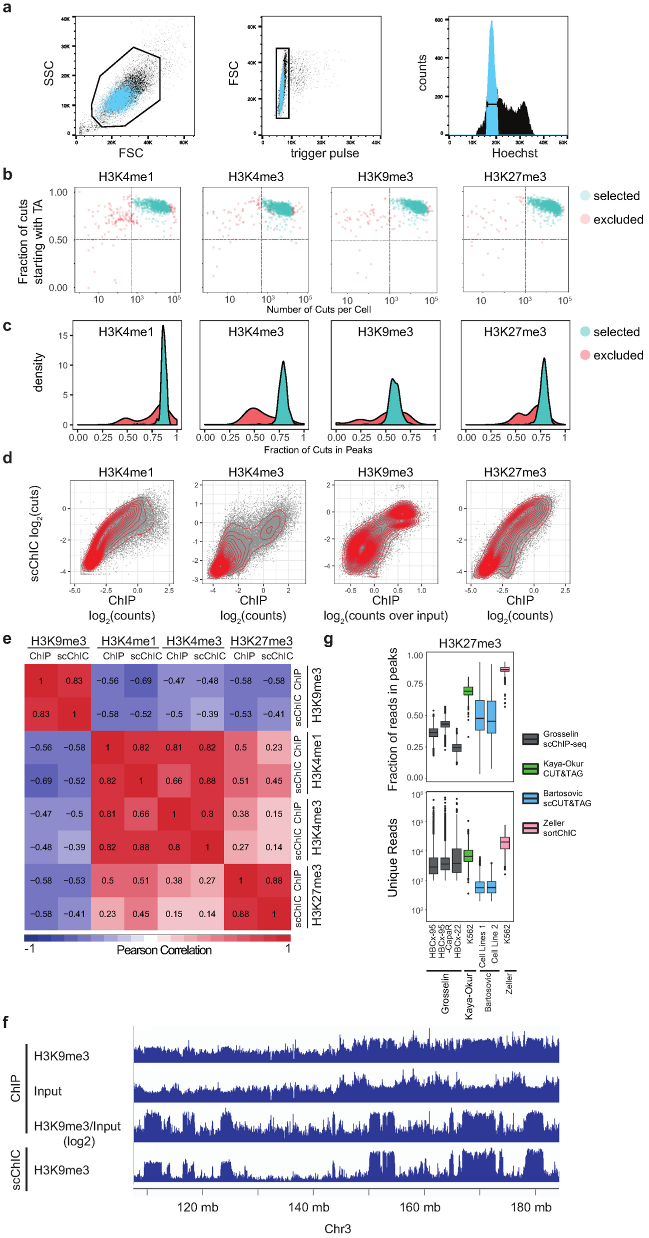
To assign a cluster pair to a double-incubated single cell, we calculated the log-likelihood for each possible pair (we had four lineages from $H_3K_9me_3$ and eight clusters from $H_3K_4me_1$, creating a 32 possible pairs) and selected the pair with the highest log-likelihood. We used the Brent method implemented in R (*optim*) to infer *w* that maximizes the log-likelihood for each pair.

Materials

Antibodies
$H_3K_4me_1$, ab8895 (Abcam), Lot: GR3206285-1
$H_3K_4me_3$, 07-473 (Merck), Lot: 3093304
$H_3K_9me_3$, ab8898 (Abcam), Lot: GR3217826-1
$H_3K_{27}me_3$, 9733S (NEB), monoclonal
Public K562 data
$H_3K_4me_1$, Peggy Farnham, ENCSR000EWC, pAb-037-050 (Diagenode)
$H_3K_4me_3$, Peggy Farnheim, ENCSR000EWA, 9751S (Cell Signaling)
$H_3K_9me_3$, Bradley Bernstein, ENCSR000APE, ab8898 (Abcam)
$H_3K_{27}me_3$, Peggy Farnheim, ENCSR000EWB, 9733S (Cell Signaling)
Public bone marrow scRNA-seq: Pseudobulk estimates merged from scRNA-seq data from Giladi et al 2018 [72]

3.6 Supplementary figures

Supplementary Figure 1

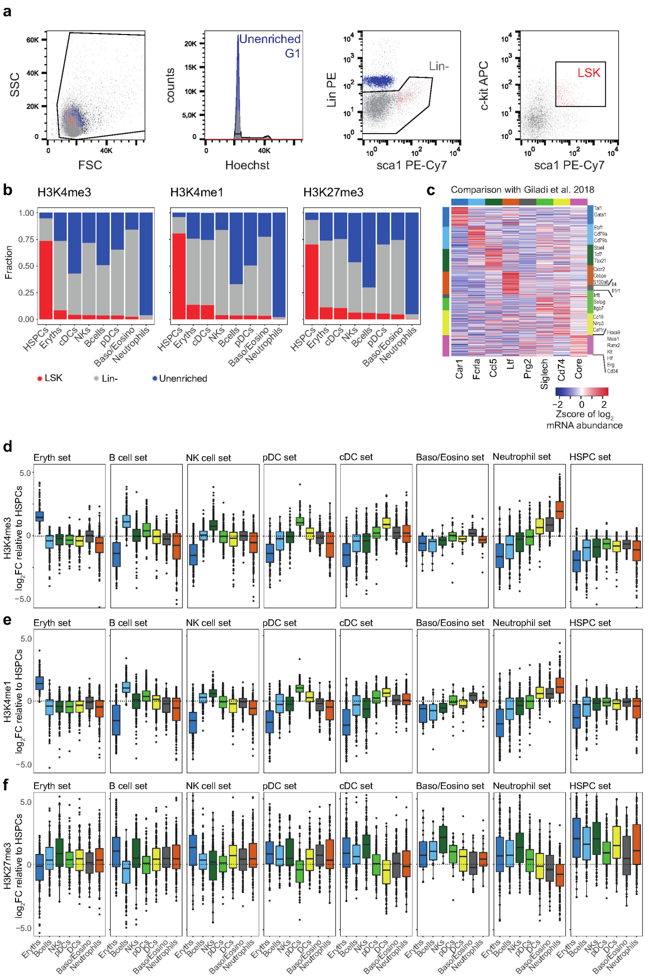


		H3K9me3 ChIP	H3K9me3 scChIC	H3K4me1 ChIP	H3K4me1 scChIC	H3K4me3 ChIP	H3K4me3 scChIC	H3K27me3 ChIP	H3K27me3 scChIC
H3K9me3	ChIP	1	0.83	−0.56	−0.69	−0.47	−0.48	−0.58	−0.58
	scChIC	0.83	1	−0.58	−0.52	−0.5	−0.39	−0.53	−0.41
H3K4me1	ChIP	−0.56	−0.58	1	0.82	0.81	0.82	0.5	0.23
	scChIC	−0.69	−0.52	0.82	1	0.66	0.88	0.51	0.45
H3K4me3	ChIP	−0.47	−0.5	0.81	0.66	1	0.8	0.38	0.15
	scChIC	−0.48	−0.39	0.82	0.88	0.8	1	0.27	0.14
H3K27me3	ChIP	−0.58	−0.53	0.5	0.51	0.38	0.27	1	0.88
	scChIC	−0.58	−0.41	0.23	0.45	0.15	0.14	0.88	1

(Caption on next page.)

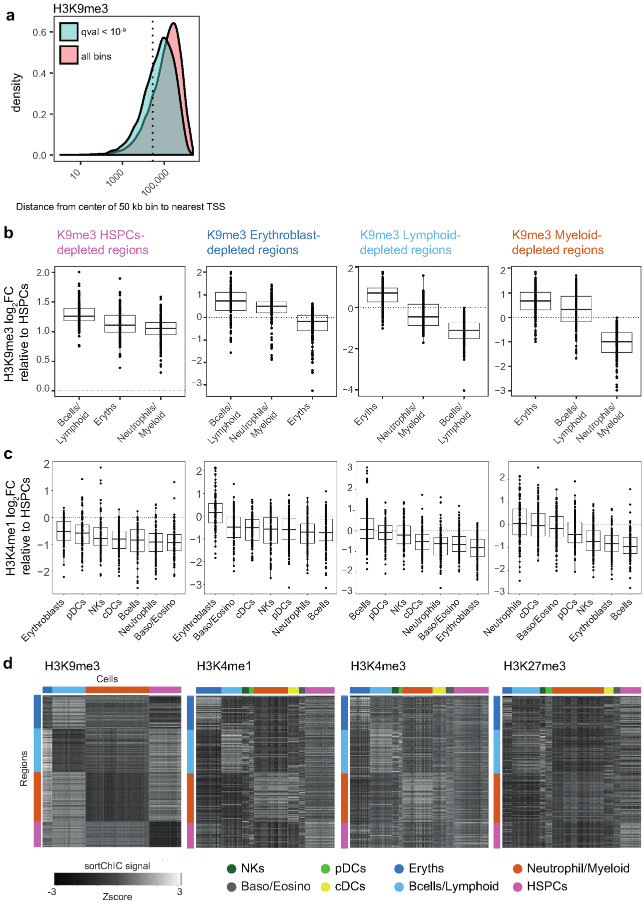
Supplementary Fig. 1: sortChIC generates high-resolution maps of histone modifications in single cells. (a) FACS plots for sorting individual K562 cells in G1 phase. (b) Fraction of cuts starting with TA (reflecting the preference of MNase to cut in an AT context) versus number of cuts mapped to the K562 genome. Cells below horizontal dotted lines and left of vertical lines are excluded from the analysis. (c) Distribution of fraction of cuts mapped to locations within peaks across cells. (d) Correlation between pseudobulk sortChIC and bulk ChIP signal using 50 kilobase (kb) bins for $H_3K_4me_1$, $H_3K_4me_3$, H_3-$K_{27}me_3$, and $H_3K_9me_3$. (e) Pearson correlation between pseudobulk sortChIC and bulk ChIP signal using 50 kb bins across the four histone marks. (f) Three tracks of $H_3K_9me_3$ ChIP-seq bulk data, one for $H_3K_9me_3$ without normalization ($H_3K_9me_3$), one for the input (Input), and one where $H_3K_9me_3$ is normalized to the input ($H_3K_9me_3$/input). Fourth track is $H_3K_9me_3$ sortChIC pseudobulk, showing that $H_3K_9me_3$ ChIP-seq requires normalizing by input to resemble sortChIC. (g) Comparison of specificity (fraction of cuts in peaks, top panel) and sensitivity (number of unique reads, bottom panel) for three alternative high throughput single cell chromatin methods with sortChIC of $H_3K_{27}me_3$ in K562 cell lines. Fraction of cuts in peaks and number of unique reads was taken from comparative analysis of $H_3K_{27}me_3$ from Bartosovic et al [32]. Boxplots show 25th percentile, median and 75th percentile, with the whiskers spanning 97% of the data.l cell type information, while dynamics in heterochromatin regions marked by $H_3K_9me_3$ reveal lineage information.

Supplementary Figure 2



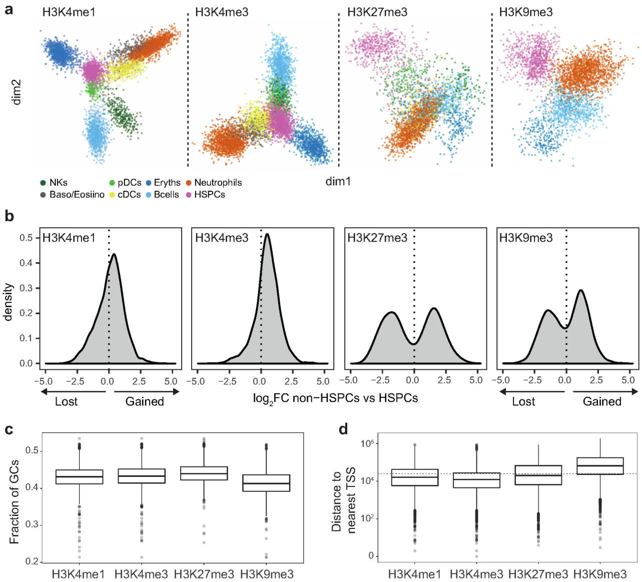
Supplementary Fig. 2: $H_3K_4me_1$ and $H_3K_4me_3$ in HSPCs prime for different blood cell fates, while $H_3K_{27}me_3$ in differentiated cell types silences genes of alternative cell fates. (a) FACS plot for sorting G1 cells of whole bone marrow (unenriched), lineage negative (Lin$^-$), and Lin$^-$,Sca1$^+$, cKit$^+$ (LSK) populations. (b) Fraction of cells in each cell type labeled by the sorted population: whole bone marrow (unenriched), lineage negative (Lin$^-$), and Lin$^-$Sca1$^+$cKit$^+$ (LSK). (c) Cell type-specific mRNA abundances for genes associated with regions in Fig. 2E using pseudobulk analysis of the Giladi et al. 2018 dataset (Methods). (d) $H_3K_4me_3$ fold changes of different cell types relative to HSPCs at cell type-specific regions. Each panel corresponds to a set of cell type-specific regions defined by the rows of one color in the heatmap of Fig. 2e. Regions are defined by +/- 5 kilobase windows centered at transcription start sites of cell type-specific genes. (e) Same as (d) but for $H_3K_4me_1$. (f) Same as (d) but for $H_3K_{27}me_3$. Boxplots show 25th percentile, median and 75th percentile, with the whiskers spanning 97% of the data.

Supplementary Figure 3



Supplementary Fig. 3: Lineage-specific loss of H3K9me3 correlates with cell type-specific increase in H3K4me1. (a) Statistically significant 50 kb regions (adjusted p-value $< 10^9$, deviance goodness-of-fit test) identified for H3K9me3, showing distribution of distances from center of 50 kb region to nearest gene. All bins are identified as 50 kb regions that have pseudobulk (counts summed across all cells) signal above background levels (Methods). Dotted line represents 25 kb, meaning the bin would overlap with a TSS. (b) Fold change in H3K9me3 relative to HSPCs for four sets of 150 regions: regions depleted in erythroblasts, lymphoid, myeloid, or HSPCs. Each region is 50 kb wide. (c) The same four sets of regions but showing fold change in H3K4me1, showing upregulation of H3K4me1 specifically in cell types that are depleted in H3-K9me3. Boxplots show 25th percentile, median and 75th percentile, with the whiskers spanning 97% of the data. (d) Heatmap of the four regions in single cells across the four marks. Rows are regions, color coded as in top of (b). Columns are cells, color coded as in bottom.

Supplementary Figure 4



e

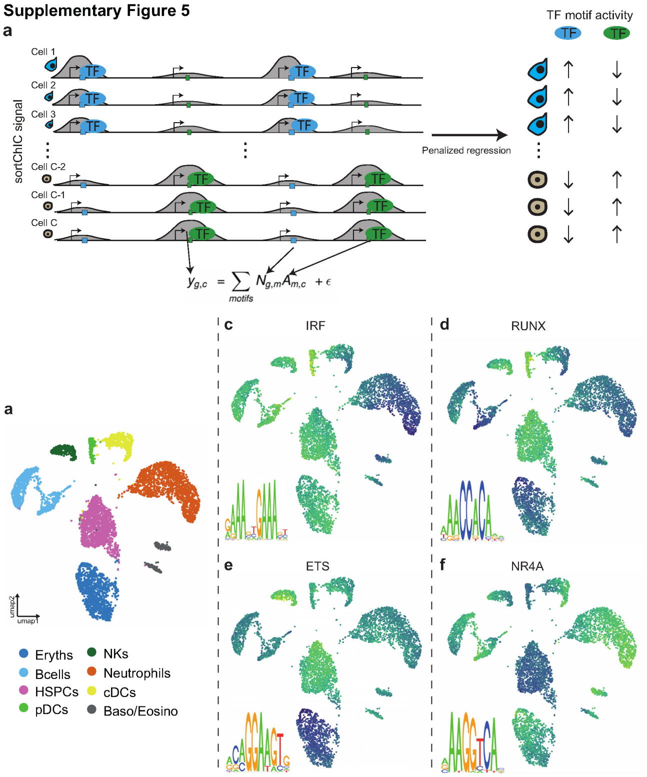
H3K9me3 regions unique to HSPCs

GO biological process	GO number	P-value	Enrichment (log2)
complement activation, classical pathway	GO:0006958	5.68E-29	5.91
phagocytosis, recognition	GO:0006910	1.86E-28	6.06
B cell receptor signaling pathway	GO:0050853	3.58E-27	5.66
phagocytosis, engulfment	GO:0006911	1.25E-26	5.46
positive regulation of B cell activation	GO:0050871	4.69E-25	4.86
defense response to bacterium	GO:0042742	2.33E-19	3.26
innate immune response	GO:0045087	5.08E-11	2.24
immunoglobulin production	GO:0002377	8.09E-08	3.13

H3K27me3 regions unique to HSPCs

GO biological process	GO number	P-value	Enrichment (log2)
locomotory behavior	GO:0007626	9.29E-09	3.23
central nervous system neuron differentiation	GO:0021953	5.14E-07	3.19
regulation of nervous system process	GO:0031644	1.74E-06	3.31
positive regulation of neuron differentiation	GO:0045666	2.10E-05	3.88
developmental growth involved in morphogenesis	GO:0060560	4.64E-04	3.33
neuromuscular process	GO:0050905	3.76E-04	3.28
memory	GO:0007613	2.81E-04	3.19
sensory perception of sound	GO:0007605	1.37E-04	3.16

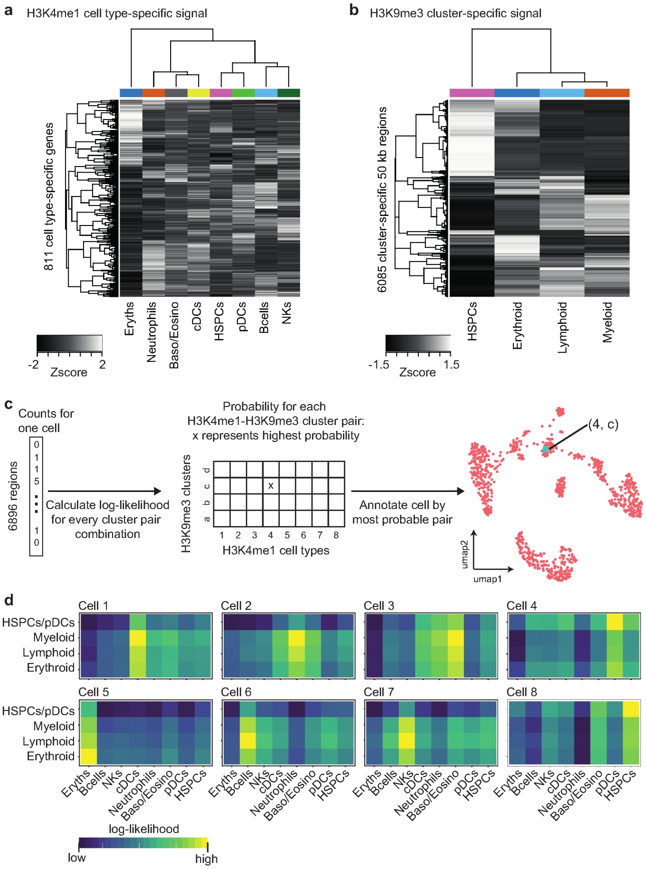
Supplementary Fig. 4: Features of active and repressive chromatin dynamics during hematopoiesis. (a) Dimensionality reduction from GLMPCA (Methods) showing the two main latent factors explaining the sortChIC data for each mark. Dim 1 contains 8.2%, 8%, 7%, and 9.5% of the total L2 norm for $H_3K_4me_1$, $H_3K_4me_3$, $H_3K_{27}me_3$, and $H_3K_9me_3$, respectively. Dim 2 contains 7.9%, 6.8%, 6.8%, and 6.8% of the total L2 norm. (b) Distribution of $\log_2$ fold changes (FC) at statistically significant changing bins (null model: a bin has constant signal across all cell types, full model: a bin has signal that depends on cell type, deviance goodness-of-fit test) between pseudobulk of non-HSPCs versus HSPCs. Bimodal distribution highlights differences originate mainly between HSPCs and non-HSPCs. (c) GC content of dynamic 50 kb bins for the four histone marks. (d) Distance to nearest TSS measured from the center of each dynamic 50 kb bin. Dotted horizontal line represents 25 kb, meaning the bin would overlap with a TSS. Boxplots show 25th percentile, median and 75th percentile, with the whiskers spanning 97% of the data. (e) Gene ontology (GO) terms of HSPC-specific $H_3K_9me_3$ (top) and $H_3K_{27}me_3$ (bottom) regions. P-value and enrichment from Fisher's exact test.



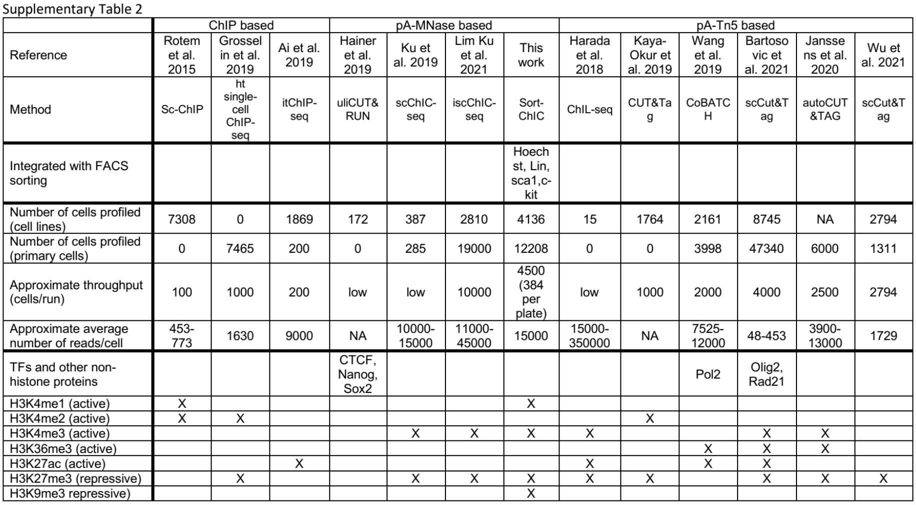
Supplementary Fig. 5: Penalized regression model reveals transcription factor motifs underlying cell type-specific chromatin dynamics.

(a) Schematic of the transcription factor (TF) activity model. The penalized regression model takes the imputed sortChIC signal in a peak as the response variable and the TFbinding motifs predicted under each peak as the explanatory variable (Method). The penalized multivariate regression infers the TF motif activity driving cell type-specific sortChIC signal. (b) UMAP of H$_3$K$_4$-me$_1$ chromatin states in single cells, colored by cell type. (c-f) UMAP where each cell is colored by the TF activity inferred from the model. Four cell type-specific TF motifs are shown.

Supplementary Figure 6



Supplementary Fig. 6: Single-incubated data from $H_3K_4me_1$ and $H_3K_9me_3$ builds a model for inferring cluster-pairs in double-incubated data. (a) Heatmap of $H_3K_4me_1$ signal across clusters for 811 cell type-specific regions (Methods). These regions come from cell type-specific genes used in Fig. 2e. (b) Heatmap of $H_3K_9me_3$ signal across clusters for 6085 cluster-specific regions (50 kb genomic window). These regions come from the statistically significantly dynamic regions of $H_3K_9me_3$ defined in Supplementary Fig. 3A. (c) Schematic of how a cluster-pair is inferred from each double-incubated cell. Each double-incubated cell has a vector of counts across 6896 regions (811 regions come from $H_3K_4me_1$, while 6085 come from $H_3K_9me_3$). We calculate the log-likelihood (Methods) of the observed double-incubated cell counts for each cluster-pair (32 cluster-pairs from 8 clusters in $H_3K_4me_1$ and 4 clusters in $H_3K_9me_3$). From the 32 log-likelihoods estimates, we assign the cell to the cluster-pair with the highest probability. (d) Examples of the 32 log-likelihood estimates from eight representative cells, shown as a 4-by-8 heatmap. Each of the four rows is a cluster from $H_3K_9me_3$; each of the eight columns is a cluster from $H_3K_4me_1$.



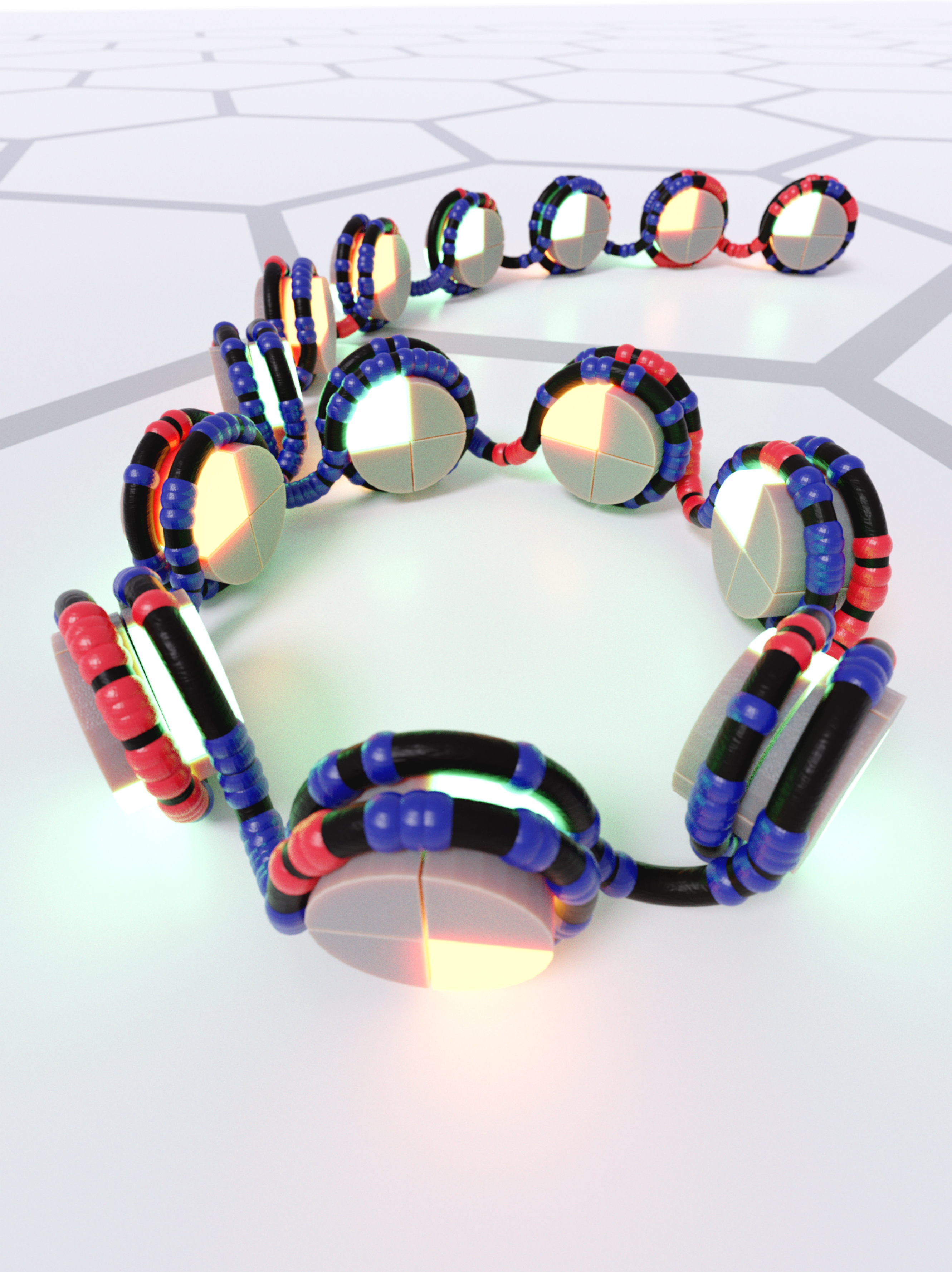
Supplementary Table 2

	ChIP based			pA-MNase based				pA-Tn5 based					
Reference	Rotem et al. 2015	Grosselin et al. 2019	Ai et al. 2019	Hainer et al. 2019	Ku et al. 2019	Lim Ku et al. 2021	This work	Harada et al. 2018	Kaya-Okur et al. 2019	Wang et al. 2019	Bartosovic et al. 2021	Janssens et al. 2020	Wu et al. 2021
Method	Sc-ChIP	ht single-cell ChIP-seq	itChIP-seq	uliCUT&RUN	scChIC-seq	iscChIC-seq	Sort-ChIC	ChIL-seq	CUT&Tag	CoBATCH	scCut&Tag	autoCUT&TAG	scCut&Tag
Integrated with FACS sorting							Hoechst, Lin, sca1,c-kit						
Number of cells profiled (cell lines)	7308	0	1869	172	387	2810	4136	15	1764	2161	8745	NA	2794
Number of cells profiled (primary cells)	0	7465	200	0	285	19000	12208	0	0	3998	47340	6000	1311
Approximate throughput (cells/run)	100	1000	200	low	low	10000	4500 (384 per plate)	low	1000	2000	4000	2500	2794
Approximate average number of reads/cell	453-773	1630	9000	NA	10000-15000	11000-45000	15000	15000-350000	NA	7525-12000	48-453	3900-13000	1729
TFs and other non-histone proteins				CTCF, Nanog, Sox2						Pol2	Olig2, Rad21		
H3K4me1 (active)	X						X						
H3K4me2 (active)	X	X							X				
H3K4me3 (active)					X	X	X	X			X	X	
H3K36me3 (active)										X	X	X	
H3K27ac (active)			X					X		X	X		
H3K27me3 (repressive)		X			X	X	X	X	X		X	X	X
H3K9me3 repressive)							X						

Supplementary Table 2: Comparison of studies on single-cell histone modification mapping.

Supplementary Table 1: Fold change estimates relative to HSPCs for different cell types. Estimates of $\log_2$ fold change between a cell type relative to HSPCs for $H_3K_4me_1$, $H_3K_4me_3$, $H_3K_{27}me_3$, and $H_3K_9me_3$ (one tab for each mark). P-values estimated from deviance goodness-of-fit test from Poisson regression. (Omitted, too large for print, can be found on the bioRxiv preprint)

Supplementary Table 3: List of barcode adaptors used in this study. (Omitted, too large for print, can be found on the bioRxiv preprint)



Chapter 4

scChIC-TAPS reveals histone modification specific DNA methylation dynamics during the cell cycle

In preparation

Christoph Geisenberger[1,*], Buys Anton de Barbanson[1,*], Jeroen de Ridder[2] and Alexander van Oudenaarden[1]

[1]Oncode Institute, Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, The Netherlands.

[2]Oncode Institute, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands.

*These authors contributed equally

4.1 Abstract

Epigenetic mechanisms regulate transcriptional output and play a vital role in cell fate decisions. While multi-omics techniques have started to provide insights into how these processes are coordinated, single-cell measurements are required to delineate correlations between epigenetic marks at the level of individual genes. In this work, we propose scChIC-TAPS , a novel method which simultaneously measures post-translational histone modifications and DNA methylation at the single-cell level. Our approach combines bisulfite-free conversion of methylated cytosines and targeted MNase digestion and resolves the local correlations of different histone modifications and DNA methylation states at base-pair resolution. Applying scChIC-TAPS in a Fucci reporter line

allows us to track epigenetic changes and their genome-wide interplay across the full cell cycle. Our data provides the first direct evidence that kinetics of replication-coupled maintenance methylation are influenced by the local chromatin environment.

4.2 Introduction

Eukaryotes package their DNA into chromatin to coordinate a number of crucial processes such as transcription, DNA repair and silencing of repetitive elements. Two of the most well-studied epigenetic marks involved in this regulation are post-translational histone modifications and the methylation of cytosine bases. Deregulation of epigenetic processes is commonly observed in a number of diseases, most notably in cancer.

In order to guarantee stable phenotypes over time, intricate mechanisms are involved in the inheritance of histone marks and DNA methylation. During every replication of the genome, old histones are disassembled in front of the replication fork and re-integrated into the leading and lagging strands in roughly equal proportions. At the same time, DNMT1 travels with the replicative machinery and copies methylation patterns from the template strand onto the newly synthesized strand. This maintenance methylation occurs in the context of symmetrical CpG dinucleotides. After DNA synthesis, methylated CpGs in the template strand are paired with unmodified CpGs in the nascent strand. DNMT1 recognized these hemimethylated sites, thereby providing a mechanistic basis for methylation inheritance.

There is sufficient evidence to suggest that the enzymatic properties of DNMT1 are reflected in genome-wide methylation patterns. For example, in-vitro experiments show that average DNA methylation is correlated with DNMT1s flanking site preference [2]. However, there is little understanding of how DNMT1 activity might be impacted by the local chromatin environment. DNMT1 itself is recruited by a number of different mechanisms, including the recognition of ubiquitin marks on histone H3, a mark placed by its partner UHRF1. Since methyltransferases cannot access nucleosome-bound DNA, it has been proposed that nucleosome insertion and DNMT1 accessibility might compete in newly synthesized regions of the genome [133]. In addition, modeling of kinetic rates of re-methylation after DNA replication has revealed large variability [37]. This suggests that the local chromatin environment might impact the type and timing of methylation maintenance [133]. Further evidence in this direction comes from the discovery of *Partially Methylated Domains* (PMDs). These large regions tend to become hypomethylated in cancer and aging and were found to co-localize with repressive, H3K9me3-marked chromatin [33, 146].

However, most evidence is correlative and there is no method to measure the epigenetic interplay in a time-resolved manner. Here, we address this shortcoming and present scChIC-TAPS , which allows read-out of histone modifications and DNA methylation from the same molecule at base-pair resolution in single cells. In addition to a thorough technical validation, we use our novel approach to measure the kinetics of DNA methylation in different chromatin contexts across the full cell cycle.

4.3 Results

Our approach builds on single-cell chromatin cleavage (scSortChIC) [195] and is outlined in Fig. 1A. Histone modification-specific antibodies are used to target MNase and the resulting fragments are ligated with barcoded adapters. After single-cell tagging, material is converted using Tet-assisted pyridine borane sequencing (TAPS) [107]. This two-step process combines enzymatic oxidation and incubation with a chemical to convert 5mC to *dihydroxyuracil* (DHU). DHU is subsequently replaced by thymidine during amplification. Of note, due to the specific conversion of 5mC, TAPS is compatible with regular, i.e., unmethylated sequencing adapters. Illumina sequencing allows the extraction of multiple pieces of information for each read: (i) single-cell identity (barcode sequence), (ii) genomic location of histone modification (mapping position) and (iii) the methylation state of the original molecule (C to T transitions). Importantly, our data provide single-molecule in addition to single-cell resolution.

Technical validation of scChIC-TAPS in K562 cells

For technical validation, we produced scChIC-TAPS data for three different histone modifications in K562 cells (H3K9me3, H3K27me3 and H3K36me3). Mapping rates ranged from 90.5% to 98.8% (Supplementary Table 1), about four times higher than single-cell bisulfite sequencing data [45]. Cells were filtered based on the number of unique cut sites, TA fraction (MNase bias) and average methylation. Between 79 % (H3K9me3) and 92% (H3K27me3) of cells

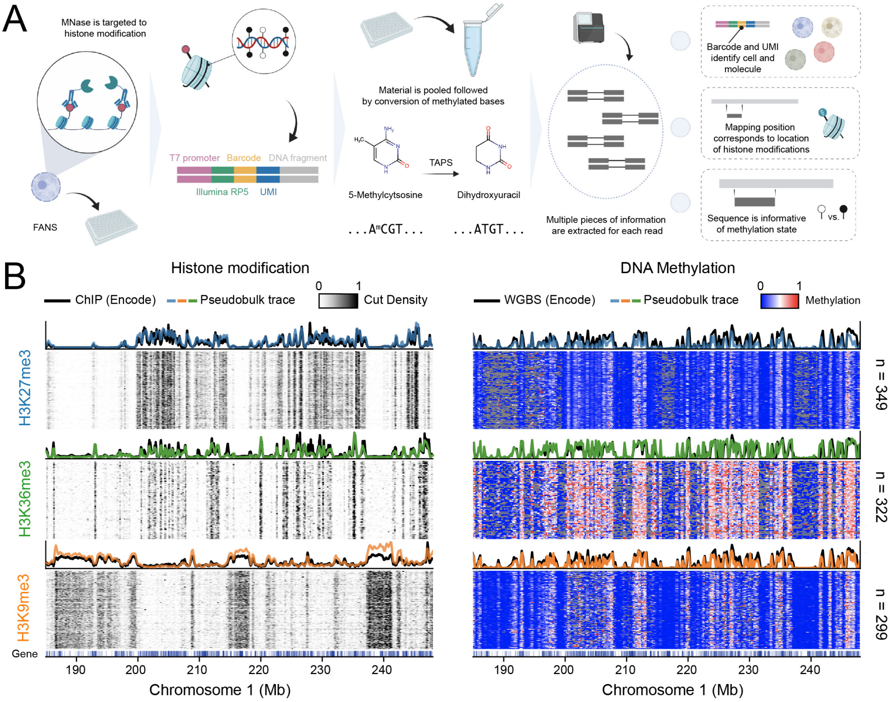


Figure 1 (A-B): scChIC-TAPS enables multiplexed profiling of histone modifications and DNA methylation in single cells A, Schematic of scChIC-TAPS . Specific antibodies coupled to MNase are targeted to the corresponding histone modifications. Single nuclei are sorted into 384-well plates and subsequent steps performed using a robotic liquid handler. First, MNase digestion is initiated through addition of Ca2+. Then, Proteinase K digestion, blunting, A-tailing and adapter ligation create barcoded fragments. Material from one plate is pooled, followed by conversion of methylated cytosines to DHU. After sequencing library amplification, DHU is replaced by thymidine (T). **B,** Heatmaps showing data for histone modifications (left) and DNA methylation (right) obtained from the same K562 single cells across a 60 Mb region on chromosome 1. Colored traces above heatmaps correspond to the averaged signal across cells and are accompanied by reference profiles (ENCODE ChIP-Seq and WGBS, respectively). Tick marks under the heatmaps indicate locations of genes.

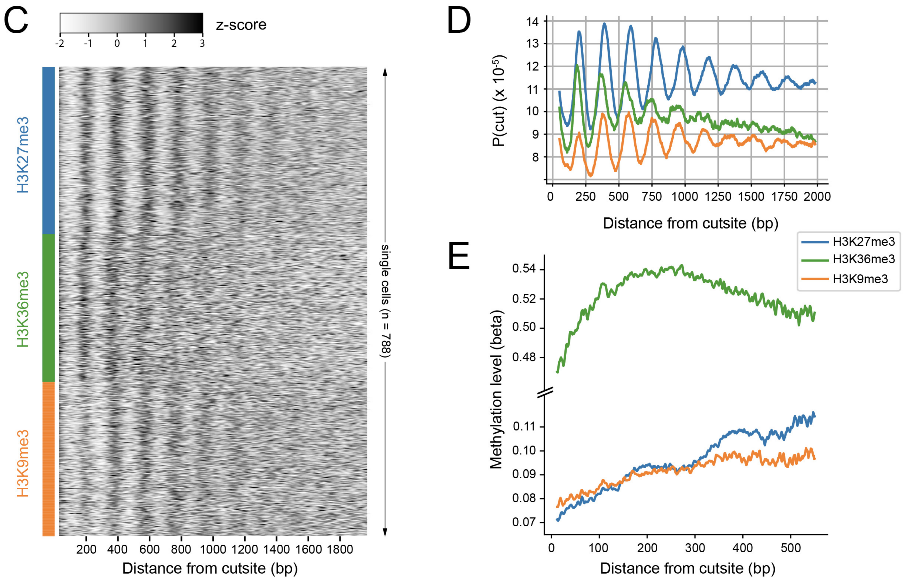


Figure 1 (C-E): scChIC-TAPS enables multiplexed profiling of histone modifications and DNA methylation in single cells C, Heatmap visualizing nucleosome spacing in single cells. Pairwise distances between cut sites (mapping position of Read 1) are calculated per single cell and z-score normalized. The striped vertical pattern corresponds to cuts located roughly one nucleosome distance away. **D,** Similar to C, this figure shows the probability of detecting another MNase cut site given a specific genomic distance aggregated across cells within the same histone modification. Oscillations show periodicities of approximately 180 to 190 bp with slight differences in peak location and signal decay between histone modifications. **E,** As described in D but taking advantage of the multiplexed nature of scChIC-TAPS , methylation levels (beta, fraction methylated) are plotted with respect to the cut site position.

passed our quality control criteria. TAPS conversion efficiency estimated based on fully methylated spike-ins ranged from 81% to 91% (Supplementary Table 1). We extracted histone and methylation profiles for the same single cells, Figure 1B shows the corresponding heatmaps for a 60 Mb region of chromosome 1. A more detailed zoom-in is provided in fig. S1. In addition to single-cell heatmaps, we aggregated signals into pseudo-bulk measurements, which we compared to bulk data published by ENCODE. The number of MNase cuts in non-overlapping genomic bins of 100 kb were compared to normalized ChIP signal, which revealed high concordance between the data sets (fig. S2, A). Genome-wide correlation (Pearsons r) ranged from 0.72 ($H_3K_{36}me_3$) to 0.78 ($H_3K_{27}me_3$), equivalent to stand-alone measurements obtained with single-cell CUT&RUN [95] and single-cell CUT&TAG [89]. Methylation values obtained by TAPS were compared to *Whole Genome Bisulfite Sequencing* (WGBS) data. For each histone mark, we assessed CpGs with at least 5x coverage in both data sets. Pseudobulk correlations were in the range of 0.98 to 0.99 (fig. S2, B), comparable to technical replicates of WGBS.

To further validate our approach, we assessed nucleosome spacing patterns resulting from MNase digestion. Figure 1C shows the relationship between cut site spacing and genomic distance as a single-cell heatmap. Figure 1 D aggregates these data per histone mark, which reveals oscillatory patterns relating to nucleosome occupancy. While phase (around 190 bp) and frequency are similar across marks supporting previously published data in K562 cells [135], spacing and signal decay show subtle differences across histone marks. Lastly, we investigated methylation values for the different histone marks. Here, average methylation within $H_3K_{27}me_3$ and $H_3K_9me_3$ fragments (8 to 10%) is much lower than compared to $H_3K_{36}me_3$ (50%). Again, these findings are in line with previously published reports on the high methylation of gene bodies [113]. Taken together, our data for K562 cells are well correlated with bulk reference data and thus accurately represent the underlying histone modification and DNA methylation landscapes. Of note, the quality of our measurements is comparable to other single-cell single-omics techniques.

Using scChIC-TAPS to measure epigenetic dynamics during the cell cycle

Next, we apply scChIC-TAPS in a setting with epigenetic dynamics. To this end, we profiled three histone modifications ($H_3K_{36}me_3$, $H_3K_{27}me_3$ and H_3K_9-me_3) paired with DNA methylation during cell cycle progression. In addition to the regulation of many genes, methylation patterns are maintained during S Phase. We therefore reasoned that our combined measurements might not only provide insights into epigenetic regulation of gene expression, but also the interplay of maintenance methylation and chromatin organization. To obtain high-resolution data, we resorted to the RPE-FUCCI reporter system. This cell line expresses an orange fluorophore during G1 phase and a green fluorophore during G2 phase, with a short window of concomitant expression of both markers in early S phase [145]. Of note, fluorophores are retained in the nucleus, which allowed the recording of FACS parameters during sorting. Cells were profiled with scChIC-TAPS in FACS space, and the cell cycle progression trajectory (pseudotime) fitted by using Wanderlust (Figure S3). The cell cycle progression trajectory is used to order cells across the cell cycle and is used as a shared manifold which allows comparing between histone modifications and previously published transcriptome data [18].

To start with, we show that $H_3K_{36}me_3$ is mainly found on expressed gene bodies, while $H_3K_{27}me_3$ is found on repressed gene bodies (S6) supporting previous research. In Figure 3 a comparison of the three histone marks across the cell cycle is shown. Briefly, the genome was divided into 10kb bins and the number of counts per cell per bin counted. Then the cell cycle was divided by using a sliding window across the estimated cell cycle progression trajectory, where each window contains 5% of the cells. This results in a matrix of genomic bin x cell cycle progression bin, which is normalized to counts per million. We concatenated the matrices of the 3 marks (Figure S5), and created an UMAP [19] from this matrix (Figure 3). The UMAP has 4 main protrusions, one for each histone mark and one for regions which do not have any of the marks. The protrusions are caused because $H_3K_{36}me_3$, $H_3K_{27}me_3$ and $H_3K_9me_3$ are mostly non-overlapping (Figure 3).

Overall, dynamics of $H_3K_{36}me_3$, $H_3K_{27}me_3$ and $H_3K_9me_3$ during the cell cycle are subtle, with the strongest changes in $H_3K_{36}me_3$ signal.

Next, we use results generated by using scRepli-seq [167] to split our 10kb genomic bins into two classes: *early* and *late* replicated. $H_3K_{36}me_3$ domains are mostly located in early replicated regions, while $H_3K_{27}me_3$ domains are found in both early and late replicated regions. Finally, most $H_3K_9me_3$ domains are late replicated Figure 2.

When the genome is being replicated, not all methylated CpGs are immediately maintained, resulting in a slight drop of mean CpG methylation. In order to study CpG maintenance dynamics near $H_3K_{36}me_3$, $H_3K_{27}me_3$ and $H_3K_9me_3$

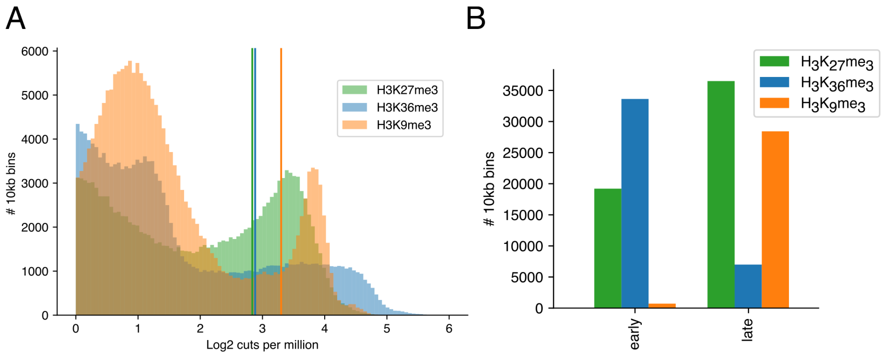


Figure 2: Replication timing of histone modified regions. A. Distribution of the amount of cuts detected per bin in counts per million. The vertical lines indicate the thresholds used to classify a region as covered by the mark. **B,** The total amount of 10kb bins with a histone covered domain which are early and late replicated. H3K36me3 domains are mostly early replicated, and H3K9me3 are mostly late replicated. H3K27me3 can be found on both early and late replicated regions.

across the cell cycle for the *early* and *late* replicated regions, we calculated the mean difference of CpG methylation for all bins over the cell cycle. We expect that regions covered by H3K36me3 are early replicated during S phase, therefore resulting in an early drop in DNA methylation. Indeed, we show that bins which are covered by H3K36me3 are early replicating and are the bins which first drop in CpG methylation, and which are also earliest maintained back to their original CpG methylation level (Figure 5). Then H3K27me3 covered domains are replicated, followed by H3K9me3 domains. A small subset of H3K9me3 domains are early replicated, but most of these early replicated H3K9me3 bins are also covered by H3K36me3 (Figure 4).

scChIC-TAPS allows to not only to calculate the mean DNA methylation level for regions across the cell cycle, but can also be used to study the dynamics at near single base resolution around nucleosomes which contain a histone modification of interest. MNase can only cut in-between nucleosomes, in the linker-DNA [124]. This means that on average the scChIC-TAPS reads cover first, a small section of linker, then cover nucleosome core-DNA, and then again linker-DNA (Figure 5 A).

We calculated the CpG methylation level for the 350bp window near the

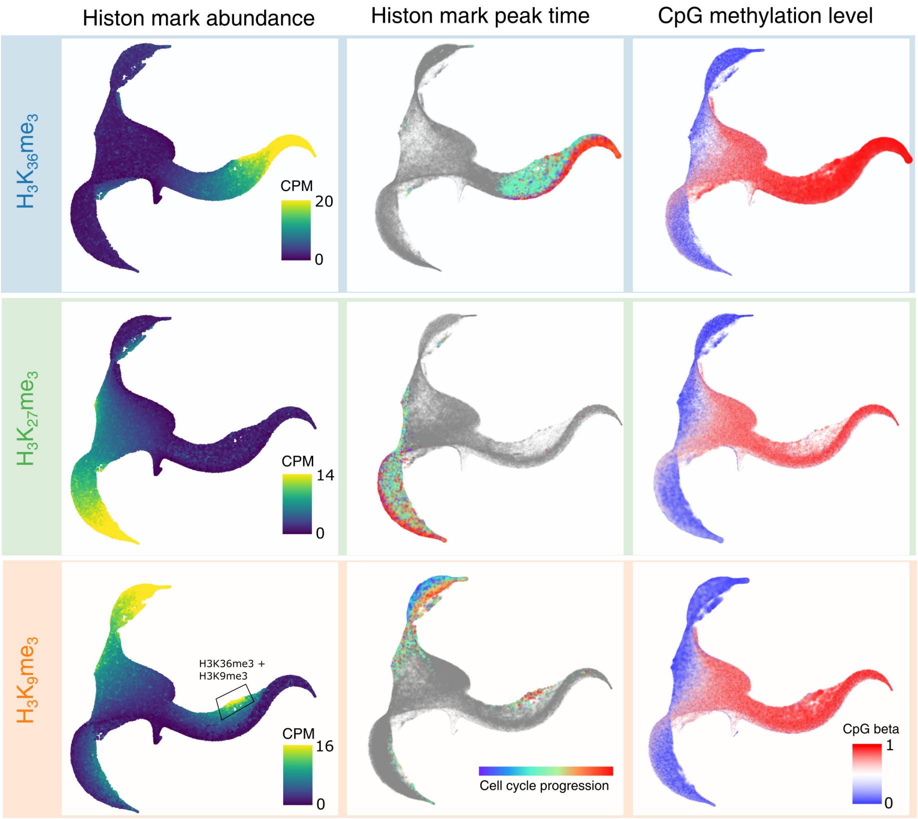


Figure 3: UMAP of 10kb bins, resampled along the cell cycle progression trajectory. Each dot represents a genomic location of 10kb, the location of the dot represents the density of $H_3K_{36}me_3$, $H_3K_{27}me_3$ and $H_3K_9me_3$ and at a more fine scale the dynamics of these marks across the cell cycle. In the first column the mean amount of detected MNase cuts in counts per million across the cycle is shown for $H_3K_{36}me_3$, $H_3K_{27}me_3$ and $H_3K_9me_3$. This indicates little overlap between the marks, a small set of bins contains both $H_3K_{36}me_3$ and $H_3K_9me_3$. In the second column the same UMAP is shown, but colors indicate the cell cycle progression time where the most counts are detected. Bins marked in grey do not contain enough counts to accurately determine the peak time. The last column shows the mean CpG methylation level detected for each bin.

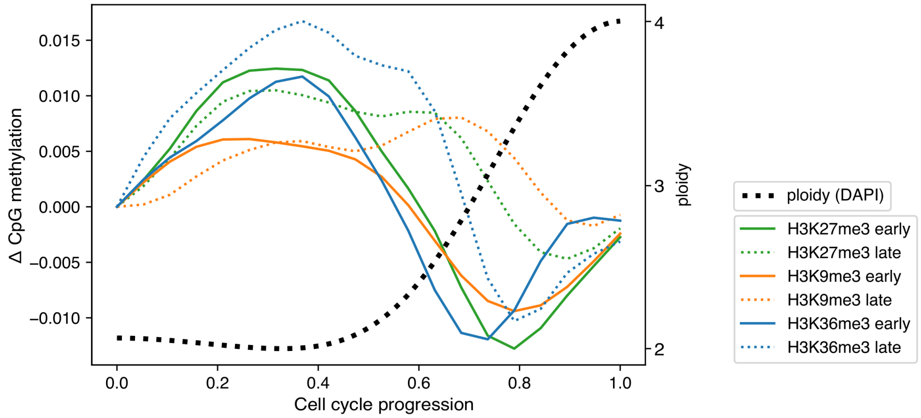


Figure 4: Dynamics of CpG methylation near $H_3K_{36}me_3$, $H_3K_{27}me_3$ and $H_3K_9me_3$ during cell cycle for early and late replicated regions. For each bin covered by the histone modification the mean CpG level is calculated for 20 pseudo time points across the cell cycle, then the value at the first pseudo time point is subtracted to align the signals and allow for a comparison. When the genome is being replicated, the CpG methylation is not always immediately maintained, resulting in a dip of CpG methylation at and after the moment of replication.

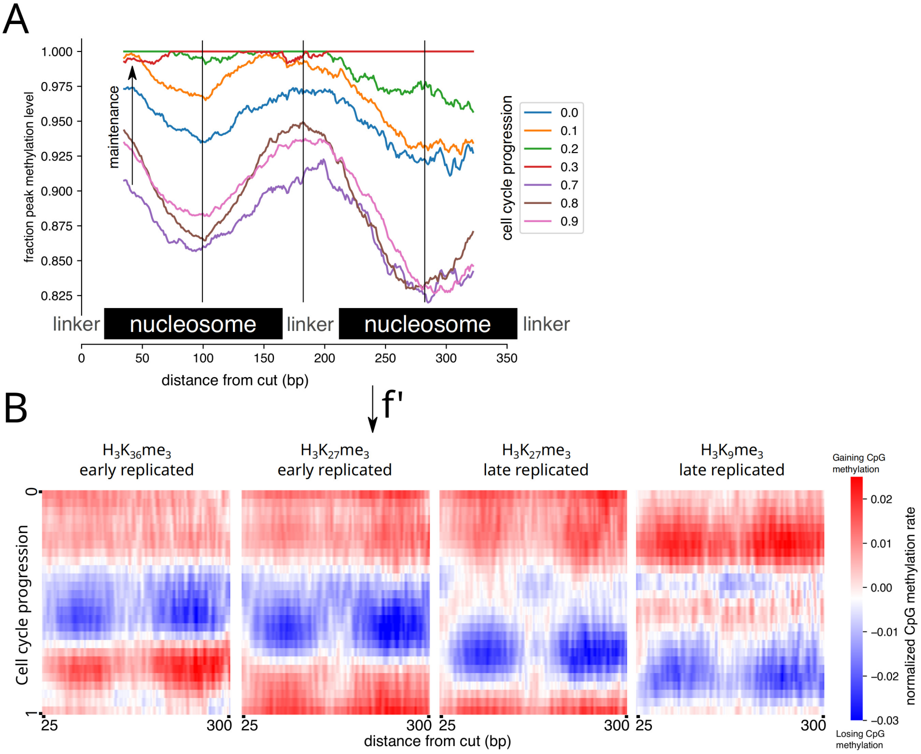


Figure 5: CpG methylation loss and maintenance dynamics relative to MNase cut. A, dynamics for $H_3K_{27}me_3$, showing the CpG methylation level for various progression stages in the cell cycle relative to the CpG methylation at the most maintained state. During DNA replication, the CpG methylation level decreases, and subsequently gradually increases due to maintenance. The largest dynamics are found near the nucleosome core-DNA, and the linker-DNA shows the least dynamics. **B,** the derivative of the CpG methylation dynamics, for $H_3K_{36}me_3$ and $H_3K_{27}me_3$ domains in early replicated regions (first 2 columns) and $H_3K_{27}me_3$ and $H_3K_9me_3$ in late replicated regions.

MNase cut sites across the cell cycle, and normalized this signal to the maximum methylation level found across the cell cycle for that position. This reduces trends caused by limits caused by nucleosome positioning as DNA methylation reduces the flexibility of the DNA which reduce the chance for a nucleosome to form, this is likely one of the reasons the linkers are more highly methylated than the core-DNA [43]. The normalization results in a 'fraction peak methylation level' or *maintenance ratio*, which reflects the DNA methylation maintenance status (a value of 1 reflects a completely maintained CpG and a 0 a completely unmaintained and unmethylated CpG). We plotted the *maintenance ratio* for various positions on the cell cycle progression trajectory (Figure 5 A). We find that the mean CpG methylation level and thus the *maintenance ratio* drops significantly after replication. The linker-DNA shows fewer dynamics relative to the core-DNA, which might be indicative of slower maintenance by DMNT1 due to reduced DNA accessibility due to the presence of the nucleosomes. The derivative of the *maintanance ratio* over the cell cycle is informative about the rate of change of the DNA-methylation levels and is plotted for H3K36me3 (early replicated), H3K27me3 (early and late replicated) and H3K9me3 (late replicated) (Figure 5 B). This shows, that for all marks the linker region has fewer dynamics, while the core-DNA has the largest dynamic range.

4.4 Discussion

We developed scChIC-TAPS , which allows measuring DNA methylation in the vicinity of histone modifications of interest. We profiled scChIC-TAPS in K562 cells, which showed that the resulting data quality is comparable to measuring the modalities separately, and both the histone and methylation measurements correlate well to representative bulk datasets. Technically, scChIC-TAPS can be further improved by more thorough degradation of DNA-linker molecules to get more accurate DNA methylation information relative to the position of the core-DNA. Currently, quite a large fraction of the cells is lost during quality control, especially the correlation analysis which removes cells which show over-digestion by the MNase protein. Improvements to the protocol in order to more precisely control the MNase digestion will lead to less QC-failed cells. While here we used FACS information for cell cycle manifold, we envision that scChIC-TAPS will be extended to include transcriptome information in the future, which allows cell types or differentiation to be used where FACS information is not available. Further, we suggest that scChIC-TAPS will be useful to investigate DNA methylation levels of transcription factor binding sites to study regulation of DNA methylation near histone modifications.

We used scChIC-TAPS in a cell cycle progression system where we profile 3 histone modifications, DNA methylation and cell cycle stage. In general, $H_3K_{36}me_3$ is replicated first, followed by $H_3K_{27}me_3$ and $H_3K_9me_3$. There is a small set of $H_3K_9me_3$ domains which overlap with are $H_3K_{36}me_3$ and are replicated early. Our results provide insights into the open question of why linker DNA is more methylated than core-DNA. It has been suggested that if DNA is methylated too much this could hinder nucleosome formation, and DNTM1 might not be able to reach all nucleosome-bound core-DNA. We show that DNA methylation in the core-DNA has a larger dynamic range than DNA methylation in linker-DNA, indicating that DNA maintenance is indeed slower in the core-DNA. Our results support the hypothesis that DNA methylation maintenance is slower for nucleosome-bound core DNA. DNA methylation influences the stiffness of the DNA, and this in turn influences nucleosome formation. Thus, scChIC-TAPS might be a useful tool to identify cases where DNA methylation is used to dynamically influence the position of the nucleosome [202].

4.5 Materials and methods

Cell Culture

K562 cells (ATCC® CCL-243TM) were cultured in RPMI 1640 GlutaMAX^{TM} medium (Gibco, cat. no. 61870036), supplemented with 5% fetal bovine serum (Gibco, cat. no. A3382001), non-essential amino acids (Gibco, cat. no. 11140050) and Pen/Strep (Gibco, cat. no. 15140122). hTERT-RPE1-FUCCI cells were grown in adherent culture with DMEM/F12 GlutaMAX^{TM} medium (Gibco, cat. no. 10565018) supplemented with Pen/Strep and 10% fetal bovine serum. TrypLE^{TM} Express Enzyme (Gibco, cat. no. 12605010) and PBS were used for passaging of RPE1-Fucci cells.

Single-cell histone profiling

Buffer composition

Component	Manufact.	Cat. No.	WB 1	WB 2	WB 3
H_2O	Invitrogen	AM9932			
HEPES	Gibco	15630080	20 mM	20 mM	20 mM
NaCl	Invitrogen	AM9760G	150 mM	150 mM	150 mM
Spermidine	Sigma	S2626-5G	0.5 mM	0.5 mM	0.5 mM
Tween-20	Sigma	P1379-100ML	0.05%	0.05%	0.05%
Protease inhibitor	Roche	5056489001	1 tablet (per 50 ml)	1 tablet (per 50 ml)	
EDTA	Invitrogen	15575020	2 mM		

* all values correspond to final concentrations

Fixation, cell permeabilization and long-term storage

All steps were performed in Protein LoBind tubes (Eppendorf, cat. no. 0030108094 and 0030122216). Cells were harvested in 15 ml tubes and washed twice with PBS. Fixation was performed by resuspending the pellet in 300 μl PBS per 10^6 cells. Then, 700 μl ice-cold absolute ethanol per 10^6 cells was added while vortexing gently. Cells were fixed at -20°C for two hours, washed twice with WB 1 and transferred to 0.5 ml tubes. Cells were stored at -80°C in WB 1 supplemented with 10% DMSO.

Antibody binding, Pa-MNase incubation and FACS

Protein A-MNase fusion protein (Pa-MNase) was expressed in bacterial culture and purified as outlined in Zeller et al. [195]. Cells were thawed, washed twice with WB 1 and resuspended in WB 1. Histone modification-specific antibodies were added to the reaction (see below for details). Incubation was performed overnight at 4°C with gentle agitation, effectively stripping the cell membrane and releasing nuclei through the presence of Tween-20. Nuclei were washed once and resuspended in 500 μl of WB 2. Pa-MNase (3 ng/ml final) and Hoechst 34580 (5 μg/ml final), were added to each sample, followed by incubation for 60 min at 4°C with gentle agitation. Cells were washed twice with WB 2, resuspended in 500 μl of WB 3 and filtered through a 70 μm strainer (Corning, cat. no. 431751) and transferred to FACS tubes .

Fluorescence-assisted nuclei sorting

Ahead of sorting, 384-well hard-shell plates were prepared for sorting by adding 10 μl of sterile filtered mineral oil (Sigma Aldrich, cat. no. 69794-500ML) per well using a Tecan Freedom EVO® liquid handler. Nuclei in WB 3 were sorted into 384-well plates on a BD Influx$^{\text{TM}}$ cell sorter. Hoechst signal was used to select for K562 cell in G1 phase. Four gates were used for RPE1-Fucci cells to sample evenly from early G1, late G1, S and G2 phase, respectively. Four to eight wells were left empty as controls in all plates. After sorting, cells were spun down for one minute at 2,000 g.

Processing of single-cell plates

All pipetting steps outlined below were performed using an Innovadyne Nanodrop II robotic liquid handler. After each dispension step, plates were sealed with aluminium sealers and spun down for one minute at 2,000 g to fuse droplets.

Pa-MNase activation and Proteinase K digest

MNase digestion was initiated by adding 100 nl of WB 3 supplemented with 2 nM $CaCl_2$ to each well. Plates were incubated for 30 min at 4°C. Digestion was stopped by dispensing 100 nl of the following solution (final concentrations): nuclease-free water; 40 mM EGTA (Thermo Fisher, cat. no. 15425795); 1.5% NP-40 and 2 mg/ml Proteinase K (Invitrogen, AM2548). Plates were incubated in PCR machines: 20 min at 4°C; 6 hours at 65°C; 2 min at 80°C; hold at 4°C. Plates were kept at -80°C until further processing.

Blunting

100 nl of the following mix was added to each well (volumes per well): 2 nl Klenow, large fragment (NEB, cat. no. M0210L); 2 nl T4 PNK (NEB, cat. no. M0201L); 5 nl dNTP solution (Promega, cat. no. U1515); 30 nl ATP 10 mM (NEB, cat. no. P0756S); 30 nl PNK Buffer 10x (NEB, cat. no. M0201L); 10 nl $MgCl_2$ 25 mM (Thermo Fisher, cat. no. 4398828); 5 nl PEG8000 50% (Promega, cat. no. V3011); 1.5 nl BSA 20 mg/ml (NEB, cat. no. B9000S); 14.5 nl nuclease-free H_2O. Incubation: 30 min at 37°C; 20 min at 75°C; hold at 4°C.

A-tailing

200 nl of the following mix was added to each well (volumes per well): 1 nl AmpliTaq 360 DNA Polymerase (Applied Biosystems, cat. no. 4398818); 2 nl T4 PNK (NEB, cat. no. M0201L); 1 nl dATP (Promega, U1205); 10 nl DTT 0.1 M (part of Invitrogen cat. no. 18064022); 14 nl Tris 1 M pH 8.0 (Invitrogen, cat. no. 15568025); 20 nl ATP 10 mM (NEB, cat. no. P0756S); 25 nl KCl 2M (Invitrogen, cat. no. AM9640G); 1 nl $MgCl_2$ 1M (Invitrogen, cat. no. AM9530G); 10 nl PEG8000 50% (Promega, cat. no. V3011); 1 nl BSA 20 mg/ml (NEB, cat. no. B9000S); 115 nl nuclease-free H_2O. Incubation: 15 min at 37°C; 10 min at 72°C; hold at 4°C.

Dispension of barcoded adapters

Per well, 50 nl of 5 μM barcoded adapter was added using a Mosquito HTS Nanolitre Liquid handler (ttplabtech). Adapters were manufactured by IDT, see below for an example sequence. Adapters contain the following features: forked sequence to prevent adapter-adapter ligations (underlined, dotted), T7 promoter (underlined, solid), RA5 Illumina primer binding site (italic), 3 random nucleotides as UMI, an 8 bp cell-specific barcode (bold) and a single-base T overhang.

Top Strand:
5’-GGTGATGCCGGTAATACGACTCACTATAG
*GGAGTTCTACAGTCCGACGAT*CNNN**ACACACTA**T

Bottom Strand:

5’-p**TAGTGTGT**NNN*GATCGTCGGACTGTAGAACTCCC*TATAGTGAGTCGTATTACCGGC*GAGCTT*

Adapter Ligation

150 nl of the following mix was added to each well (volumes per well): 25 nl T4 Ligase 400,000 U/ml (NEB, cat. no. M0202L); 3 nl $MgCl_2$ 1M (Invitrogen, cat. no. AM9530G); 45 nl DTT 0.1 M (part of Invitrogen cat. no. 18064022); 20 nl ATP 10 mM (NEB, cat. no. P0756S); 5 nl PEG8000 50% (Promega, cat. no. V3011); 1 nl BSA 20 mg/ml (NEB, cat. no. B9000S); 51 nl nuclease-free H_2O. Incubation: 20 min at 4°C; 16 hours at 16°C; 10 at 65°C; hold at 4°C.

Pooling of plates

Plates were inverted and placed in pooling plates (Clickbio VBLOK200) pre-coated with 3 ml of sterile filtered mineral oil. Plates were spun for two minutes at 500 g and the liquid phase transferred to fresh 1.5 ml Eppendorf tubes. Carry-over mineral oil was removed with the following washing procedure: 500 μl of n-Butanol were added, tubes inverted multiple times and spun down for one minute at 5,000 g. The butanol phase containing mineral oil was taken off with a P1000 pipette. This procedure was repeated for a total of three times. Then, 500 μl of ether were added. Tubes were vortexed and spun down using a table-top centrifuge. After removal of ether with a P1000 pipette, tubes were left open briefly at room temperature to allow evaporation of left-over ether. Next, DNA was purified by incubating for 10 minutes with 0.8x volumes of Ampure XP beads (Beckman Coulter, cat. no. A63881) pre-diluted 1:4 in bead binding buffer (1 M NaCl, 20% PEG8000, 20 mM Tris pH 8.0, 1 mM EDTA). Beads were pelleted and washed twice with 80% ethanol . Beads were air-dried and resuspended in 19 μl of nuclease-free H_2O. The supernatant was transferred to a fresh 0.5 ml Eppendorf tube. Material was stored at -20°C until further processing.

Methylation profiling

Preparation of spike-ins

To produce fully methylated lambda phage DNA, the following reaction was assembled in 0.5 ml DNA lo-bind Eppendorf tubes: 1 μg of unmethylated lambda phage DNA (Promega, cat. no. D1521); 5 μl NEB Buffer 2 10x (NEB, cat. no. M0226S); 1 μl SAM 32 mM (NEB, cat. no. M0226S); 2 μl M.SssI 4,000 U/ml (NEB, cat. no. M0226S); topped up to 50 μl with nuclease-free H_2O. Incubation: 2 hours at 37°C. After 2 hours an additional 1 μl of SAM and 0.5 μl of M.SssI were added followed by further incubation for 2 hours at 37°C. DNA was cleaned with 1x volume of Ampure XP beads. The above reaction, including the top-up of enzyme and SAM, was repeated with the purified material as input, followed by a final 1x volume Ampure XP bead cleanup and elution in 20 μl of nuclease-free H_2O. Next, methylated DNA was subjected to NlaIII restriction with the following reaction: 1 μl NlaIII 10,000 U/ml (NEB, cat. no. R0125S); 5 μl CutSmart Buffer 10x (NEB, cat. no. R0125S); 24 μl of nuclease-free H_2O. Incubation: 2 hours at 37°C; 20 min at 65°C; hold at 4°C. Material was cleaned up with 1x volumes of Ampure XP beads and the concentration was measured with a Qubit 3 Fluorometer (Invitrogen). Pre-annealed adapter (see below for sequence) was added to the sample in a ratio of 10:1 (based on the measured concentration and assuming full digestion to 180 bp fragments). Next, ligation was performed by addition of the following: 2.5 μl T4 DNA ligase 400,000 U/ml (NEB, cat. no. M0202L); 5 μl T4 DNA ligase buffer 10x (NEB, cat. no. M0202L); volume topped up to 50 μl with nuclease-free H_2O. Ligation was performed for 20 minutes at room temperature followed by heat inactivation for 10 min at 65°C. Material was cleaned up twice with 0.8x volumes of Ampure XP beads. Fully methylated and adapter-ligated spike-ins were diluted to a concentration of 7 pg/μl.

NlaIII adapter top Strand:

5’-GGTGATGCCGGTAATACGACTCACTATAG*GGAGTTCTACAGTCCGACGAT*
CNNN**ACACACTA**CATG

NlaIII adapter bottom Strand:

5’-p**TAGTGTGT**NNN*GATCGTCGGACTGTAGAACTCCC*
TATAGTGAGTCGTATTACCGGC

TET1 enzyme production

Catalytic domain of mouse Ten-eleven translocation methylcytosine dioxygenase 1 (mTET1CD) was expressed as outlined by Liu et al. [107]. Briefly, FLAG-tagged protein was expressed in Expi293F cells (Gibco, cat. no. 13479756). After lysis, protein is bound with Anti-Flag M2 Affinity Gel (Sigma, cat. no. A2220) and purified on gravity chromatography columns (Bio-Rad, cat. no. 7321010) according to the manufacturer's specifications. Protein is concentrated on Amicon® Ultra-4 Centrifugal Filter units (Merck, cat. no. UFC803024) followed by buffer exchange with Bio-Spin® P-30 Gel Columns (Bio-Rad, cat. no. 7326231). Protein was stored at -80°C in 20 mM HEPES pH 8.0, 150 mM NaCl, 1 mM DTT and 30% Glycerol

TAPS conversion and clean-up

Reaction buffer for TAPS consists of (final concentrations): 167 mM HEPES (Gibco, cat. no. 15630080); 333 mM NaCl (Invitrogen, cat. no. AM9760G); 3.3 mM alpha-Ketoglutarate (Sigma-Aldrich, cat. no. K3752-5G); 6.67 mM L-ascorbic acid (Sigma-Aldrich, cat. no. 95210-50G); 4 mM ATP (part of Thermo Fisher Scientific, R0441); 8.33 mM DTT (part of Invitrogen cat. no. 18064022). The following reaction was assembled on ice: 19 μl of pooled material, 1 μl of methylated lambda spike-in, 15 μl of TAPS reaction buffer, 3.33 μl of 1.5 mM Fe^{2+} solution, 12 μl of mTET1CD. Samples were incubated for 80 min at 37°C. Then, 1 μl of Proteinase K 20 mg/ml was added per reaction, followed incubation for 15 min at 55°C. Next, samples were cleaned up with 2x volumes of Ampure XP DNA beads and eluted in 19.67 μl of nuclease-free H_2O. The above reaction and Proteinase K digest were repeated once followed by a clean-up with 2x volumes of Ampure XP DNA beads and elution in 33.75 μl. Sample was transferred to fresh 1.5 ml Eppendorf tubes. Then, 10 μl of NaAc 3M pH 4.3 (produced in-house) and 6.25 μl of pyridine borane solution 10 M (Sigma Aldrich, cat. no. 179752-5G) were added to the reaction mix. Samples were incubated for 16 hours at 37°C in a thermal shaker set to 850 rpm.

After pyridine borane incubation, reactions were cleaned up with oligo clean & concentrator columns (Zymo, cat. no. D4060) according to the manufacturer's protocol with the following adaptations: samples were topped up to 200 μl with nuclease-free H_2O and 400 μl of oligo-binding buffer and 800 μl of ethanol were used per column. Samples were eluted twice with pre-warmed (60°C), nuclease-free H_2O. Then, volumes were reduced to 9.6 μl in a Speed-Vac chamber. Cleaned-up samples were kept at -20°C until library preparation.

Sequencing library preparation

In-vitro transcription (IVT)

TAPS-converted and cleaned up samples were subjected to in-vitro transcription (IVT) by adding 14.4 μl of IVT reaction mix (2.4 μl UTP, 2.4 μl TTP, 2.4 μl GTP, 2.4 μl ATP, 2.4 μl Buffer 10x, 2.4 μl Enzyme; all part of MEGAscriptTM T7 Transcription Kit, Invitrogen, cat. no. AMB13345) followed by incubation for 14 hours at 37°C (with lid temperature set to 70°C). Next, 6 μl of H_2O and 3 μl of Turbo DNAse (part of MEGAscriptTM T7 Transcription Kit) were added and samples incubated for 15 min at 37°C to digest template DNA. Amplified RNA (aRNA) was fragmented by adding 7.88 μl of fragmentation buffer (200 mM Tris-Acetate, pH 8.1; 500 mM KaOAc; 150 mM MgOAc) followed by incubation for 90 s at 94°C. Samples were immediately chilled on ice and 4.13 μl of 0.5 M EDTA pH 8.0 (Invitrogen, cat. no. 15575020) was added to capture Mg^{2+}. Then, aRNA was cleaned with 0.8x volumes of RNAClean XP beads (Beckman Coulter, cat. no. A63987) and eluted in 6 μl of nuclease-free H_2O. In order to assess RNA yield and quality, 1 μl of aRNA was run on a Bioanalyzer (Agilent RNA 6000 Pico Kit, cat. no. 5067-1513).

Reverse transcription and library amplification

After quality control, 5 μl of aRNA were combined with 0.5 μl of 10 mM dNTP solution (Promega, cat. no. U1515) and 1 μl of random hexamer RT primer 20 μM (sequence: GCCTTGGCACCCGAGAATTCCANNNNNN, IDT). Samples were heated to 65°C for 5 minutes and then immediately chilled on ice. 6.5 μl of primed sample were combined with 2 μl First Strand Buffer 5x, 1 μl DTT 0.1 M, 0.5 μl of SuperScriptII 200 U/μl (all part of Invitrogen cat. no. 18064022) and 0.5 μl of RNAseOUT (Invitrogen, cat. no. 10777019). Incubation: 10 min at 25°C; 60 min at 42°C; hold at 4°C. Then, 2 μl of barcoded RPIx primer (see below for example) was added to each sample. Library PCR is performed by adding 11 μl nuclease-free H_2O, 25 μl of NEBNext Ultra II Q5 Master Mix 2x (NEB, cat. no. M0492L) and 2 μl of 10 μM RP1 primer (see below for sequence). Samples are amplified with 10 to 13 cycles of PCR, dependent on histone modification and aRNA yield. PCR settings: 30 s at 98°C; 10 to 13 x [10 s at 98°C, 30 s at 60°C, 30 s at 72°C]; 10 min at 72°C; hold at 4C. Amplified DNA was cleaned with two subsequent 0.8x AMPure XP bead cleanups and eluted in 15 μl of nuclease-free H_2O. Concentration and size distribution were measured on a Qubit 3 Fluorometer and Bioanalyzer (Agilent High Sensitivity DNA kit, cat. no. 5067-4626), respectively. Samples were pooled and sequenced on the Illumina NextSeq2000 platform.

Barcoded RPIx primer (IDT):
5'-CAAGCAGAAGACGGCATACGAGAT-[6bp]
GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

RP1 library PCR primer (IDT):
5'- AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA

Trimming, Demultiplexing, Mapping and deduplication

The scChiC mapping and counting workflow used is identical to to the one used in Zeller et al. [195], and additional steps are performed to perform molecule consensus and methylation calling.

Sequenced reads were demultiplexed using SingleCellMultiOmics demux.py. Next remaining adapter sequences were removed using cutadapt. A custom reference was prepared by combining the following assemblies:

Assembly	**Comment**
Human Ensembl assembly version 97 (Hg38)	K562 and RPE cell lines are human derived
Escherichia Lambda phage acc. J02459.1	used as methylation detection spike-in
Cutibacterium acnes KPA171202	cell culture contaminant
Escherichia coli strain RHB09-C15	cell culture contaminant
Escherichia coli str. K-12 substr. MG1655	MNase protein production contaminant

The trimmed reads were paired-end mapped using BWA mem with default parameters. The resulting mapped reads were quality filtered and deduplicated using SingleCellMultiOmics using the following command:

```
bamtagmultiome.py -method chic sorted.bam -o tagged.bam --
    ↪ multiprocess --one_contig_per_process
```

Molecule consensus calling

In the molecule consensus calling step, the information from reads derived from the same original DNA-template is aggregated. See fig. S4 for a schematic overview. Aggregation starts by clustering reads with the same UMI, starting coordinate, strand and haplotype. These clusters are assumed to be derived from the same template molecule. The information contained in these reads are aggregated in a single consensus alignment and base-calls for each covered base. The aggregated alignment and base-calls will be referred to as consensus molecule from here on.

The generation of a consensus molecule starts by merging the information of the two mates of a paired end read, then the PCR and IVT duplicates are merged resulting in a final consensus molecule.

Paired end reads are merged, for positions where both mates overlap the base call with the highest phred score is selected For dovetailing alignments, the overhanging ends are not considered. Bases with a base calling confidence phred score of 15 are not considered at all.

Finally, the information for each IVT and PCR duplicates are merged. For each covered position the most common base is selected using majority voting over all reads (IVT/PCR duplicates) which cover a location, resulting in a base-call for each covered position of the consensus molecule. Ties are resolved by inserting an ambiguous base call ((N)).

Methylation calling

Methylation calling was performed on the generated consensus base-calls by an extension module of the SingleCellMultiOmics package, which performs methylation calling on consensus molecules. For every covered cytosine (C) of the consensus molecule, a methylation call is performed. When a molecule consensus base call is a C the site is considered unmethylated and for a T the site is considered methylated. A G, A or N base-call result in an ambiguous methylation-call. To avoid incorrect methylation calls due to the presence of SNVs, genomic locations with known $C \rightarrow T$ variants in the cell-line were not taken into account during methylation calling.

Methylation calling, distance to cut aware and feature annotation

To generate methylation calls relative to the cut location of the consensus molecule, SingleCellMultiOmics was extended with a script *tapsTabulator.py* which generates methylation calls which include the location of the MNase cut.

```
tapsTabulator.py tagged.bam -context Z -method chic -
   ↪ features annotations.gtf -min_phred_score 15 |  gzip
   ↪ > Z.annot.tsv.gz
```

TAPs conversion efficiency estimation

The conversion efficiency rate is estimated. This is the ratio of converted CpGs vs total covered CpGs on the lambda-phage genome.

```
estimateTapsConversionEfficiency.py ./sorted.bam -o
   ↪ conv_efficiency -ref [reference_fasta_path] -method
   ↪ chic
```

The full source-code for SingleCellMultiOmics is available at
https://github.com/BuysDB/SingleCellMultiOmics
Additional code with code specific to this manuscript is available at
https://github.com/cgeisenberger/taps-analysis

4.6 Supplementary Figures

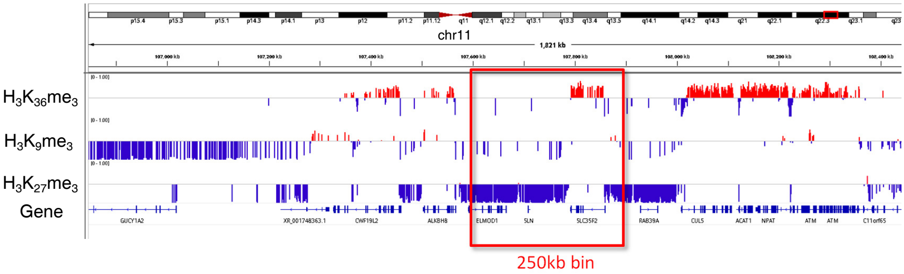


Figure S1: Binning CpG Methylation in large bins results in a higher mean methylation level for $H_3K_{36}me_3$ relative to repressive marks $H_3K_9me_3$ and $H_3K_{27}me_3$. IGV screenshot of a region of chromosome 11 with tracks showing the mean methylation level for $H_3K_{36}me_3$, $H_3K_9me_3$ and $H_3K_{27}me_3$. The beta values range between zero and one, values above one are colored red. An 250kb example bin is depicted with a rectangle.

Cell line	Modification	TAPs conv.	Mapping Rate	Unique
K562	$H_3K_9me_3$	90.9%	98.81%	66.54%
	$H_3K_{27}me_3$	87.0%	93.95%	72.64%
	$H_3K_{36}me_3$	81.0%	90.47%	26.92%
RPE	$H_3K_9me_3$	93.7%	88.33%	43.99%
	$H_3K_{27}me_3$	93.8%	93.26%	39.94%
	$H_3K_{36}me_3$	94.7%	91.14%	36.01%

Table S1: TAPs conversion rates estimated using lambda phage spike-in molecules, the mapping rate (total mapped / demultiplexed reads) and the fraction of unique reads.

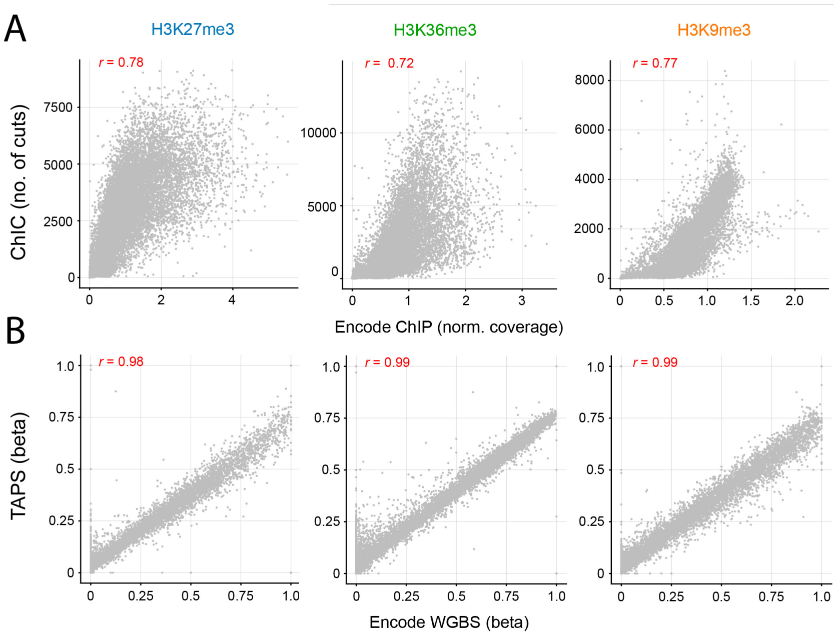


Figure S2: Comparisons of scChIC-TAPS to ENCODE. A, scatterplots comparing scChIC-TAPS MNase cuts (H3K36me3, H3K27me3 and H3K9me3) to ENCODE ChIP. **B,** scatterplots comparing scChIC-TAPS CpG methylation beta values near H3K36me3, H3K27me3 and H3K9me3 to ENCODE WGBS.

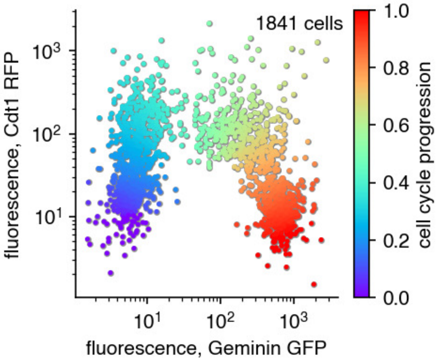


Figure S3: FACS trajectory of cell cycle. Each dot represents a single cell. The position of the cell relates to the measured FACS properties which are influenced by the FUCCI reporter. The color of each cell shows the estimated cell-cycle progression timing of the cell.

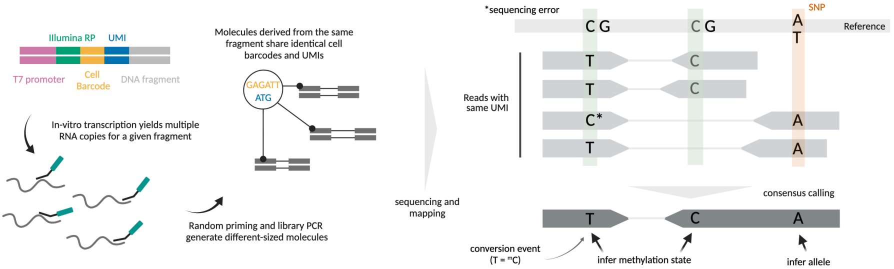

Illumina RP
UMI
T7 promoter
Cell Barcode
DNA fragment
In-vitro transcription yields multiple RNA copies for a given fragment
Molecules derived from the same fragment share identical cell barcodes and UMIs
GAGATT
ATG
Random priming and library PCR generate different-sized molecules
sequencing and mapping
*sequencing error
SNP
Reference
C G
C G
A
T
Reads with same UMI
T
T
C*
T
C
C
A
A
consensus calling
T
C
A
conversion event (T = ^{m}C)
infer methylation state
infer allele

Figure S4: Schematic overview of consensus and methylation calling process. Sequence and alignment information from molecules with the same cell barcode, UMI, starting coordinate, strand (and optionally haplotype) is pooled in order to obtain high confidence consensus base calls from which methylation information can be extracted.

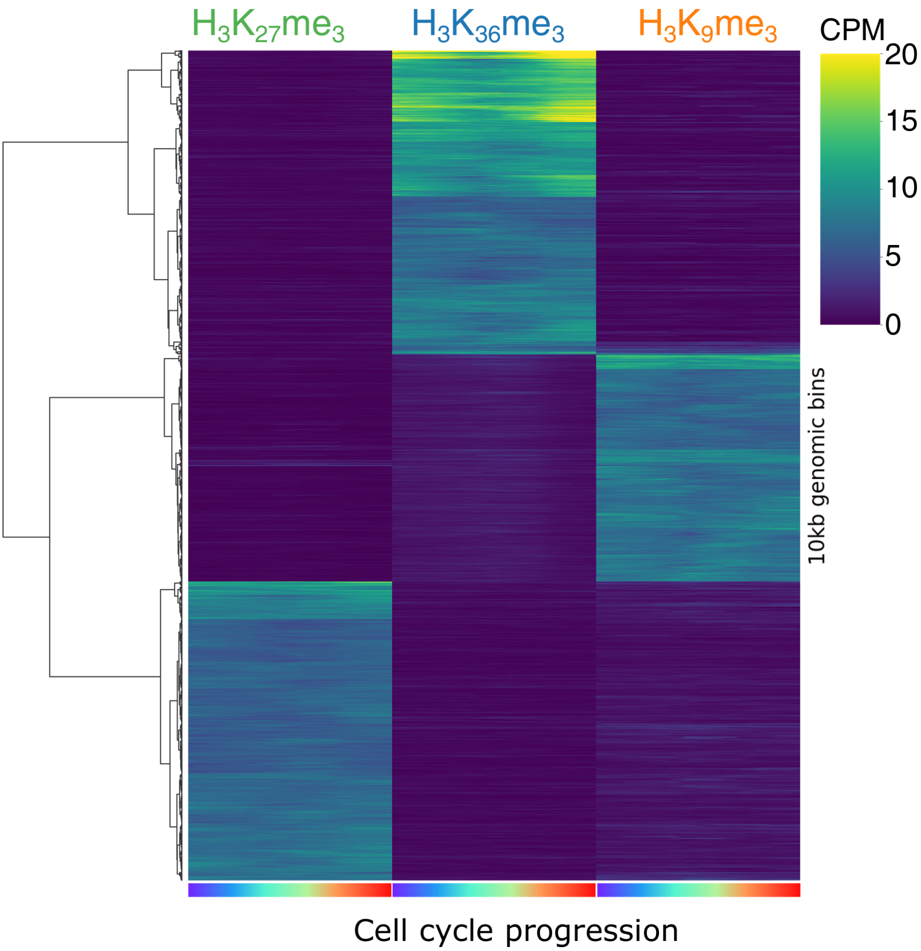


Figure S5: Matrix of ChIC signal at 10kb resolution, resampled along the cell cycle progression trajectory.

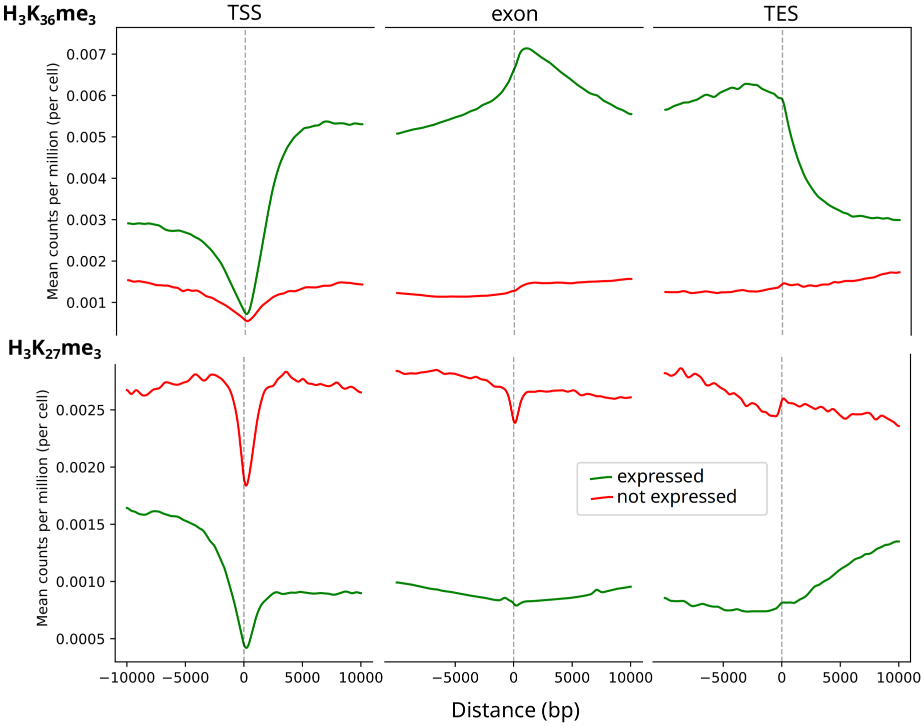


Figure S6: $H_3K_{36}me_3$ and $H_3K_{27}me_3$ distributions on expressed and not expressed genes. The mean count per million reads per cell per base around transcription start sites (TSS), intron/exon boundaries and transcription end sites (TES). Counts are separated into two classes: expressed and not expressed. This class is based on the Battich 2020 dataset [18], genes with at least 20 total transcripts are classified as expressed. A depletion of histone modifications is visible at the transcription start site (TSS), while intron exon boundaries show an enrichment of $H_3K_{36}me_3$. Expressed gene bodies show an enrichment for $H_3K_{36}me_3$ in comparison to not-expressed genes. Inversely for $H_3K_{27}me_3$.

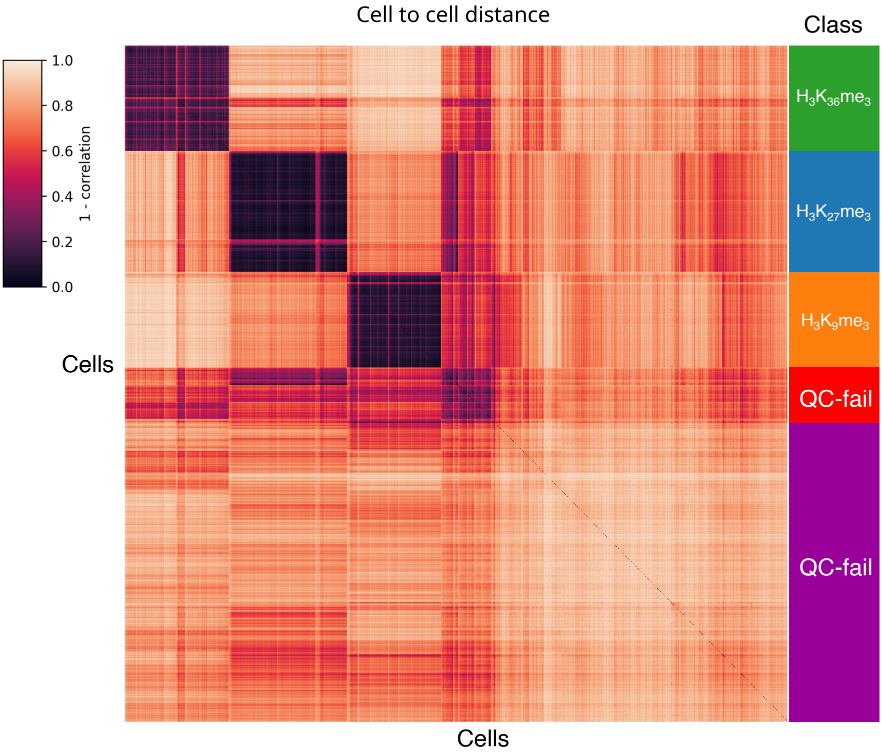


Figure S7: Selection of properly MNase-digested cells by correlation analysis Cell to cell correlation distance matrix clustered using hierarchical clustering. Clusters consisting of a mixture of histone modifications are classified as QC-fail.

Chapter 5

Discussion

Novel biological discoveries are often linked to technological advancement. In this thesis, we demonstrate the application of single cell multi-omics protocols in the context of colorectal carcinoma organoids, blood formation and cell cycle. Apart from technical developments required to measure the here presented modalities from a single cell, data analysis is needed to leverage the information generated during such a multi-omics experiment. Through data integration of three independent lineage measurements from single cell *Whole Genome Sequencing* (WGS), we can reconstruct highly structured events during clonal evolution of colorectal carcinoma organoids (chapter 2). Joint profiling analysis of the interplay between histone modifications marking active chromatin or heterochromatin during blood formation suggests a hierarchical chromatin regulation program, wherein heterochromatin states define lineages, while euchromatin states establish cell types (chapter 3). Integrated data analysis of methylation and histone modifications during cell cycle shows a direct link of local chromatin environment influencing replication-coupled methylation maintenance (chapter 4). In summary, single multi-omics data analysis can aid in providing insights into the order of chromatin changes, expression changes and phenotype and vice versa.

Technological improvements will ease data analysis

While single cell multi-omics data allows for direct measurements of relationship between modalities, the techniques are still in an early stage and improvements in sequencing quality and experimental protocols are going to ease data analysis and yield even more biological insights. Higher quality and longer read sequencing, for example, would greatly improve variant calling and resolve more complex copy number aberrations due to the possibility to phase variants over longer distances. Longer read sequencing additionally enables to phase methylation information over longer distances and allows for the discovery of more allele specific differences. Single cell sequencing protocols currently rely on strong amplification, but reducing amplification bias would improve methylation and variant calling as well as detecting barcode bleed-through, which is often a problem in single cell experiments. Further technical advancement will be focused on the integration of more measurements, such as multiple histone modifications from the same histone to study bi-valency.

Hierarchial analysis and clustering

In chapter 2, hierarchical multi-modal lineage tracing is discussed. Before establishing a lineage hierarchy, clustering of single cells provides useful insights into the clonal events. Resolving the lineage questions with the techniques at hand relies on, however, on biological meaningful clustering of the measured modalities, such as lineage, SNVs, indels, and structural variants. This kind of multi-modal hierarchical analysis is still in early development and can be much improved upon. Currently, the confidence of individual measurements and lack of a measurement are rarely taken into account. Allowing sparse measurements is currently only possible when analyzing a single modality at a time. To obtain a relevant clustering, biological meaningful distances in both, the copy number space and in SNV space, are required. However, these are currently not available. Especially unexplored is tuning the weights of the various modalities when combining distance matrices of multiple modalities. Biologically plausible distance metrics are also important for histone mark analysis, as clustering is usually used to identify groups of cells with the same epigenetic state. For each of these groups, the corresponding pseudobulk[1] is analyzed. When cells are not assigned to a biologically meaningful group, the subsequent pseudobulk is not biologically relevant and could potentially lead to invalid conclusions. These wrong assignments can happen when cells are clustered on a technical artifact, for example, over-digestion or amplification bias. Additionally, having the same cell type in two different clusters or two cell types in one cluster can lead to wrong lineage conclusions. More research is required, to estimate and correct these biases properly, for example by modeling and correcting for MNase digestion efficiency.

Solving hierarchical analysis and simultaneously clustering multiple modalities can not only be applied to lineage questions, but also to establish causal

---

[1] sum of all the data of the cells assigned to a cluster

relationship between modalities, such as the order of epigenetic modifications leading to transcription. Does deposition of $H_3K_4me_3$ precede the decrease of DNA methylation at promoters? Questions about causality are interesting to ask when many combinations of modalities are measured, allowing to further unravel the relationship between various epigenetic marks and transcription and ultimately transcriptional regulation.

Internal validation through complementary measurements

In order to obtain a reliable clustering, it is beneficial to use complementary measurements, which ensure that biological claims and results are supported by multiple lines of evidence. Apart from applying this internal control scheme to copy number and lineage, as shown in chapter 2, this is also applicable when other modalities have been measured. For example, in the scChIC-TAPs project in chapter 4, the FACS parameters provide an additional modality which is complementary to the manifold created using the single cell ChIC data. In the sortChIC analysis in chapter 3, the FACS parameters provide an extra layer of evidence regarding bone-marrow cell-typing. Complementary measurements are most powerful when their technical biases are unrelated, and allow tuning model parameters by using a complementary measurement for cross-validation.

Future implementations of complementary measurements could for example include a protocol allowing to measure sortChIC and transcriptome, allowing the cell typing to be performed on both the transcriptome, histone modification level and FACS properties.

Overcoming data sparsity

A lot of the difficulties in single cell multi-omics data analysis arise from noisy measurements. The most prevalent is dropout, resulting in missing data-points. Less sparse data allows for the detection of more subtle dynamics/changes (e.g. cell cycle). With development of sophisticated analysis most measurement problems can potentially be overcome, technical improvements are the main contributor to reduce data sparsity. Reducing material clean-up and handling before amplification could decrease measurement dropout. Progress made in reducing dropouts in multi-omics protocols will be a big step in boosting single cell measurements to quantitative measurements. Such improvements, together with more sensitive DNA sequencing and less material amplification, would lead, for example, to higher lineage mark recovery rates.

Using negative measurements positively

Some techniques collect negative measurements besides only positive ones. Examples of such protocols are bisulfite sequencing and the TAPs protocol, presented in chapter 4. For TAPs sequencing, both, methylated and unmethylated bases are read out, which allows for the calculation of methylation ratios (beta values). Another example is genotyping of gSNV-phased somatic variants. The negative readout is the detection of the wild type allele, which allows genotyping even when only a single allele is covered. In this line, a negative ChIC read-out would be very beneficial for pointing to locations where a measured histone mark is not present. For the ChIC-TAPS integration, this will allow methylation to be read out in regions where the histone mark of interest is not present, simplifying the generation of a background model (methylation profile for regions where the histone mark is not present).

Quantitative single cell measurements

Nearly all genomic single-cell protocols generate relative measurements, with the readout being a relative signal for each genomic location. Such a relative readout is limited in the information it carries. For example, it is not possible to retrieve information about genome wide changes, such as the doubling of the genome (in copy number analysis), or the genome wide loss of a histone mark. In order to compare absolute difference between cells a quantitative readout is required which is able to capture an absolute signal, like the number of copies of DNA in a cell, the absolute amount of modified nucleosomes etc. One of the ways to obtain an absolute signal is to normalize a relative signal to a reference signal, for example, by spiking in a molecule with known concentration, which is also done in various bulk protocols. So far this has been difficult to implement as the spike-ins need to be subjected to the exact same procedure (and thus biases) as the genomic DNA. Novel efforts aim at determining scaling factors for global changes without the need for spike-ins [87].

Integrating biological knowledge into data-driven science

All projects discussed in this thesis rely heavily on data-driven science, and less on overarching models of molecular cell biology. While a data-driven approach is less biased, it might be beneficial to integrate more prior knowledge into the data-driven algorithms to speed up the scientific lead discovery process. Implementing a-priori biological information was successfully implemented for the scChiC-TAPs analysis in chapter 4, where the known methylation patterns around histones allowed us to tune the methylation caller. Additionally, integration of known cell cycle replication data allowed us to show differences between early and late replicating regions. However, biological knowledge integration does not always yield additional insights. The integration of, for example, pathway analysis with copy number aberrations in chapter 2, did not yield insightful cancer related results. The rapidly advancing field of (transformer based) artificial intelligence algorithms will advance data driven science by fast and precise embedding of experimental results into huge databases of existing knowledge, allowing for faster lead finding and hypothesis generation.

Bibliography

[1] C. Abbosh et al. “Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution”. In: *Nature* 545.7655 (2017), pp. 446–451.

[2] S. Adam, H. Anteneh, M. Hornisch, V. Wagner, J. Lu, N. E. Radde, P. Bashtrykov, J. Song, and A. Jeltsch. “DNA sequence-dependent activity and base flipping mechanisms of DNMT1 regulate genome-wide DNA methylation”. In: *Nature Communications* 11.1 (2020), pp. 1–15.

[3] S. Ai, H. Xiong, C. C. Li, Y. Luo, Q. Shi, Y. Liu, X. Yu, C. Li, and A. He. “Profiling chromatin states using single-cell itChIP-seq”. In: *Nature Cell Biology* 21.9 (2019), pp. 1164–1172.

[4] K. Akashi, D. Traver, T. Miyamoto, and I. L. Weissman. “A clonogenic common myeloid progenitor that gives rise to all myeloid lineages”. In: *Nature* 404.6774 (2000), pp. 193–197.

[5] G. Allfrey, R. Faulkner, and A. E. Mirsky. “Possible Role in the Regulation of Rna Synthesis *”. In: *Biochemistry* 315.1938 (1964), pp. 786–794.

[6] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle. “DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning”. In: *Genome Biology* 18.1 (2017), pp. 1–13.

[7] C. Angermueller et al. “Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity”. In: *Nature Methods* 13.3 (2016), pp. 229–232.

[8] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. “Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets”. In: *Molecular Systems Biology* 14.6 (2018), pp. 1–13.

[9] P. Arnold, I. Erb, M. Pachkov, N. Molina, and E. Van Nimwegen. “MotEvo: Integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences”. In: *Bioinformatics* 28.4 (2012), pp. 487–494.

[10] A. Auton et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74.

[11] G. A. Auwera et al. “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline”. In: *Current Protocols in Bioinformatics* 43.1 (Oct. 2013), pp. 1–33.

[12] C. Baccin, J. Al-Sabah, L. Velten, P. M. Helbling, F. Grünschläger, P. Hernández-Malmierca, C. Nombela-Arrieta, L. M. Steinmetz, A. Trumpp, and S. Haas. “Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization”. In: *Nature Cell Biology* 22.1 (2020), pp. 38–48.

[13] S. Bae and B. J. Lesch. “H3K4me1 Distribution Predicts Transcription State and Poising at Promoters”. In: *Frontiers in Cell and Developmental Biology* 8.May (2020), pp. 1–11.

[14] B. Bakker et al. “Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies”. In: *Genome Biology* 17.1 (2016), pp. 1–15.

[15] K. Barton, N. Muthusamy, C. Fischer, C. N. Ting, T. L. Walunas, L. L. Lanier, and J. M. Leiden. “The Ets-1 transcription factor is required for the development of natural killer cells in mice”. In: *Immunity* 9.4 (1998), pp. 555–563.

[16] M. Bartosovic, M. Kabbe, and G. Castelo-Branco. *Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues*. 2021.

[17] E. Bártová, J. Kreǰcí, A. Harničarová, G. Galiová, and S. Kozubek. “Histone modifications and nuclear architecture: A review”. In: *Journal of Histochemistry and Cytochemistry* 56.8 (2008), pp. 711–721.

[18] N. Battich, J. Beumer, B. De Barbanson, L. Krenning, C. S. Baron, M. E. Tanenbaum, H. Clevers, and A. Van Oudenaarden. “Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies”. In: *Science* 367.6483 (2020), pp. 1151–1156.

[19] E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. “Dimensionality reduction for visualizing single-cell data using UMAP”. In: *Nature Biotechnology* 37.1 (2019), pp. 38–47.

[20] S. Behjati et al. “Genome sequencing of normal cells reveals developmental lineages and mutational processes”. In: *Nature* 513.7518 (2014), pp. 422–425.

[21] C. Beisel and R. Paro. “Silencing chromatin: Comparing modes and mechanisms”. In: *Nature Reviews Genetics* 12.2 (2011), pp. 123–135.

[22] O. Bell, V. K. Tiwari, N. H. Thomä, and D. Schübeler. “Determinants and dynamics of genome accessibility”. In: *Nature Reviews Genetics* 12.8 (2011), pp. 554–564.

[23] Y. Benjamini and T. P. Speed. “Summarizing and correcting the GC content bias in high-throughput sequencing”. In: *Nucleic Acids Research* 40.10 (2012), pp. 1–14.

[24] A. C. Bester, M. Roniger, Y. S. Oren, M. M. Im, D. Sarni, M. Chaoat, A. Bensimon, G. Zamir, D. S. Shewach, and B. Kerem. “Nucleotide deficiency promotes genomic instability in early stages of cancer development”. In: *Cell* 145.3 (2011), pp. 435–446.

[25] S. Bian et al. “Single-cell multiomics sequencing and analyses of human colorectal cancer”. In: *Science* 362.6418 (2018), pp. 1060–1063.

[26] A. P. Bird. “Use of restriction enzymes to study eukaryotic DNA methylation. II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern”. In: *Journal of Molecular Biology* 118.1 (1978), pp. 49–60.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent Dirichlet Allocation”. In: *J. Mach. Learn. Res.* 3.null (2003), pp. 993–1022.

[28] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. “Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data”. In: *Bioinformatics* 28.3 (2012), pp. 423–425.

[29] V. Boeva, A. Zinovyev, K. Bleakley, J. P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. “Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization”. In: *Bioinformatics* 27.2 (2011), pp. 268–269.

[30] C. L. Bohrson et al. “Linked-read analysis identifies mutations in single-cell DNA-sequencing data”. In: *Nature Genetics* 51.4 (Apr. 2019), pp. 749–754.

[31] A. C. Bolhaqueiro et al. “Ongoing chromosomal instability and karyotype evolution in human colorectal cancer organoids”. In: *Nature Genetics* 51.5 (2019), pp. 824–834.

[32] L. A. Boyer et al. “Polycomb complexes repress developmental regulators in murine embryonic stem cells”. In: *Nature* 441.7091 (2006), pp. 349–353.

[33] A. B. Brinkman et al. “Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation”. In: *Nature Communications* 10.1 (2019).

[34] J. D. Buenrostro, M. R. Corces, C. A. Lareau, B. Wu, A. N. Schep, M. J. Aryee, R. Majeti, H. Y. Chang, and W. J. Greenleaf. “Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation”. In: *Cell* 173.6 (2018), 1535–1548.e16.

[35] W. Buntine and A. Jakulin. “Applying Discrete PCA in Data Analysis”. In: *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI2004)* (2004), pp. 59–66.

[36] R. A. Burrell et al. “Replication stress links structural and numerical cancer chromosomal instability”. In: *Nature* 494.7438 (2013), pp. 492–496.

[37] L. Busto-Moner, J. Morival, H. Ren, A. Fahim, Z. Reitz, T. L. Downing, and E. L. Read. “Stochastic modeling reveals kinetic heterogeneity in post-replication DNA methylation”. In: *PLoS Computational Biology* 16.4 (2020), pp. 1–23.

[38] J. C. Campbell, A. Hindle, and E. Stroulia. “Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data”. In: *The Art and Science of Analyzing Software Data* 3 (2015), pp. 139–159.

[39] A. K. Casasent, A. Schalck, R. Gao, E. Sei, A. Long, W. Pangburn, T. Casasent, F. Meric-Bernstam, M. E. Edgerton, and N. E. Navin. “Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing”. In: *Cell* 172.1-2 (2018), 205–217.e12.

[40] M. Catoni, J. M. Tsang, A. P. Greco, and N. R. Zabet. “DMRcaller: A versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts”. In: *Nucleic Acids Research* 46.19 (2018).

[41] C. Chen, D. Xing, L. Tan, H. Li, G. Zhou, L. Huang, and X. S. Xie. “Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI)”. In: *Science* 356.6334 (2017), pp. 189–194.

[42] L. F. Cheow et al. “Single-cell multimodal profiling reveals cellular epigenetic heterogeneity”. In: *Nature Methods* 13.10 (2016), pp. 833–836.

[43] R. K. Chodavarapu et al. “Relationship between nucleosome positioning and DNA methylation”. In: *Nature* 466.7304 (2010), pp. 388–392.

[44] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples”. In: *Nature Biotechnology* 31.3 (2013), pp. 213–219.

[45] S. J. Clark, S. A. Smallwood, H. J. Lee, F. Krueger, W. Reik, and G. Kelsey. “Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq)”. In: *Nature Protocols* 12.3 (2017), pp. 534–547.

[46] A. Cloutier, C. Guindi, P. Larivée, C. M. Dubois, A. Amrani, and P. P. McDonald. “Inflammatory Cytokine Production by Human Neutrophils Involves C/EBP Transcription Factors”. In: *The Journal of Immunology* 182.1 (2009), pp. 563–571.

[47] J. Cohen. “Sorting out chromosome errors”. In: *Science* 296.5576 (2002), pp. 2164–2166.

[48] C. K. Collings, P. J. Waddell, and J. N. Anderson. “Effects of DNA methylation on nucleosome stability”. In: *Nucleic Acids Research* 41.5 (2013), pp. 2918–2931.

[49] D. F. Conrad et al. “Variation in genome-wide mutation rates within and between human families”. In: *Nature Genetics* 43.7 (2011), pp. 712–714.

[50] C. A. Davis et al. “The Encyclopedia of DNA elements (ENCODE): Data portal update”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D794–D801.

[51] T. Davoli, A. W. Xu, K. E. Mengwasser, L. M. Sack, J. C. Yoon, P. J. Park, and S. J. Elledge. “Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome”. In: *Cell* 155.4 (2013), p. 948.

[52] F. B. Dean, J. R. Nelson, T. L. Giesler, and R. S. Lasken. “Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification”. In: *Genome Research* 11.6 (2001), pp. 1095–1099.

[53] J. Demeulemeester et al. “Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing”. In: *Genome Biology* 17.1 (2016), pp. 1–15.

[54] M. A. Depristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. In: *Nature Genetics* 43.5 (2011), pp. 491–501.

[55] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. “PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors”. In: *Genome Biology* 16.1 (2015), pp. 1–20.

[56] S. M. Dewhurst et al. “Tolerance of whole- genome doubling propagates chromosomal instability and accelerates cancer genome evolution”. In: *Cancer Discovery* 4.2 (2014), pp. 175–185.

[57] S. S. Dey, L. Kester, B. Spanjaard, M. Bienko, and A. Van Oudenaarden. “Integrated genome and transcriptome sequencing of the same cell”. In: *Nature Biotechnology* 33.3 (2015), pp. 285–289.

[58] X. Dong, L. Zhang, B. Milholland, M. Lee, A. Y. Maslov, T. Wang, and J. Vijg. “Accurate identification of single-nucleotide variants in whole-genome-amplified single cells”. In: *Nature Methods* 14.5 (2017), pp. 491–493.

[59] R. J. Dress et al. “Plasmacytoid dendritic cells develop from Ly6D+ lymphoid progenitors distinct from the myeloid lineage”. In: *Nature Immunology* 20.7 (2019), pp. 852–864.

[60] J. Drost et al. “Sequential cancer mutations in cultured human intestinal stem cells”. In: *Nature* 521.7550 (2015), pp. 43–47.

[61] J. Edmonds. “Optimum branchings”. In: *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics* 71B.4 (1967), p. 233.

[62] E. R. Fearon. “Molecular genetics of colorectal cancer”. In: *Annual Review of Pathology: Mechanisms of Disease* 6 (2011), pp. 479–507.

[63] E. R. Fearon, S. R. Hamilton, and B. Vogelstein. “Clonal analysis of human colorectal tumors”. In: *Science* 238.4824 (1987), pp. 193–197.

[64] N. Feldman, A. Gerson, J. Fang, E. Li, Y. Zhang, Y. Shinkai, H. Cedar, and Y. Bergman. “G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis”. In: *Nature Cell Biology* 8.2 (2006), pp. 188–194.

[65] P. R. Freire and O. M. Conneely. “NR4A1 and NR4A3 restrict HSC proliferation via reciprocal regulation of C/EBPa and inflammatory signaling”. In: *Blood* 131.10 (2018), pp. 1081–1093.

[66] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. “A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands”. In: *Proceedings of the National Academy of Sciences of the United States of America* 89.5 (1992), pp. 1827–1831.

[67] X. Fu, C. Zhang, and Y. Zhang. “Epigenetic regulation of mouse preimplantation embryo development”. In: *Current Opinion in Genetics and Development* 64 (2020), pp. 13–20.

[68] Y. Fu, C. Li, S. Lu, W. Zhou, F. Tang, X. S. Xie, and Y. Huang. “Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.38 (2015), pp. 11923–11928.

[69] T. Garvin, R. Aboukhalil, J. Kendall, T. Baslan, G. S. Atwal, J. Hicks, M. Wigler, and M. C. Schatz. “Interactive analysis and assessment of single-cell copy-number variations”. In: *Nature Methods* 12.11 (2015), pp. 1058–1060.

[70] C. Gawad, W. Koh, and S. R. Quake. “Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.50 (2014), pp. 17947–17952.

[71] M. Gerlinger et al. “Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing”. In: *New England Journal of Medicine* 366.10 (Mar. 2012), pp. 883–892.

[72] A. Giladi et al. “Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis”. In: *Nature Cell Biology* 20.7 (2018), pp. 836–846.

[73] K. Grosselin et al. “High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer”. In: *Nature Genetics* 51.6 (2019), pp. 1060–1066.

[74] S. J. Hainer, A. Bošković, K. N. McCannell, O. J. Rando, and T. G. Fazzio. “Profiling of Pluripotency Factors in Single Cells and Early Embryos”. In: *Cell* 177.5 (2019), 1319–1329.e11.

[75] M. A. Hall, N. J. Slater, C. G. Begley, J. M. Salmon, L. J. Van Stekelenburg, M. P. McCormack, S. M. Jane, and D. J. Curtis. “Functional but Abnormal Adult Erythropoiesis in the Absence of the Stem Cell Leukemia Gene”. In: *Molecular and Cellular Biology* 25.15 (2005), pp. 6355–6362.

[76] K. D. Hansen, B. Langmead, and R. A. Irizarry. “BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions”. In: *Genome Biology* 13.10 (2012).

[77] A. Harada, K. Maehara, T. Handa, Y. Arimura, J. Nogami, Y. Hayashi-Takanaka, K. Shirahige, H. Kurumizaka, H. Kimura, and Y. Ohkawa. “A chromatin integration labelling method enables epigenomic profiling with lower input”. In: *Nature Cell Biology* 21.2 (2019), pp. 287–296.

[78] J. Hård et al. “Conbase: A software for unsupervised discovery of clonal somatic mutations in single cells through read phasing”. In: *Genome Biology* 20.1 (2019), pp. 1–18.

[79] N. D. Heintzman et al. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”. In: *Nature Genetics* 39.3 (2007), pp. 311–318.

[80] K. Honda et al. “IRF-7 is the master regulator of type-I interferon-dependent immune responses.” eng. In: *Nature* 434.7034 (Apr. 2005), pp. 772–777.

[81] K. Hornik and B. Grün. “topicmodels: An R Package for Fitting Topic Models”. In: *Journal of Statistical Software.* 40.13 (2011), pp. 1–30.

[82] M. Hosokawa, Y. Nishikawa, M. Kogawa, and H. Takeyama. “Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics”. In: *Scientific Reports* 7.1 (2017), pp. 3–4.

[83] F. S. Howe, H. Fischl, S. C. Murray, and J. Mellor. “Is H3K4me3 instructive for transcription activation?” In: *BioEssays* 39.1 (2017), pp. 1–12.

[84] International HapMap Consortium. “International HapMap Consortium. The International HapMap Project.” In: *Nature* 426.6968 (2003), pp. 789–796.

[85] E. Izumchenko et al. “Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA”. In: *Nature Communications* 6 (2015), pp. 1–13.

[86] D. H. Janssens, M. P. Meers, S. J. Wu, E. Babaeva, S. Meshinchi, J. F. Sarthy, K. Ahmad, and S. Henikoff. “Automated CUT&amp;Tag profiling of chromatin heterogeneity in mixed-lineage leukemia”. In: *bioRxiv* (Jan. 2021), p. 2020.10.06.32894

[87] H. Jin, L. H. Kasper, J. D. Larson, G. Wu, S. J. Baker, J. Zhang, and Y. Fan. “ChIPseqSpikeInFree: A ChIP-seq normalization approach to reveal global changes in histone modifications without spike-in”. In: *Bioinformatics* 36.4 (2020), pp. 1270–1272.

[88] C. A. Kapourani and G. Sanguinetti. “Higher order methylation features for clustering and prediction in epigenomic studies”. In: *Bioinformatics* 32.17 (2016), pp. i405–i412. arXiv: `1603.08386`.

[89] H. S. Kaya-Okur, S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff. “CUT&Tag for efficient epigenomic profiling of small samples and single cells”. In: *Nature Communications* 10.1 (2019), pp. 1–10.

[90] J. D. Kim, C. Faulk, and J. Kim. “Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1”. In: *Nucleic Acids Research* 35.10 (2007), pp. 3442–3452.

[91] M. Kimura. “The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations.” eng. In: *Genetics* 61.4 (Apr. 1969), pp. 893–903.

[92] K. J. Knudsen et al. “ERG promotes the maintenance of hematopoietic stem cells by restricting their differentiation”. In: *Genes and Development* 29.18 (2015), pp. 1915–1929.

[93] O. Kopper et al. “An organoid platform for ovarian cancer captures intra- and interpatient heterogeneity”. In: *Nature Medicine* 25.5 (2019), pp. 838–849.

[94] F. Krueger and S. R. Andrews. “Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications”. In: *Bioinformatics* 27.11 (2011), pp. 1571–1572.

[95] W. L. Ku, K. Nakamura, W. Gao, K. Cui, G. Hu, Q. Tang, B. Ni, and K. Zhao. “Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification”. In: *Nature Methods* 16.4 (2019), pp. 323–325.

[96] W. L. Ku, L. Pan, Y. Cao, W. Gao, and K. Zhao. “Profiling single-cell histone modifications using indexing chromatin immunocleavage sequencing”. In: *Genome Research* (2021), pp. 1–12.

[97] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel. “Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors”. In: *Genome Research* 27.11 (2017), pp. 1885–1894.

[98] J. Kuipers, M. A. Tuncel, P. Ferreira, K. Jahn, and N. Beerenwinkel. “Single-cell copy number calling and event history reconstruction”. In: *bioRxiv* (2020), pp. 1–24.

[99] D. Lähnemann et al. *Eleven grand challenges in single-cell data science*. Vol. 21. 1. Genome Biology, 2020, pp. 1–35.

[100] B. Langmead and S. L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359.

[101] D. Lara-Astiaso et al. “Chromatin state dynamics during blood formation HHS Public Access”. In: *Science* 345.6199 (2014), pp. 943–949.

[102] S. F. Levy, J. R. Blundell, S. Venkataram, D. A. Petrov, D. S. Fisher, and G. Sherlock. “Quantitative evolutionary dynamics using high-resolution lineage tracking”. In: *Nature* 519.7542 (2015), pp. 181–186.

[103] B. Li and C. N. Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.” In: *BMC bioinformatics* 12 (Jan. 2011), p. 323.

[104] H. Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: *arXiv preprint arXiv* 00.00 (2013), p. 3. arXiv: `1303.3997`.

[105] H. Li and R. Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[106] T. Liu, Z. Chen, B. Zhang, W. Y. Ma, and G. Wu. “Improving text classification using local Latent Semantic Indexing”. In: *Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004* (2004), pp. 162–169.

[107] Y. Liu, P. Siejka-Zielińska, G. Velikova, Y. Bi, F. Yuan, M. Tomkova, C. Bai, L. Chen, B. Schuster-Böckler, and C. X. Song. “Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution”. In: *Nature Biotechnology* 37.4 (2019), pp. 424–429.

[108] Z. Lu, C. C. Hong, G. Kong, A. L. Assumpção, I. M. Ong, E. H. Bresnick, J. Zhang, and X. Pan. “Polycomb Group Protein YY1 Is an Essential Regulator of Hematopoietic Stem Cell Quiescence”. In: *Cell Reports* 22.6 (2018), pp. 1545–1559.

[109] L. J. Luquette, C. L. Bohrson, M. A. Sherman, and P. J. Park. “Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance”. In: *Nature Communications* 10.1 (2019).

[110] I. C. Macaulay, C. P. Ponting, and T. Voet. “Single-Cell Multiomics: Multiple Measurements from Single Cells”. In: *Trends in Genetics* 33.2 (2017), pp. 155–168.

[111] I. C. Macaulay, M. J. Teng, W. Haerty, P. Kumar, C. P. Ponting, and T. Voet. "Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq". In: *Nature Protocols* 11.11 (2016), pp. 2081–2103.

[112] M. Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17 (2011), pp. 10–12.

[113] A. K. Maunakea et al. "Conserved role of intragenic DNA methylation in regulating alternative promoters". In: *Nature* 466.7303 (2010), pp. 253–257.

[114] T. R. Mayo, G. Schweikert, and G. Sanguinetti. "M 3 D: A kernel-based test for spatially correlated changes in methylation profiles". In: *Bioinformatics* 31.6 (2015), pp. 809–816.

[115] N. McGranahan and C. Swanton. "Biological and therapeutic impact of intratumor heterogeneity in cancer evolution". In: *Cancer Cell* 27.1 (2015), pp. 15–26.

[116] N. McGranahan and C. Swanton. "Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future". In: *Cell* 168.4 (2017), pp. 613–628.

[117] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis". In: *Nucleic Acids Research* 33.18 (2005), pp. 5868–5877.

[118] T. S. Mikkelsen et al. "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells". In: *Nature* 448.7153 (2007), pp. 553–560.

[119] D. Mooijman, S. S. Dey, J. C. Boisset, N. Crosetto, and A. Van Oudenaarden. "Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction". In: *Nature Biotechnology* 34.8 (2016), pp. 852–856.

[120] M. J. Muraro et al. "A Single-Cell Transcriptome Atlas of the Human Pancreas". In: *Cell Systems* 3.4 (2016), 385–394.e3.

[121] N. Navin et al. "Tumour evolution inferred by single-cell sequencing". In: *Nature* 472.7341 (2011), pp. 90–95.

[122] D. Nicetto and K. S. Zaret. "Role of H3K9me3 heterochromatin in cell identity establishment and maintenance". In: *Current Opinion in Genetics and Development* 55 (2019), pp. 1–10.

[123] M. Nieser et al. "Loss of Chromosome 18 in Neuroendocrine Tumors of the Small Intestine: The Enigma Remains". In: *Neuroendocrinology* 104.3 (2017), pp. 302–312.

[124] M. Noll. "Subunit structure of chromatin". In: *Nature* 251.5472 (1974), pp. 249–251.

[125] P. C. Nowell. "The clonal evolution of tumor cell populations". In: *Science* 194.4260 (1976), pp. 23–28.

[126] H. N. Nowyhed, T. R. Huynh, A. Blatchley, R. Wu, G. D. Thomas, and C. C. Hedrick. "The Nuclear Receptor Nr4a1 Controls CD8 T Cell Development Through Transcriptional Suppression of Runx3". In: *Scientific Reports* 5 (2015), pp. 2–10.

[127] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data". In: *Biostatistics* 5.4 (2004), pp. 557–572.

[128] S. H. Orkin. "Diversification of haematopoietic stem cells to specific lineages". In: *Nature Reviews Genetics* 1.1 (2000), pp. 57–64.

[129] S. H. Orkin and L. I. Zon. "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology". In: *Cell* 132.4 (2008), pp. 631–644.

[130] F. Paul et al. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors". In: *Cell* 163.7 (2015), pp. 1663–1677.

[131] F. M. Pauler, M. A. Sloane, R. Huang, K. Regha, M. V. Koerner, I. Tamir, A. Sommer, A. Aszodi, T. Jenuwein, and D. P. Barlow. "H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome". In: *Genome Research* 19.2 (2009), pp. 221–233.

[132] A. H. Peters et al. "Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability". In: *Cell* 107.3 (2001), pp. 323–337.

[133] N. Petryk, S. Bultmann, T. Bartke, and P. A. Defossez. "Staying true to yourself: Mechanisms of DNA methylation maintenance in mammals". In: *Nucleic Acids Research* 49.6 (2021), pp. 3020–3032.

[134] R. Poplin et al. "A universal snp and small-indel variant caller using deep neural networks". In: *Nature Biotechnology* 36.10 (2018), p. 983.

[135] S. Pott. "Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells". In: *eLife* 6 (2017), pp. 1–19.

[136] A. M. Ranzoni et al. "Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis". In: *Cell Stem Cell* 28.3 (2021), 472–487.e7.

[137] T. J. Richmond and C. A. Davey. "The structure of DNA in the nucleosome core". In: *Nature* 423.6936 (2003), pp. 145–150.

[138] P. F. Rodrigues, L. Alberti-Servera, A. Eremin, G. E. Grajales-Reyes, R. Ivanek, and R. Tussiwand. “Distinct progenitor lineages contribute to the heterogeneity of plasmacytoid dendritic cells”. In: *Nature Immunology* 19.7 (2018), pp. 711–722.

[139] F. Rohart, A. Eslami, N. Matigian, S. Bougeard, and K. A. Lê Cao. “MINT: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms”. In: *BMC Bioinformatics* 18.1 (2017), pp. 1–13.

[140] E. M. Ross and F. Markowetz. “OncoNEM: Inferring tumor evolution from single-cell sequencing data”. In: *Genome Biology* 17.1 (2016), pp. 1–14.

[141] A. Rotem, O. Ram, N. Shoresh, R. A. Sperling, A. Goren, D. A. Weitz, and B. E. Bernstein. “Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state”. In: *Nature Biotechnology* 33.11 (2015), pp. 1165–1172.

[142] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. “PyClone: Statistical inference of clonal population structure in cancer”. In: *Nature Methods* 11.4 (2014), pp. 396–398.

[143] A. Roth et al. “Clonal genotype and population structure inference from single-cell tumor sequencing”. In: *Nature Methods* 13.7 (2016), pp. 573–576.

[144] R. Sachidanandam et al. “A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms”. In: *Nature* 409.6822 (2001), pp. 928–933.

[145] A. Sakaue-Sawano et al. “Visualizing Spatiotemporal Dynamics of Multicellular Cell-Cycle Progression”. In: *Cell* 132.3 (2008), pp. 487–498.

[146] A. Salhab et al. “A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains”. In: *Genome Biology* 19.1 (2018), pp. 9–11.

[147] G. Satas, S. Zaccaria, G. Mon, and B. J. Raphael. “SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses”. In: *Cell Systems* 10.4 (2020), 323–332.e8.

[148] A. T. Satpathy, C. G. Briseño, X. Cai, D. G. Michael, C. Chou, S. Hsiung, D. Bhattacharya, N. A. Speck, and T. Egawa. “Runx1 and Cbfβ regulate the development of Flt3+ dendritic cell progenitors and restrict myeloproliferative disorder”. In: *Blood* 123.19 (2014), pp. 2968–2977.

[149] C. M. Sawai, V. Sisirak, H. S. Ghosh, E. Z. Hou, M. Ceribelli, L. M. Staudt, and B. Reizis. “Transcription factor Runx2 controls the development and migration of plasmacytoid dendritic cells”. In: *Journal of Experimental Medicine* 210.11 (2013), pp. 2151–2159.

[150] S. Saxonov, P. Berg, and D. L. Brutlag. “A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.5 (2006), pp. 1412–1417.

[151] M. Schmid, T. Durussel, and U. K. Laemmli. “ChIC and ChEC”. In: *Molecular Cell* 16.1 (2004), pp. 147–157.

[152] D. E. Schones, K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. “Dynamic Regulation of Nucleosome Positioning in the Human Genome”. In: *Cell* 132.5 (2008), pp. 887–898.

[153] L. M. Scott, C. I. Civin, P. Rorth, and A. D. Friedman. “A novel temporal expression pattern of three C/EBP family members in differentiating myelomonocytic cells”. In: *Blood* 80.7 (1992), pp. 1725–1735.

[154] M. Sen, D. Mooijman, A. Chialastri, J. C. Boisset, M. Popovic, B. Heindryckx, S. M. Chuva de Sousa Lopes, S. S. Dey, and A. van Oudenaarden. “Strand-specific single-cell methylomics reveals distinct modes of DNA demethylation dynamics during early mammalian development”. In: *Nature Communications* 12.1 (2021), pp. 1–10.

[155] S. P. Shah et al. “The clonal and mutational evolution spectrum of primary triple-negative breast cancers.” eng. In: *Nature* 486.7403 (Apr. 2012), pp. 395–399.

[156] N. Shivapurkar, A. Maitra, S. Milchgrub, and A. F. Gazdar. “Deletions of chromosome 4 occur early during the pathogenesis of colorectal carcinoma”. In: *Human Pathology* 32.2 (2001), pp. 169–177.

[157] A. M. Sidore, F. Lan, S. W. Lim, and A. R. Abate. “Enhanced sequencing coverage with digital droplet multiple displacement amplification”. In: *Nucleic Acids Research* 44.7 (2015), pp. 1–9.

[158] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp. “Detecting DNA cytosine methylation using nanopore sequencing”. In: *Nature Methods* 14.4 (2017), pp. 407–410.

[159] J. Singer, J. Kuipers, K. Jahn, and N. Beerenwinkel. “Single-cell mutation identification via phylogenetic inference”. In: *Nature Communications* 9.1 (2018), pp. 1–8.

[160] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K. A. L. Cao. “DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays”. In: *Bioinformatics* 35.17 (2019), pp. 3055–3062.

[161] P. J. Skene, J. G. Henikoff, and S. Henikoff. “Targeted in situ genome-wide profiling with high efficiency for low cell numbers”. In: *Nature Protocols* 13.5 (2018), pp. 1006–1019.

[162] S. A. Smallwood, H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey. “Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity”. In: *Nature Methods* 11.8 (Aug. 2014), pp. 817–820.

[163] T. Smith, A. Heger, and I. Sudbery. “UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy”. In: *Genome Research* 27.3 (2017), pp. 491–499.

[164] A. Sottoriva et al. “A big bang model of human colorectal tumor growth”. In: *Nature Genetics* 47.3 (2015), pp. 209–216.

[165] G. J. Spangrude, S. Heimfeld, and I. L. Weissman. “Purification and characterization of mouse hematopoietic stem cells.” eng. In: *Science (New York, N.Y.)* 241.4861 (July 1988), pp. 58–62.

[166] J. Starmer and T. Magnuson. “Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains”. In: *BMC Bioinformatics* 17.1 (2016), pp. 1–10.

[167] S. Takahashi, H. Miura, T. Shibata, K. Nagao, K. Okumura, M. Ogata, C. Obuse, S. ichiro Takebayashi, and I. Hiratani. “Genome-wide stability of the DNA replication program in single mammalian cells”. In: *Nature Genetics* 51.3 (2019), pp. 529–540.

[168] J. Tang et al. “The genomic landscapes of individual melanocytes from human skin”. In: *Nature* 586.7830 (2020), pp. 600–605.

[169] H. Telenius, N. P. Carter, C. E. Bebb, M. Nordenskjöld, B. A. Ponder, and A. Tunnacliffe. “Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer”. In: *Genomics* 13.3 (1992), pp. 718–725.

[170] L. Teytelman, B. Özaydin, O. Zill, P. Lefrançois, M. Snyder, J. Rine, and M. B. Eisen. “Impact of chromatin structures on DNA processing for genomic analyses”. In: *PLoS ONE* 4.8 (2009), pp. 1–11.

[171] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry. “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model”. In: *Genome Biology* 20.1 (2019), pp. 1–16.

[172] S. Turajlic et al. “Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal”. In: *Cell* 173.3 (2018), 581–594.e12.

[173] A. Unnikrishnan, W. M. Freeman, J. Jackson, J. D. Wren, H. Porter, and A. Richardson. *The role of DNA methylation in epigenetics of aging*. 2019.

[174] R. Vaisvila et al. “Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA”. In: *Genome Research* 31.7 (2021), pp. 1280–1289.

[175] N. D. Vanderkraats, J. F. Hiken, K. F. Decker, and J. R. Edwards. “Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes”. In: *Nucleic Acids Research* 41.14 (2013), pp. 6816–6827.

[176] E. Vanneste, N. Van Der Aa, T. Voet, and J. R. Vermeesch. “Aneuploidy and copy number variation in early human development”. In: *Seminars in Reproductive Medicine* 30.4 (2012), pp. 302–308.

[177] B. Vilagos et al. “Essential role of EBF1 in the generation and function of distinct mature B cell types”. In: *Journal of Experimental Medicine* 209.4 (2012), pp. 775–792.

[178] C. Vincent-Fabert, N. Platet, A. Vandevelde, M. Poplineau, M. Koubi, P. Finetti, G. Tiberi, A. M. Imbert, F. Bertucci, and E. Duprez. “PLZF mutation alters mouse hematopoietic stem cell function and cell cycle progression”. In: *Blood* 127.15 (2016), pp. 1881–1885.

[179] M. Walter, A. Teissandier, R. Pérez-Palacios, and D. Bourc’his. “An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells”. In: *eLife* 5 (2016), pp. 1–30.

[180] F. Wang, L. Qiao, X. Lv, A. Trivett, R. Yang, J. J. Oppenheim, D. Yang, and N. Zhang. “Alarmin human α defensin HNP1 activates plasmacytoid dendritic cells by triggering NF-κB and IRF1 signaling pathways.” eng. In: *Cytokine* 83 (July 2016), pp. 53–60.

[181] Q. Wang, H. Xiong, S. Ai, X. Yu, Y. Liu, J. Zhang, and A. He. “CoBATCH for High-Throughput Single-Cell Epigenomic Profiling”. In: *Molecular Cell* 76.1 (2019), 206–216.e7.

[182] R. Wang, D. Y. Lin, and Y. Jiang. “SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing”. In: *Cell Systems* 10.5 (2020), 445–452.e6.

[183] Y. Wang et al. “Clonal evolution in breast cancer revealed by single nucleus genome sequencing”. In: *Nature* 512.7513 (2014), pp. 155–160.

[184] Y. Wang et al. “Clonal evolution in breast cancer revealed by single nucleus genome sequencing”. In: *Nature* 512.7513 (2014), pp. 155–160.

[185] B. Wen, H. Wu, Y. Shinkai, R. A. Irizarry, and A. P. Feinberg. “Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells”. In: *Nature Genetics* 41.2 (2009), pp. 246–250.

[186] T. Wilhelm, M. Said, and V. Naim. “Dna replication stress and chromosomal instability: Dangerous liaisons”. In: *Genes* 11.6 (2020), pp. 1–35.

[187] S. J. Wu et al. “Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression”. In: *Nature Biotechnology* (2021).

[188] X. Xu et al. “Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor”. In: *Cell* 148.5 (2012), pp. 886–895.

[189] R. Yaeger et al. “Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer”. In: *Cancer Cell* 33.1 (2018), 125–136.e3.

[190] X. Yang, H. Han, D. D. DeCarvalho, F. D. Lay, P. A. Jones, and G. Liang. “Gene body methylation can alter gene expression and is a therapeutic target in cancer”. In: *Cancer Cell* 26.4 (2014), pp. 577–590.

[191] J. Yeung, M. Florescu, P. Zeller, B. A. de Barbanson, and A. van Oudenaarden. “Deconvolving multiplexed histone modifications in single cells”. In: *bioRxiv* (2021).

[192] S. Zaccaria and B. J. Raphael. “Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL”. In: *Nature Biotechnology* (2020).

[193] H. Zafar, N. Navin, K. Chen, and L. Nakhleh. “SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data”. In: *Genome Research* 29.11 (2019), pp. 1847–1859.

[194] H. Zafar, Y. Wang, L. Nakhleh, N. Navin, and K. Chen. “Monovar: Single-nucleotide variant detection in single cells”. In: *Nature Methods* 13.6 (2016), pp. 505–507.

[195] P. Zeller, J. Yeung, A. De Barbanson, H. Viñas Gaza, M. Florescu, and A. Van Oudenaarden. “Hierarchical chromatin regulation during blood formation uncovered by single-cell sortChIC”. In: *bioRxiv* (2021), p. 2021.04.26.440606.

[196] A. Zemach, I. E. McDaniel, P. Silva, and D. Zilberman. “Genome-wide evolutionary analysis of eukaryotic DNA methylation”. In: *Science* 328.5980 (2010), pp. 916–919.

[197] W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt. “Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements”. In: *Genome Biology* 16.1 (2015), pp. 0–19. arXiv: `1308.2134`.

[198] X. Zhang et al. “CUTseq is a versatile method for preparing multiplexed DNA sequencing libraries from low-input samples”. In: *Nature Communications* 10.1 (2019).

[199] W. Zhou, H. Q. Dinh, Z. Ramjan, D. J. Weisenberger, C. M. Nicolet, H. Shen, P. W. Laird, and B. P. Berman. “DNA methylation loss in late-replicating domains is linked to mitotic cell division”. In: *Nature Genetics* 50.4 (2018), pp. 591–602.

[200] C. Zong, S. Lu, A. R. Chapman, and X. S. Xie. “Genome-wide detection of single-nucleotide and copy-number variations of a single human cell”. In: *Science* 338.6114 (2012), pp. 1622–1626.

[201] E. C. Zook, K. Ramirez, X. Guo, G. van der Voort, M. Sigvardsson, E. C. Svensson, Y. X. Fu, and B. L. Kee. “The ETS1 transcription factor is required for the development and cytokine-induced expansion of ILC2”. In: *Journal of Experimental Medicine* 213.5 (2016), pp. 687–696.

[202] Zoonen, Renger, H. Schiessel, and J. V. Noort. “CpG Methylation in Nucleosome Positioning”. In: (2018).

Chapter 6

Addendum

6.1 Samenvatting

Voor zowel eencellige als complexe meercellige organismen geldt dat er geen enkele individuele cel gelijk aan een ander is. Bij de mens en andere dieren zijn weefsels samengesteld uit een verscheidenheid van celtypes met verschillende functies, die op hun beurt zijn samengesteld uit cellen die van elkaar verschillen op meerdere moleculaire niveaus. Onderzoek naar enkele cellen bestudeert de verschillen in feno- en genotype en de drijvende krachten achter de verschillen tussen afzonderlijke cellen. In alle cellen wordt voortdurend variatie geïntroduceerd, waardoor cellen steeds meer van elkaar gaan verschillen, en elke cel unieker wordt. Variatie tussen cellen kan geïntroduceerd tijdens celdelingen in de vorm van onder andere DNA-mutaties, veranderingen in DNA-methylering en door middel van histon modificaties. Ziekten zoals kanker kunnen beginnen in een enkele cel die door celdelingen snel divergeert van zijn voorouderlijke cel en aanleiding geeft tot een heterogene tumor.

Dit proefschrift bevat drie onderzoeks-hoofdstukken. De technologische basis voor elk hoofdstuk is het meten van meerdere modaliteiten in enkele cellen door middel van enkele cel sequencing technieken.

In het eerste hoofdstuk bestuderen we klonale dynamica in colorectale kanker door een colonkanker organoïde model systeem te laten evolueren gedurende een periode van 26 weken, waarbij we simultaan de klonale grootte, veranderingen in het aantal chromosomen, en enkelvoudige nucleotide varianten in individuele cellen in kaart brengen. De geïntegreerde metingen maken het mogelijk de volgorde van gebeurtenissen waarin chromosomale afwijkingen optreden te reconstrueren en maken het mogelijk veranderingen te vinden die meerdere malen in parallel binnen dezelfde populatie cellen zijn ontstaan. We observeren een terugkerend verlies van chromosoom 4, dat alleen voorkomt na verlies van chromosoom 18 en we laten zien dat dit overeenkomt met klinische waarnemingen in dikkedarmkanker-patiënten.

In het tweede hoofdstuk wordt een nieuwe techniek *scSort-ChIC* geïntroduceerd. Deze techniek kan worden gebruikt om histon modificaties te meten in enkele cellen, en dit kan worden gekoppeld aan FACS-informatie waarmee het celtype van individuele cellen kan worden bepaald. scSort-ChIC wordt gebruikt om actieve en repressieve histon modificaties in kaart te brengen in het proces waar bloed gevormd wordt (hematopoëse). *scSort-ChIC* wordt toegepast op zowel bloedstamcellen als volwassen bloedcellen in het beenmerg van de muis. Tijdens de differentiatie verwerven bloedstamcellen verschillende actieve chromatine toestanden die afhankelijk van de bestemming van de cel wordt geregeld door celtype specifieke transcriptie factoren. De meeste regio's op het genoom die tijdens de differentiatie repressieve histon-markeringen krijgen of verliezen, doen dit onafhankelijk van de celtype- bestemming van de cel. Het simultaan meten van de histon modificaties $H_3K_4me_1$ en $H_3K_9me_3$ toont aan dat celtypes binnen de myeloïde lijn verschillend actief chromatine hebben, maar gelijkaardig repressief chromatine dat specifiek is voor de myeloide lijn. Dit suggereert hiërarchische chromatine regulatie tijdens hematopoëse: het repressieve chromatine definieert differentiatie trajecten en afstamming, terwijl actief chromatine de celtypen bepaald.

In het derde hoofdstuk wordt een nieuwe techniek geïntroduceerd om zowel histon markeringen, DNA-methylatie en FACS-eigenschappen simultaan in een enkele cel te meten. Deze combinatie van metingen is nog nooit eerder uitgevoerd. De methode wordt grondig gevalideerd, en toegepast op een systeem waarbij de positie van elke cel in de celcyclus precies kan worden gemeten. Deze cel-cyclus informatie wordt gebruikt om gegevens van meerdere histonmarkeringen te integreren en hun gedrag gedurende de celcyclus te vergelijken. We vinden dat DNA-methylering in gebieden die bedekt zijn met nucleosomen langzamer hersteld wordt dan gebieden die vrij zijn van nucleosomen.

6.2 Curriculum Vitae

Buys de Barbanson was born on the 11th of April 1991 in Groningen, the Netherlands. After finishing high-school (VWO) in Warffum, he obtained a Bioinformatics bachelors degree from the Hanze Hogeschool Groningen. During his bachelors he was an intern at KeyGene NV in the Sequence Analysis group, led by Antoine Janssen. After obtaining his Bachelor of Applied Science degree, he enrolled in the Bioinformatics Master's program at the universities of Leiden and Delft. During his master program he was an shared intern in the Lab of Jeroen de Ridder and Alexander van Oudenaarden. In October 2016 he started his PhD at the Hubrecht Institute in the group of Prof. Dr. A van Oudenaarden. The results of the research are described in this thesis. After his PhD he founded Barbanson Biotech, a Bioinformatics consultancy company.

6.3 List of publications

Molecular characterization of Barrett's esophagus at single-cell resolution

Georg A. Busslinger, **Buys de Barbanson**, Rurika Oka, Bas L. A. Weusten, Michiel de Maat, Richard van Hillegersberg, Lodewijk A. A. Brosens, Ruben van Boxtel, Alexander van Oudenaarden, Hans Clevers *Proceedings of the National Academy of Sciences, 2021, 118 (47) e2113061118; DOI: 10.1073/pnas.2113061118*

Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies

Nico Battich, Joep Beumer, **Buys de Barbanson**, Lenno Krenning, Chloé S. Baron, Marvin E. Tanenbaum, Hans Clevers, Alexander van Oudenaarden *Science 367(6482):1151-1156 (2020) DOI: 10.1126/science.aax3072*

Eleven grand challenges in single-cell data science

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, **Buys de Barbanson**, *[..]* & Alexander Schönhuth
Genome Biology volume 21, Article number: 31 (2020) DOI:10.1186/s13059-020-1926-6

Snake Venom Gland Organoids

Yorick Post, Jens Puschhof, Joep Beumer, Harald M. Kerkkamp, Merijn A.G. de Bakker, Julien Slagboom, **Buys de Barbanson** *[..]* Alexander van Oudenaarden *[..]* Hans Clevers
Cell, Volume 180, Issue 2, 2020 DOI: 10.1016/j.cell.2019.11.038

Three-dimensional analysis of single molecule FISH in human colon organoids

Manja Omerzu, Nicola Fenderico, **Buys de Barbanson**, Joep Sprangers, Jeroen de Ridder, Madelon M Maurice
Biology Open 8:bio.042812 2019 DOI:10.1242/bio.042812

In preparation

Integration of multiple lineage measurements from the same single cell reconstructs parallel tumor evolution

Lennart Kester*, **Buys de Barbanson***, Anna Lyubimova, Li-Ting Chen, , Valérie van der Schrier, Anna Alemany, Dylan Mooijman, Josi Peterson-Maduro, Jarno Drost, Jeroen de Ridder and Alexander van Oudenaarden
Accepted with minor changes

Hierarchical chromatin regulation during blood formation uncovered by single-cell sortChIC

Peter Zeller*, Jake Yeung*, **Buys de Barbanson**, Helena Viñas Gaza, Maria Florescu, and Alexander van Oudenaarden
Under review

scChIC-TAPS reveals histone modification specific DNA methylation dynamics during the cell cycle

Christoph Geisenberger, **Buys de Barbanson**, Jeroen de Ridder and Alexander van Oudenaarden
In preparation

Simultaneous detection of full-length transcriptome and histone modifications

Peter Zeller, Marloes Blotenburg, **Buys de Barbanson**, Fredrik Salmén and Alexander van Oudenaarden
In preparation