From DNA sequences to cell types by detecting regulatory genomic regions in sequencing data

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Tommaso Andreani

geb. am 07.11.1986 in Pesaro

Mainz, 2019

Three o’clock is always too late or too early for anything you want to do

Jean-Paul Sartre

1. Abstract

One of the big questions in biology today is to understand which genetic and epigenetic factors are involved in the regulation of gene expression, and in which cases their deregulation can contribute to the development of abnormal phenotypes or diseases. Innovations in genome sequencing techniques and corresponding data processing algorithms have enabled unbiased interrogation of the different genomic and epigenomic components of transcription at nucleotide resolution. Therefore, it is now possible to use and integrate different types of data for both bulk and single-cell samples, and to understand the molecular components of gene expression regulation using ad-hoc reproducible computational analysis.

As an interdisciplinary field, bioinformatics takes advantage of different quantitative disciplines, such as statistics and machine learning. This allows the implementation of detailed analyses to support and elucidate specific fundamental discoveries, and also to test unexpected predictions coming from exploratory data analysis. In particular, the use of bioinformatics is a necessity in the study of the genomic basis of gene regulation given the complexity of the data produced. Thus, the application of existing and the development of novel bioinformatics methods improves the interpretation of new data by integrating several data types from multiple sources.

In this thesis I applied and developed bioinformatics methods to help investigate basic biological questions in the genomic study of epigenetic gene regulation: i) I created a pipeline for whole-genome bisulfite sequencing data analysis to improve the understanding of the way genes and DNA sequences are demethylated by GADD45 proteins and how this might be linked to a key stage of development in mouse embryonic stem cells (mESCs), ii) I developed a metric based on the Gini index to evaluate unsupervised clustering results obtained using several computational methods that were tested to identify various types of peripheral blood mononuclear cells (PBMCs) from single-cell ATAC-seq samples in which the labels of the cells were not provided and iii) I developed an algorithm to extract variable regions in ChIP-seq data that can improve the identification of target-specific binding sites of different proteins in several cell lines of the ENCODE project. Together, these three studies are a significant contribution to the improvement of the interpretation of genomic data for the study of epigenetic gene regulation by bioinformatics.

2. Zusammenfassung

Eine der aktuellen, großen Fragen der Biologie ist es zu verstehen, welche genetischen und epigenetischen Faktoren an der Regulation der Genexpression beteiligt sind und in welchen Fällen ihre Deregulierung zur Entwicklung von abnormalen Phänotypen oder Krankheiten beitragen kann. Innovationen bei Genomsequenzierungstechniken und entsprechenden Datenverarbeitungsalgorithmen ermöglichten objektive Analysen der verschiedenen genomischen und epigenomischen Komponenten der Transkription in der Auflösung von einzelnen Nukleotiden. Daher ist es jetzt möglich, verschiedene Daten sowohl für Bulk- als auch für Einzelzellproben zu integrieren und die molekularen Komponenten der Genexpressionsregulation zu verstehen, mithilfe einer reproduzierbaren rechnerischen ad-hoc-Analyse.

Als interdisziplinäres Feld nutzt die Bioinformatik verschiedene quantitative Disziplinen wie Statistik und maschinelles Lernen. Dies ermöglicht die Implementierung von detaillierten Analysen zur Unterstützung und Aufklärung spezifischer, fundamentaler Entdeckungen, sowie zur Prüfung unerwarteter Vorhersagen, die sich aus der explorativen Datenanalyse ergeben. Insbesondere ist die Bioinformatik notwendig für die Untersuchung der genomischen Grundlagen der Genregulation angesichts der Komplexität der erzeugten Daten. Die Anwendung bestehender und die Entwicklung neuer bioinformatischer Methoden verbessern die Interpretation neuer Daten zu, indem verschiedene Datentypen aus mehreren Quellen integriert werden.

In dieser Dissertation habe ich bioinformatische Methoden angewendet und entwickelt, um grundlegende biologische Fragen in der genomischen Erforschung der epigenetischen Genregulation zu untersuchen: i) habe ich eine Pipeline für die Datenanalyse der Bisulfitsequenzierung im gesamten Genom erstellt, um zu verstehen, wie Gene und DNA-Sequenzen von Gadd45-Proteinen demethyliert werden und wie dies mit einem der wichtigsten Entwicklungsstadien in embryonalen Maus-Stammzellen (mESCs) zusammenhängt, ii) entwickelte ich eine Metrik auf der Grundlage des Gini-Index, um die Ergebnisse von unüberwachten Clusterings von verschiedenen Berechnungsmethoden zu bewerten, die zur Trennung von periphere mononukleäre Blutzellen (PBMCs) von einzelzell-ATAC-seq-Proben angewendet wurden, von denen die Markierungen der Zellen nicht vorhanden waren, und iii) habe ich einen Algorithmus entwickelt, mit dem variable Regionen in ChIP-seq-Daten extrahiert werden können, um die Identifizierung Protein-spezifischer Bindungsstellen in mehreren Zelllinien des ENCODE-Projekts zu verbessern. Zusammen sind diese drei Studien ein signifikanter Beitrag durch die Bioinfromatik zur Verbesserung der Interpretation von Genomdaten in Hinblick auf die Untersuchung der epigenetischen Genregulation.

3. Introduction

3.1 Molecular basis of epigenetic gene regulation

The genetic information of living organisms is composed of nucleic acid molecules (known individually as nucleotides, which together form DNA) that are packed together with different types of histone proteins; this compact complex of DNA and proteins is called chromatin. A long-standing question of molecular biology research today has been how chromatin can de-condenses and become accessible by transcriptional machinery to promote gene expression. Several mechanisms were proposed to link higher-order chromatin structure to gene expression activation, but the conclusions were mostly driven by association analysis and, therefore, lack real mechanistic insights. For example, during transcription, various proteins and regulatory DNA elements combine together to maintain the physiological expression of genes, but how this is coordinated in the cell for every gene is still not understood. A popularly accepted hypothesis is the interaction between enhancers and promoters at accessible chromatin regions that allowed the development of several predictive computational methods (SIngh et al. 2019, Talukder et al. 2019, Zeng et al. 2018, Belokopytova et al. 2019, Gao et al. 2019).

While DNA is supposed to remain stable and immutable throughout the lifetime of a cell, epigenetic modifications at the nucleotide or histone protein level are reversible and dynamic (Poetsch et al. 2011, Bradburi 1992, Ramchadani et al. 1999). Therefore, in order to better understand the promotion of gene expression, it is important to elucidate how the deposition of these modifications into the genome occurs as well as the molecular process involved in their removal. The role of epigenetics in the development of multicellular organisms was proposed several years ago when pioneering studies demonstrated developmental arrest in mice lacking key enzymes such as DNA methyltransferase (DNMT) involved in DNA methylation or Polycomb Repressive Complex 2 Subunit SUZ12, which methylates 'Lys-9' (H3K9me) and 'Lys-27' (H3K27me) of histone H3 (Butler et al. 2012) and is involved in chromatin silencing. Studies like these have motivated the development of sequencing techniques capable of detecting all the genomic regions modified by such enzymes in an unbiased manner. Sequencing technologies such as ChIP-seq, ATAC-seq and Bisulfite-seq allow the investigation of the epigenome and the regulatory elements such as promoters, enhancers, and insulators that promote or inhibit gene expression.

The first layer of epigenetic gene regulation can be found directly on the DNA, when a methyl group is attached to the fifth carbon of a cytosine, in a process called methylation (5-methylcytosine (5mC)). This DNA modification is particularly abundant in mammals, especially in regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence

of bases along its 5' → 3' also named CpG. Approximately 70–80% of CpGs show a methyl group at the carbon five of the cytosine (Greenberg & Bourc'his, 2019). In the second layer, DNA accessible regions, which are usually enhancers or promoters, can be directly bound by transcription factors and other regulatory proteins involved in gene expression. The following layer is comprised of histone proteins. Histones are found in eukaryotic cells and consist of eight subunits, which are known as histone octamers. Approximately 145–147 base pairs (bp) of DNA are wrapped in 1.8 helical turns of a histone octamer of four highly evolutionarily conserved histone proteins: H2A, H2B, H3 and H4. Histones are the fundamental unit of DNA packaging in the genome and are named nucleosomes when wrapped with DNA. The final or highest layer of epigenetic gene regulation is made up of a group of proteins involved in post-translational modifications of the histone tails. Methyl groups are added to the lysine or arginine "tails" that protrude from the histone proteins, which then affects how loosely or tightly the DNA is wrapped around it, thereby determining the chromatin structure and DNA accessibility for transcription. The proteins adding the methyl-groups to histone tails are named Trithorax group (TrxG) and Polycomb (PcG) and they can modulate the accessibility to the DNA via repressive or activating mechanisms, are involved in the maintenance of cellular identity, and are required for normal differentiation across various species such as Drosophila and Mouse (Li et al. 2016, Aloia et al. 2013). It is also important to mention the role of RNAs in epigenetic gene regulation. Their role can be either at the DNA level, in which long RNA sequences that do not undergo translation named long non-coding RNA (lncRNA) are recruited at CpG sites to promote demethylation (Arab et al. 2014, Arab et al. 2019) or at the chromatin level, where they can remodel the chromatin to regulate the expression of nearby genes (Zhang et al. 2016). In this thesis, I will focus on the first two layers of epigenetic gene regulation: DNA methylation and DNA accessibility.

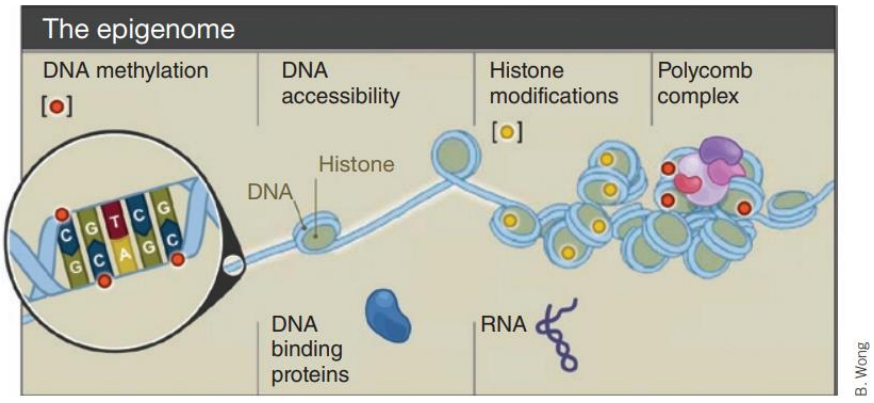


Fig. 1 Layers of genome organization. Genome function and cellular phenotypes are influenced by DNA methylation and the protein-DNA complex known as chromatin. In mammals, DNA

methylation occurs on cytosine bases, primarily in the context of CpG dinucleotides. Accessible chromatin that is hypersensitive to DNase I digestion marks promoters and functional elements bound by transcription factors or other regulatory proteins. Histone modifications, associated proteins such as Polycomb repressors and noncoding RNAs constitute an additional layer of chromatin structure that affects genome function in a context-dependent manner (Bernstein, 2010).

3.1.1 Promoters and distal regulatory elements

The transcription of prokaryotic and eukaryotic genes requires the presence of DNA sequences and protein complexes, which both together are necessary to initiate gene expression. These DNA sequences are called promoters and are 100-1000 base pairs (bp) in length and are located in the vicinity of the transcription start site of genes. Promoters play an important role to activate gene expression and they differ in nucleotide composition and type of proteins that bind to them. Among the various class of proteins involved in the process of transcription, transcription factors are proteins involved in converting, or transcribing, DNA into RNA. Transcription factors include a wide number of proteins, excluding RNA polymerase, that initiate and regulate the transcription of genes. One distinct feature of transcription factors is that they have DNA-binding domains that give them the ability to bind to specific sequences of DNA called enhancer or promoter.

In prokaryotic cells, a promoter requires an RNA polymerase, the sigma factor as well as the consensus sequences “TATAAT” and “TTGACA” to trigger gene activation. In eukaryotic cells, several other accessory proteins, i.e. the general transcription factors (GTF) complex, are required to initiate transcription. According to the type of gene, eukaryotic promoters are bound by different RNA polymerases. For example, RNA polymerase I is required to transcribe genes encoding 18S, 5.8S, and 28S ribosomal RNAs; RNA polymerase II is required to transcribe genes encoding messenger RNA and certain small nuclear RNAs; RNA polymerase III is required to transcribe transfer RNA, 5s ribosomal RNAs, and other small RNAs. Independent of the type of gene, there are different consensus sequences that allow the polymerase to recognize the transcription start site. Among these, TATA box and the B recognition elements (BRE) are two of the most well-characterized sequences. The first consensus sequence is found in the core promoter region of archaea and eukaryotic genes, while the latter is a cis-regulatory element that is found immediately near the TATA box, and consists of 7 nucleotides usually bound by the transcription factor II B (TFIIB). BRE and TATA box have various effects on transcription. Several other characteristic sequences are typical of eukaryotic promoters. In fact, most of the mammalian promoters are CpG dinucleotide rich and multiple structures that are CG rich, such as G-quadruplex and R-loop, can form during transcription.

Promoters can also work in cooperation with other classes of cis-regulatory elements, named enhancers. These DNA elements are located in great

distance to the transcription start site, from 1 kilobase (kb) up to 1 million bp, and play an important role when associated to the promoter of a gene. One of the best-characterized loci is the mouse β-globin that contains a cluster of β-chain variants of haemoglobin that are developmentally regulated by multiple elements spanning a region of 100 kb. These regulatory elements include the locus control region (LCR), which acts as an enhancer and is located approximately 25 kb upstream of the εy-globin gene (HBE1), and a group of DNase I hypersensitive sites located approximately 20 kb downstream of the locus (3′HS1) and upstream of the LCR (–60HS). Expression of the β-globin gene (HBB) requires the presence of specific transcription factors (erythroid transcription factor (GATA1) and Krüppel-like factor 1 (KLF1)) that mediate the clustering of these elements with the β-globin gene promoters. This occurs through long-range interactions to form what has been described as an 'active chromatin hub' (Tolhuis et al. 2002).

Enhancers are usually located at low methylated regions in the genome (Stadler et al. 2011) and can harbour different DNA modifications that can be the product of oxidative DNA demethylation (Shen et al 2013). Enhancer DNA demethylation is mediated by the Ten eleven translocation (TET) enzymes in mESCs (Lu et al. 2014) and by the growth arrest DNA damage 45 (GADD45) proteins in mouse embryonic fibroblasts (MEF) (Shen et al 2013, Lu et al. 2014, Schafer et al. 2018).

3.1.2 DNA methylation

DNA methylation is one of the best-studied epigenetic marks. Due to its prevalence within mammalian genomes, the DNA modification "5mC" is considered the fifth nucleic acid-base. It is implicated in the epigenetics of genomic imprinting and X-chromosome inactivation and is highly abundant at CpG sites. Classically, methylation at CpG sites is associated with transcriptional repression. Individual mutations in DNA methyltransferase (DNMT) genes lead to irreversible DNA methylation defects as also pathological phenotypes, early embryonic lethality, sterility, developmental syndrome and cancer (Zhang et al. 2017).

There are different classes of DNMT enzymes in mammals, which play different functional roles in epigenetic regulation. DNMT1 is involved in the maintenance of DNA methylation by methylating the hemimethylated DNA strands after DNA replication. DNMT3a and DNMT3b are de-novo methyltransferases (Okano et al. 1999). Additionally, DNMT3a works on maintaining DNA methylation, suggesting a redundant role for this type of enzyme (Feng et al. 2010). DNMT3L is a nuclear protein with similar structural characteristics to DNA methyltransferases but does not act as a DNA methyltransferase since it does not contain the amino acid residues

necessary for methyltransferase activity. However, it does stimulate *de novo* methylation by DNMT3a is required for the establishment of maternal genomic imprints. Recently, a methyltransferase from the *Gm14490* gene was discovered (Barau 2016). This gene was identified as a pseudogene but was eventually characterized as a genuine functional gene that encodes DNMT3C–a novel *de novo* DNA methyltransferase. In contrast to other DNMTs, this enzyme is distinguished for its selectivity in methylating the promoters of evolutionarily young retrotransposons only in the context of fetal spermatogenesis. DNMT3A/B are not able to compensate for the function of DNMT3C.

3.1.3 DNA demethylation

DNA methylation is a reversible process, which is called DNA demethylation. There are two types of DNA demethylation: global demethylation that happens broadly throughout the whole genome and locus-specific demethylation that occurs within particular sequences in a context-dependent manner. The loss or removal of 5mC can occur in different ways. One way is considered to be a passive mechanism: 5mC can be lost during DNA replication when DNA methylation maintenance machinery is absent. Another way is regarded as an active mechanism: an enzymatic process that removes or modifies the methyl group from 5mC via TET enzymes. These enzymes are methylcytosine dioxygenases and iteratively oxidize 5mC to 5-hydroxymethylcytosine (5hmC), 5-formyl-cytosine (5fC) and 5-carboxylcytosine (5caC). All the oxidized forms can promote DNA demethyl-ation during replication. Additionally, in the case of 5fC and 5caC, an enzyme by the name of thymine DNA glycosylase (TDG) can remove the nitrogen base, which is then followed by the activity of the base excision repair pathway (Kohli et al. 2013). TDG activity can be boosted by other DNA repair proteins, such as the Small Ubiquitin-like Modifier (SUMO) (Steinacher et al. 2019) and Neil DNA glycosylase (Schomacher et al. 2016). In particular, Neil protein enhances the removal of 5fC and 5caC by promoting the substrate turnover to TDG. Along this line, another class of enzymatic adapters, named GADD45, plays an important role in active DNA demethylation mediated by DNA repair. During active demethylation, GADD45 has a double function: it enhances the removal of 5fC and 5aC when interacting with TDG (Cortellino 2011, Li et al. 2015), and it boosts TET activity and as a result 5mC to 5hmC oxidation (Kienhöfer et al. 2015).

The activation-induced cytidine deaminase (AID) enzyme has also been reported to deaminate 5-methylcytosine to thymine. This creates a T:G mismatch, which can be repaired by the activity of TDG through DNA repair pathways (Cortellino et al. 2011). In mESCs AID was shown to be involved in DNA demethylation together with TET1, MBD4 and GADD45a. This demethylation system was reported to be not necessary for generating the overall bimodal methylation pattern of mESCs but appeared to be involved in

resetting methylation patterns during somatic-cell reprogramming (Sabag et al. 2014).

DNA methylation is dynamic during early development. Following fertilization, the global loss of DNA methylation in the maternal and paternal genomes corresponds with the gaining of totipotency. At this stage of the development, the maternal genome (red pronucleus; red line, Fig 2) undergoes passive DNA demethylation throughout several rounds of DNA replication. The paternal genome (blue pronucleus; blue lines, Fig 2) undergoes active demethylation before DNA replication (Gu et al. 2011) in the zygote ensues, although it is debatable how this occurs. In fact, it has been shown that the initial loss of paternal 5mC does not require 5hmC formation (Amouroux et al. 2016). Concomitant with global loss of paternal 5mC, 5hmC (blue dotted line) and the further oxidation derivatives (5fC and 5caC; blue dashed line Fig. 2) are enriched.

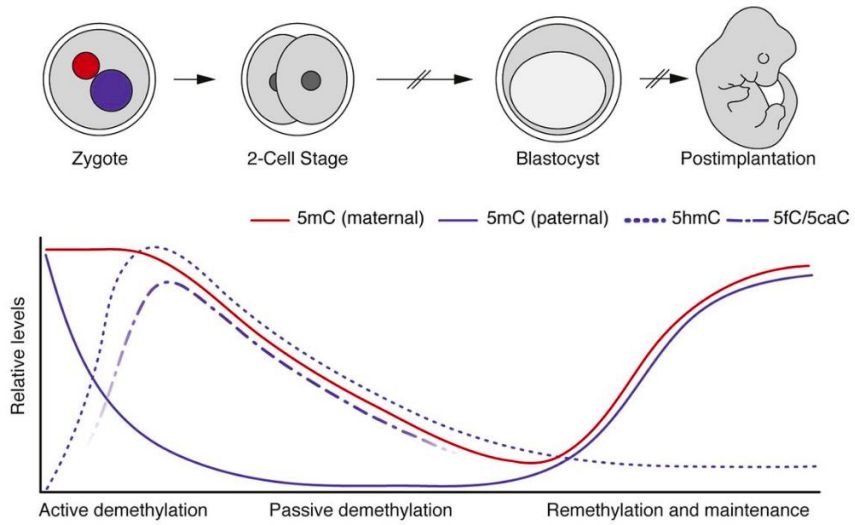


Fig. 2. DNA demethylation dynamics and maintenance in preimplantation and postimplantation embryos. Although selected loci are restored to unmodified cytosines, the bulk of paternal 5hmC is passively diluted, paralleling demethylation of the maternal genome (Messerschmidt et al., 2014).

During the transition to the primed pluripotent state, *de novo* methylation is responsible for a global gain of this epigenetic mark and the cells loose fate potential (Auclair et al., 2014; Seisenberger et al., 2012; Smith et al., 2012) (Fig. 2). Furthermore, *de novo* methylation occurs also post-implantation.

3.1.4 Other DNA modifications

Epigenetics refers to a variety of processes that have heritable effects on gene expression programs without changes in DNA sequence. Key players in epigenetic control are the already described chemical modifications to DNA, such as the well-known 5mC and its TET oxidative derivative products, 5hmC, 5fC, and 5caC. DNA modifications can have regulatory potential if they can affect particular steps of gene expression and can be found in the form of DNA damage.

A novel DNA modification in the form of DNA damage was found in mESC by monitoring the oxidative activity of TET enzymes on thymines, which formed 5-hydroxymethyluracil (5hmU) (Pfaffeneder et al. 2014). This DNA modification has been reported to affect protein-binding to DNA (Greene et al. 1986) and may also be a key intermediate in the generation of site-specific mutations, as it can be excised by DNA glycosylases to create potentially mutagenic abasic sites (Kawasaki et al. 2017).

Another DNA modification on adenines, named N6-Methyladenine, has also been reported after monitoring the activity of two enzymes: methyltransferase N6AMT1, which deposits this modification, and demethylase ALKBH1, which removes it (Xiao et al. 2018).

8-Oxoguanine (8-oxoG) is found in the form of DNA lesion, caused by reactive oxygen species (ROS) modifying guanine. During DNA oxidation, 8-oxoG can result in a mismatched pairing with adenine, which can subsequently cause G to T and C to A substitutions in the genome. In humans, it is primarily repaired by (8-oxoguanine-DNA glycosylase) OGG1. Even if 8-oxoG is known as a DNA lesion, it has been reported to have an epigenetic function in the activation of murine splenocytes (Seifermann et al. 2017).

Approximately more than 40 DNA modifications documented as DNA damages are documented across the entire spectrum of living organisms most of which have neither been characterized with genome-wide sequencing techniques nor have been associated with a particular regulatory role or disease predisposition. Recently, the DNAmod database was developed to collect all the currently known potential DNA modifications characterized in all species even those observed using synthetic approaches (Sood et al. 2019). Thus, several other DNA modifications similar to 5mC and its oxidative derivative products might have important roles in gene regulation and embryonic development.

3.2 Gadd45 protein family

The Growth Arrest and DNA Damage-inducible 45 (GADD45) proteins are stress sensors that are activated under various stimuli. There are three different genes coding for these isoforms, which are located in different chromosomal locations along the mouse genome: *Gadd*45a is located at chromosome 6 and its protein product is 165 amino acids in length; *Gadd*45b–initially named *Myd118*–is located at chromosome 10 and its protein product is 145 amino acids in length; *Gadd45g*–initially named *Ddit2*–is located at chromosome 13 and its protein product is 159 amino acids in length. All three GADD45 proteins are highly acidic and can be located both nuclear and cytoplasmic. The first functional role attributed to the *Gadd45* genes was discovered in mammals, where its expression increased after growth cessation signals or treatment with DNA-damaging agents (Hildesheim et al. 2002). GADD45 proteins do not have any enzymatic function but they usually work as adapters for effectors proteins. Since their first discovery, abundant research on the *Gadd45* genes and protein family has demonstrated its importance in environmental stress response and epigenetic gene regulation at a molecular level.

3.2.1 Gadd45 proteins role & function

The expression of all *Gadd45* genes is induced upon exposure to several genotoxic or oxidative agents and the different family members are implicated in a variety of responses, i.e. cell cycle checkpoints, apoptosis, and DNA repair (reviewed in Liebermann and Hoffman 2008, Tamura et al. 2012). In a stress response, p53 is activated, which causes transcriptional upregulation of *Gadd45* that acts as a stress response protein and interacts with MTK1 to initiate the JNK/p38-mediated apoptotic pathway (Chi et al. 2004). GADD45 proteins also mediate cell-cycle regulation through interactions with PCNA (Kelman et al. 1998, Azam et al. 2001), the cyclin-dependent kinase inhibitor p21 (Kearsey, 1995), and the Cdk/cyclin B complex (Zhan et al. 1999; Jin et al. 2002, Vairapandi et al. 2002).

Several model systems were used to investigate the role of GADD45 proteins during development. Gadd45a-null mice exhibit several abnormalities including genomic instability, increased radiation carcinogenesis, and a low frequency of exencephaly (Hollander et al. 1999) and Gadd45γ-deficiency in mice causes male to female sex reversal (Gierl et al. 2012). Gadd45αγ *Xenopus* morphants show reduced neural cell proliferation and downregulation of pan-neural and neural crest markers (Kauffman et al. 2011). All these observations suggest that GADD45 proteins are important for cellular differentiation during development.

The molecular roles of GADD45 proteins are connected to specific phenotypic outcomes. Adult mouse brain cells lacking Gadd45β exhibit specific deficits in neural activity and exhibit reduced proliferation of neural progenitors and dendritic growth of newborn neurons in the adult hippocampus (Ma et al. 2009). Mechanistically, Gadd45β is required for activity-induced DNA demethylation of specific promoters and expression of corresponding genes critical for adult neurogenesis, including brain-derived neurotrophic factor and fibroblast growth factor (Ma et al. 2009). In the context of ageing, double-knockout mice for *Gadd45a/Ing1* exhibit phenotypic characteristics of progeria. In fact, mice lacking both proteins exhibit reduced life span, kyphosis, weight reduction, ovarian atrophy, female infertility, bone marrow fattening, and skin senescence. All of these phenotypic characteristics are partially attributed to de-regulation of active DNA demethylation at C/EBP (CCAAT/enhancer-binding protein)-dependent super-enhancers in mouse embryonic fibroblasts (Schafer et al. 2018).

3.2.2 Gadd45 proteins in active DNA demethylation

GADD45 proteins play an important role in active DNA demethylation in order to regulate gene expression (Barreto et al. 2007, Schafer et al. 2018, Arab et al. 2014 and Arab et al. 2019). This role was first identified in a screen, based on activation of a methylated reporter plasmid gene. Further analysis in *Xenopus* showed that GADD45α interacts with the nucleotide excision repair (NER) factor XPG to promote DNA repair and hence mediating DNA demethylation (Barreto et al. 2007). Other studies supported the role of GADD45 proteins in locus-specific DNA demethylation. For example, during the demethylation of rDNA-coding genes, Gadd45α is recruited at the rDNA promoters by TAF12. At the chromatin level, Gadd45α can be recruited by Ing1 at H3K4me3 -containing loci to promote gene-specific DNA demethylation (Schafer et al. 2013). GADD45 is also an RNA binder and can interact with long non-coding RNAs to promote demethylation. For example, GADD45α binds the long non-coding RNA *TARID* to facilitate demethylation of the *Tcf21* locus (Arab et al. 2014). Furthermore, Gadd45α functions as an R-loop reader and recruits TET to CGIs promoters (Arab et al. 2019).

GADD45α was reported to co-localize with TET enzymes in the nucleus (Kienhöfer et al 2015), giving further evidence of its proteins' dual role in enhancing both the oxidative activity of TET and the DNA repair activity of TDG (Fig. 3).

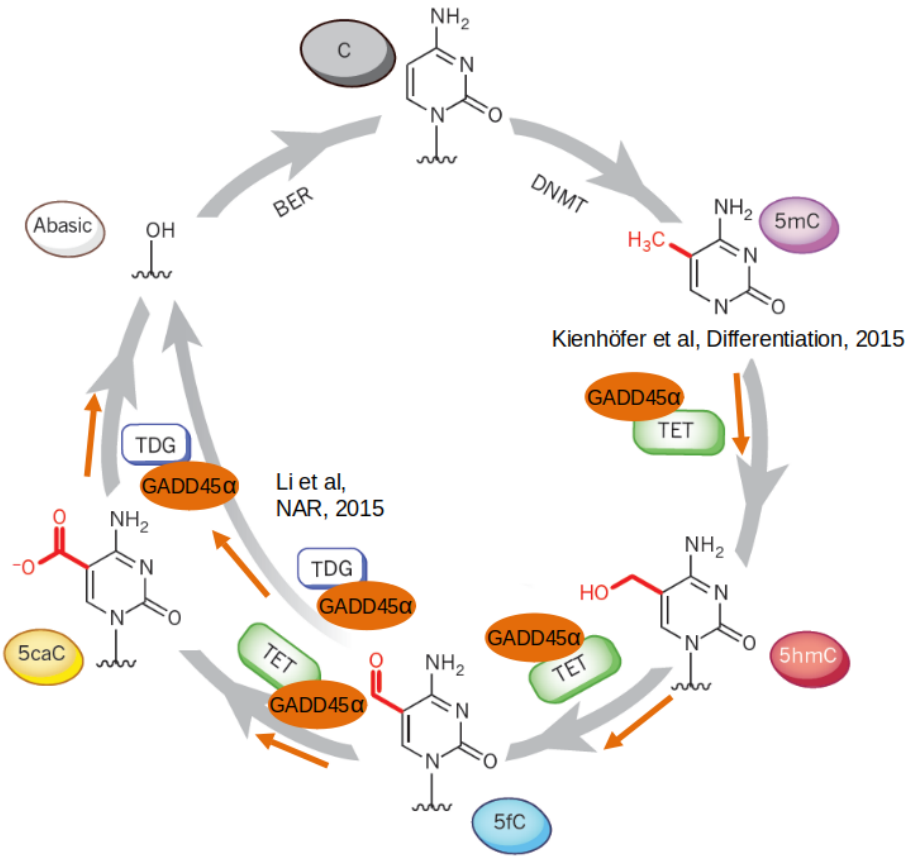


Fig. 3 A biochemically validated pathway for modification and de-modification of C within DNA is shown. 5mC bases, introduced by DNA methyltransferase (DNMT) enzymes, can be oxidized iteratively to 5hmC, 5fC and 5caC. In the pathway of active modification (AM) followed by passive dilution (PD), 5hmC is diluted in a replication-dependent manner to regenerate unmodified C. For clarity, PD of highly oxidized 5hmC, 5fC and 5caC is not depicted. In the pathway of AM followed by active restoration (AR), 5fC or 5caC is excised by Tdg, generating an abasic site as part of the base excision repair (BER) process that regenerates unmodified C. In orange is represented the role of Gadd45. (Adapted from Kohli et al. 2013).

In the context of gene regulation, TET enzymes are required to promote 5mC oxidation at enhancers and genomic sites harbouring TET oxidative DNA products such as 5hmC, 5fC, and 5caC (Lu et al 2014). As previously mentioned, ING1 functions in DNA demethylation by directing GADD45α to H3K4me3, a chromatin mark that is typically found at promoters. Mouse embryonic fibroblasts of Gadd45α/Ing1 knockout mouse showed DNA hypermethylation at C/EBP (CCAAT/enhancer-binding protein)-dependent super-enhancers. Thus, promoter-enhancer interactions might be governed by mechanisms that promote TET-mediated active DNA demethylation and this interaction could be a potential mechanism that involves the demethylation of mC at CG regions to promote gene expression activation.

A specific structure of CG-rich protomers is an R-loop, a three-stranded nucleic acid structure composed of a DNA:RNA hybrid with an associated non-

template single-stranded DNA. When an R-loop is located at the promoter of a CGIs rich gene, it functions as a molecular barcode for GADD45α. In the case of specific genes, the interaction between GADD45α and the R-loops can recruits TET enzymes to trigger active DNA demethylation (Arab et al., 2019).

Overall, active DNA demethylation potentially regulates the expression of genes involved in the immediate response to external stimuli during development as also during terminal osteogenic differentiation of adipose-derived mesenchymal stem cells (Zhang et al. 2011). The interaction of GADD45 with both TET and the base excision repair factor TDG demonstrates its dual ability in promoting active demethylation (Kienhöfer et al. 2015, Cortellino 2011, Li et al. 2015). Its role in epigenetic gene regulation is shown phenotypically in fear memory formation, aversive learning and ageing (Li et al 2019, Rey et al. 2019, Schafer et al. 2018).

3.3 Mouse embryonic stem cells (mESCs) as a model of early embryonic development

Mouse embryonic stem cells (mESCs) are pluripotent stem cells that can be obtained from the inner cell mass (ICM) of the blastocyst. In mice, a blastocyst forms 3.5 days after fertilization when a first differentiation event occurs: the outer layer of cells commits the trophectoderm (TE) becoming part of the placenta and separates from the ICM (Fig. 4).

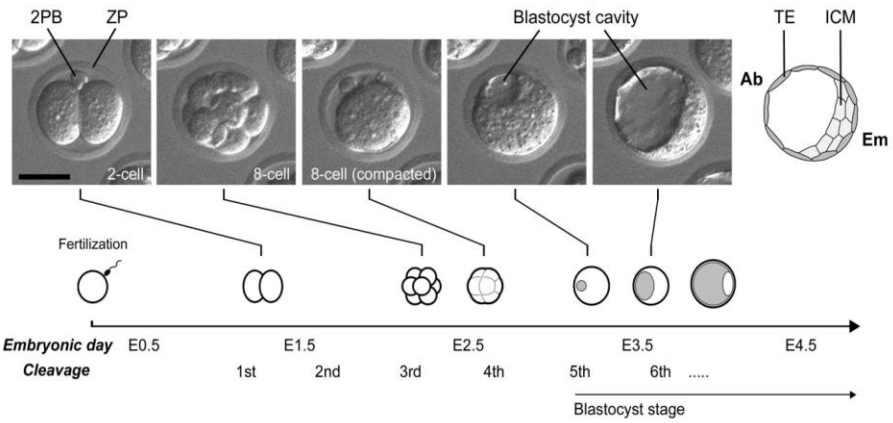


Fig. 4 Different stages of mouse preimplantation development during the first five days after fertilization. The embryonic day (E) is used to describe the developmental stage of embryos, which correspond to days after mating of parents. The timing of developmental progression in this diagram is based on embryos that are cultured in vitro. 2PB, second polar body; ZP, zona pellucida; TE, trophectoderm; ICM, inner cell mass; Ab, abembryonic side; Em, embryonic side. Scale bar, 50 µm. (Marikawa and Alarcón, 2009)

mESCs extracted from the ICM can give rise to any cell type of the body and diminish as they differentiate to other cell types with more limited developmental potential. mESCs are a useful resource to study the biology of early developmental processes, in fact, several studies have shown that in vitro experiments resemble the in vivo situation of early embryonic development (Macfarlan et al. 2012). Thus, mESCs are a good model to study cellular differentiation during the early stages of embryonic development as also pluripotency in particular cell culture conditions.

Pluripotent mESCs can give rise to progeny representing all types of cells in an organism. This was confirmed in a study in which chimeric mice were generated by injecting mESCs into either eight cell-stage embryos or blastocysts. The injected cells contributed to all embryonic cell lineages, including germ cells (Bradley et al. 1984). This ability to generate chimeric organisms is now regarded as one of the hallmarks of ‘‘naïve’’ pluripotency (Nichols & Smith 2009), which is distinguished from the ‘‘primed’’ pluripotency exhibited by epiblast stem cells (EpiSCs) in which the cells are ready for lineage specification and commitment in response to stimuli (Brons et al. 2007; Tesar et al. 2007).

The cell culture medium is an important factor to maintain the cells pluripotent or to lead them to differentiate. To keep mESCs in the ‘‘naïve’’ pluripotent state in serum-free-like conditions two small-molecule kinase (Mek and GSK3) inhibitors (“2i” - PD0325901 and CHIR99021) are used. Leukemia inhibitory factor (LIF) is an interleukin 6 class cytokine that affects cell growth by inhibiting differentiation. When LIF levels drop, the cells differentiate. The addition of 2i to LIF medium prevents pluripotent cells from differentiation. Conversely, it is possible to induce the differentiation of mESCs into various lineages with specific medium supplementation. In the absence of LIF and upon overexpression of Gata4 or Gata6, mESCs can differentiate into extraembryonic endoderm cells (XEN) (Fujikura et al. 2002). Another example was reported for Oct3/4-null mESCs differentiating into Trophectoderm (TE). The Oct3/4 gene encodes a transcription factor specifically expressed in embryonic stem cells (Okamoto et al. 1990; Rosner et al. 1990; Scholer et al. 1990), and its downregulation in mESCs exclusively induces their differentiation into TE cells (Niwa et al. 2000). In another study it has been shown that late epiblast cells and primitive ectoderm cells can give rise to primed pluripotent stem cell lines (EpiSCs) in culture containing FGF2, Activin and MEF feeder (Brons et al. 2007; Tesar et al. 2007).

Cultures of mESCs contain a rare cell population named “2C-like” cells with characteristics similar to the two-cell embryo. These cells display molecular features of totipotent cleavage-stage cells and can even contribute to both embryonic and extraembryonic derivatives, including trophoblast, in chimeras. 2C-like cells can be used as a model of 2-cell stage embryo development. Characteristics of 2C-like cells are the high expression of the leucine tRNA

primer (MERVL) retrotransposon and zinc finger and SCAN domain-containing protein 4 (Zscan4).

The methylation dynamics during early development can be studied in mESCs. In fact, mESC cultures show different 5mC signatures. Male 2i mESCs are globally hypomethylated compared to conventional mESCs maintained in serum. In serum, female mESCs are hypomethylated similarly to male mESCs in 2i, and DNA methylation is further reduced in 2i. The methylome of male mESCs in serum parallels postimplantation blastocyst cells, while 2i stalls mESCs in a hypomethylated, ICM-like state (Habibi et al., 2013).

3.3.1 mESC regulation of pluripotency & Differentiation

Transcription factor binding sites (TFBSs) and DNA methylation play an important role in embryonic development. In mESCs, transcription factors (TFs) such as Oct4 (also known as Pou5f1), Nanog and Sox2 are involved in the maintenance of pluripotency. These TFs regulate a substantial portion of target genes that frequently encode transcription factors, many of which are developmentally important homeodomain proteins (Boyer et al. 2005).

Oct4 and Sox2 proteins, also orchestrate germ layer fate selection. Oct4 suppresses neural ectodermal differentiation and promotes mesendodermal differentiation; Sox2 inhibits mesendodermal differentiation and promotes neural ectodermal differentiation. Differentiation signals continuously and asymmetrically modulate Oct4 and Sox2 protein levels, altering their binding pattern in the genome, and leading to cell fate choice (Thomson et al. 2011).

DNA methylation plays an important role in the regulation of mESCs differentiation (Dawlaty et al. 2014; Jackson et al. 2004; Lei et al. 1996; Okano et al. 1999; Sakaue et al. 2010). In fact, naive mESCs were found to have minimal methylation variability at enhancers, whereas primed mESCs exhibited 17%-86% methylation variability at enhancer elements located at the H3K4me1 chromatin mark. This methylation variability was independent of the cell cycle, suggesting that DNA methylation variance at enhancers is a unique feature of primed pluripotency. Consequently, the oscillation in the expression of TET and DNMT enzymes is observed for the proper regulation of differentiation and pluripotency in mESCs (Rulands et al., 2018).

3.3.2 mESC and two-cell stage development

As previously explained, a characteristic of mESCs is their capability to differentiate into any type of tissue or remain pluripotent. Additionally, mESCs cultures contain a transient group of cells, named 2C-like cells. These cells express high levels of transcripts found in two-cell (2C) embryos, in which the blastomeres are totipotent (Macfarlan et al., 2012) have the ability to

contribute to both embryonic and extraembryonic tissues. Other gene expression signatures of 2C-like mESCs, that are also present in two-cell embryos, are the MuERV-L family of retroviruses and their corresponding long terminal repeat (LTR) promoters (Mt2_mm). Furthermore, other previously characterized genes, whose expression is restricted to the 2–4-cell stage of development, are also expressed in the 2C-like cells. These include Zscan4, Tcstv1/3, Eif1a, Gm4340 (also known as Thoc4), Tdpoz1–5 and Zfp352 (Macfarlan et al., 2012).

The ability of mESCs to transiently express features similar to 2C stage embryos allows them to test candidate proteins potentially involved in the regulation of this developmental stage. The discovery of the DUX transcription factor exemplifies this well. Overexpression of Dux in mESCs promotes the 2C-like state (De Iaco et al. 2017, Hendrickson et al. 2017). DUX transcription factor binds to the promoters of 2C stage-specific transcripts and activates the expression of 2C stage marker genes such as ZSCAN4 and the retroviral elements called the MERVL/HERVL family. Furthermore, Dux knockout in mESCs prevents the cells from cycling to the 2C-like state, and zygotic depletion of Dux leads to impaired early embryonic development and defective zygote genome activation (ZGA) (De Iaco et al., 2017). Conversely, Dux zygotic KO (Z-KO) mice and maternal and zygotic KO (MZ-KO) embryos can survive to adulthood despite showing reduced developmental potential. Furthermore, MZ-KO embryos revealed that loss of DUX has minimal effects on ZGA (Chen et al. 2019).

In another study, the function of the developmental pluripotency-associated 2 (DPPA2) and DPPA4 proteins was identified as potential positive regulators of ZGA gene expression and thus of the 2C-like and 2 cell stage development. Specifically, DPPA2 and DPPA4 were found to be upstream regulators of Dux by binding to the Dux promoter and gene body to drive its expression and initiate the expression of 2C stage-specific transcripts such as Zscan4c (Eckersley-Maslin et al. 2019). Another study confirmed the role of DPPA2 and DPPA4 in activating Dux and thereby regulating the 2C-like state by modulating the expression of the young LINE-1 elements, whose mechanisms remain elusive (De Iaco et al. 2019).

3.4 Sequencing technologies to study epigenetic gene regulation

The different molecular layers that contribute to the epigenetic regulation of gene expression are of fundamental importance for the understanding of how cells and living organisms can perform their function. In the last decade, improvements in the development of genome-wide sequencing technologies

have enabled to sequence most of the molecular features that are linked to epigenetic regulation of gene expression. Sequencing techniques capable to identify in an unbiased way protein-DNA interactions (ChIP-seq), open regulatory DNA regions (ATAC-seq) and 5mC DNA modifications (Bisulfite-seq) have enabled the disentangling of the molecular circuitry responsible for gene expression regulation. The description of which protein regulates which gene, the regulatory DNA element involved in this process and which DNA modification will be responsible to promote gene activation or repression will allow to further improve the understanding of the signalling cascade that leads to a particular phenotype.

Nowadays different industrial competitors are involved in the development of high-throughput sequencing techniques to allow the study of epigenetic gene regulation. Illumina/Solexa released the Genome Analyzer II in 2006 gaining in output and reductions in cost. As a consequence, Illumina machines currently dominate the high throughput sequencing market and are the most used to perform a test for new sequencing techniques (Reuter et al., 2015). Life Technologies commercialized Ion Torrent's semiconductor sequencing technology in 2010. The speed of sequencing, 2–8 hr depending on the machine and chip used, make these sequencers particularly useful for clinical applications rather than academic usage. Pacific Biosciences developed Single-molecule real-time (SMRT) sequencing. The latest chemistry version P6-C4 from 2014 allows to sequence reads up to 15000 base pairs in length. Important, this sequencing technology allows directly detects epigenetic modifications by measuring kinetic variation during base incorporation. Finally, Oxford Nanopore is a single-molecule strategy that has made significant progress in recent years. Sequencing is accomplished by measuring characteristic changes in current that are induced as the bases are threaded through the pore by a molecular motor protein. A single 18 hr run can produce >90 Mb of data from around 16,000 total reads, with median and maximum read lengths of 6 kb and >60 kb, respectively (Ashton et al., 2015). As with all single-molecule sequencing methodologies, error rates are high. Jain and colleagues most recently reported insertion, deletion, and substitution rates of 4.9%, 7.8%, and 5.1%, respectively.

Even if tremendous efforts have been made in the field of sequencing technologies, there are still limitations that need to be considered during the adoption of sequence-based techniques to study epigenetic gene regulation. From a technological perspective, accuracy and coverage across the genome are still problematic, particularly for GC-rich regions. Furthermore, the short read lengths produced by most current platforms is limiting our ability to accurately characterize large repeat regions, many indels, structural variant and trans-splicing mechanisms, leaving significant portions of the genome inaccurately explored.

3.4.1 ChIP-seq

The spring and summer of 2007 were among the most important for the development of chromatin immunoprecipitation followed by sequencing (ChIP-seq), a genomic technique that allows the identification of where a protein is binding along the DNA. Three important works (Johnson et al. 2007, Mikkelsen et al. 2007, Barski et al. 2007), pioneered the establishment of this technology that became later useful in projects such as the encyclopedia of DNA elements (ENCODE). Today, the web portal of this project collected ~9000 experiments for ChIP-seq for several different proteins involved in the regulation of gene expression and this number is growing every year.

The protocol of a ChIP-seq assay can be described in a few crucial steps. ChIP-Seq typically starts with crosslinking of DNA-protein complexes. Samples are then fragmented and treated with an exonuclease to trim unbound oligonucleotides. Protein-specific antibodies are used to immunoprecipitate the DNA-protein complex. The DNA is extracted and sequenced, giving high-resolution sequences of the protein-binding sites.

However, not all the binding sites exhibit the same specificity for a given protein target. In fact, binding sites across the genome are not functionally equivalent and are influenced on particular types variation at the sequence level (in the form of single nucleotide polymorphism) as also are influenced on the expression level of the nearby gene. In fact, high proteins-turnover sites are associated with lower transcriptional output (Lickwar et al., 2012) even under similar rates of occupancy while sites stably bound by the same factor are associated with efficient transcriptional activation (Furey 2012).

Data coming from this sequencing technology can be processed with standard pipeline such as the one proposed from the ENCODE consortium (https://github.com/ENCODE-DCC/chip-seq-pipeline2).

3.4.2 ATAC-seq

Another sequencing technology emerged in the last few years, is the Assay for Transposon Accessible Chromatin-seq (ATAC-seq). This technique is a transposon-based insertion sequencing technology and it is used to map open chromatin regions (Buenrostro et al., 2013) along the genome. The method relies on next-generation sequencing (NGS) library construction using the hyperactive transposase Tn5. NGS adapters are loaded onto the transposase, which allows simultaneous fragmentation of chromatin and integration of those adapters into open chromatin regions. The ATAC-seq protocol can be applied to small numbers of cells or even single cells. Regulatory information is especially revealing when compared across many

individual genomes or within a single genome across many cells or tissue types.

The power of this technology is that at the same time it is possible to map all the accessible regions of a cell irrespective of the active chromatin mark present or the nucleosome position. ATAC-seq can recapitulate all the genomic position mapped with other sequencing techniques that aim to map the nucleosome position (MNAse-seq), regulatory non-promoters regions (FAIRE-seq) and regulatory promoters regions that are sensitive to cleavage by DNase I (DNAse-seq) (Fig. 5).

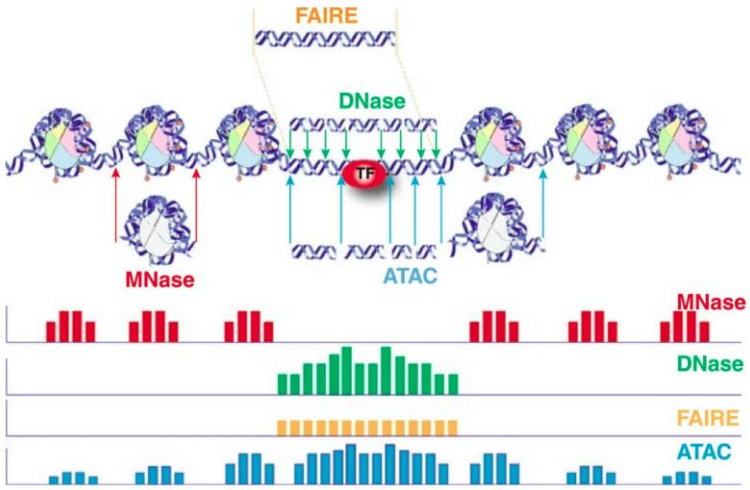


Fig. 5. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by coloured arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments (Tsompana and Buck, 2014).

ATAC-seq protocols require two main steps. As a first step, the cells need to be harvested. Since the number of cells used in ATAC-Seq assays is crucial for the transposition reaction and size distribution of the generated DNA fragments, counting the cells is important. Furthermore, cells should be intact and in a homogeneous, single-cell suspension. After harvesting, cells are lysed with a nonionic detergent to yield pure nuclei. In the second step, there is the transposition reaction in which The resulting chromatin is then fragmented and simultaneously tagmented with sequencing adapters using the Tn5 transposase to generate the ATAC-Seq library. After purification, the library can be amplified by PCR using barcoded primers. The resulting library can then be analyzed by qPCR or next-generation sequencing.

As for ChIP-seq, the data coming from this sequencing technology can be processed with standard pipeline such as the one proposed from the ENCODE consortium (https://github.com/ENCODE-DCC/atac-seq-pipeline).

3.4.2.1 single-cell ATAC-seq

ATAC-seq has been also improved at the single-cell level. In fact, single-cell platforms such as the Fluidigm C1 array (Buenrostro et al., 2018), droplet-based 10X and Bio-Rad (Lareau et al. 2019) and Split-pool (Cusanovich 2018) have allowed the application of the ATAC-seq protocol to single cells. The application of the technology to separate the cells of choice is followed by the implementation of the ATAC-seq protocol. The usage of these technologies allowed to identify regulatory elements involved in the differentiation of human blood cells in normal and upon stimulatory conditions (Buenrostro et al., 2018; Lareau et al. 2019) and also the identification of tissue-specific regulatory elements in the mouse (Cusanovich 2018). It does not exist at the moment a unified framework that is able to analyze scATAC-seq data. However, the major steps can be obtained from the literature and computational tools can be applied according to the data type and the number of cells (Chen et al. 2019).

3.4.4 Bisulfite-seq

Genome-wide 5mC distribution can be obtained using different chemistry protocols, the most popular of which is the whole-genome bisulfite sequencing (WGBS). The idea behind this sequencing technology is that DNA methylation status of single cytosines is obtained by treating the DNA with sodium bisulfite before sequencing. Sodium bisulfite is a chemical compound that converts unmethylated cytosines into uracil. The cytosines that haven't converted in uracil are methylated. After sequencing, the unmethylated cytosines appear as thymines. WGBS allows the sequencing of all the cytosines of a genome irrespective of the genomic context in which they are located. Given the high cost of this technique, a variant of WGBS is the reduced representation bisulfite sequencing (RRBS) (Meissner et al. 2005). This version differs from WGBS in the early steps of the protocol, in which CpG are extracted using restriction enzymes that can recognize only CpG. Once the CpG are separated the bisulfite conversion and sequencing will be restricted at CpG rich DNA regions. Methylated cytosines with characteristics different from the CpG, such as those like CHG or CHH (where H = A,T), won’t be sequenced and this will save money and time in the data processing.

There are also other sequencing technologies to study 5mC that are bisulfite free. Methylated DNA immunoprecipitation sequencing (MeDIP-seq) relies on the use of some antibody to precipitate fragments that can recognize all methylated cytosines or only those in the CpG context. Other methods have recently been developed for the application of the technique to low (160ng) DNA concentrations such as MBD-seq and MethylCap-seq which use a methyl CpG binding domain protein to precipitate DNA fragments containing methylated CpG sites, with a preference for those fragments with a high density of methylated sites (Hardcastle 2013). Furthermore, another bisulfite free technique takes advantage of an enzyme that recognizes unmethylated CpG islands (Staševskij et al. 2017). This methodology is named tethered-

oligonucleotide-primed (TOP) sequencing (seq) and uses an engineered version of the M.SssI methyltransferase and a synthetic analogue of the AdoMet cofactor to tag the unmodified and hemimethylated CG sites and the whole TOP-seq reaction conditions are adapted for Ion Torrent sequencing.

As for ChIP-seq and ATAC-seq, the data coming from WGBS and RRBS can be processed with standard pipeline such as the one proposed from the ENCODE consortium (https://github.com/ENCODE-DCC/dna-me-pipeline).

3.5 Artifacts and variation in ChIP-seq results

The interest in the application of technology can lead to an understanding of potential factors that contribute to evaluating its reliability and consistency across several experiments and conditions. In the case of ChIP-seq, its application during the ENCODE project (The Encode Consortium 2012) established its popularity and several laboratories started to use it as a standard in order to understand the function of specific proteins and transcription factors when binding into the genome.

When it came to using ChIP-seq to ask a specific biological question, several groups have found inconsistency in the binding of the protein under investigation. Also, particular genomic regions were found to be susceptible to a high number of reads mapping into them irrespective of the type of protein investigated. It is therefore important to identify such artifactual regions as also create a sort of flags that allow the users to be ready to further validate the results obtained from these assays.

There two main types of artifactual regions in ChIP-seq assays. The first is named “Blacklisted regions” (Amemiya et al., 2019) and as normal routine ChIP-seq results are usually interrogated for their presence before downstream interpretative analysis. They are also adopted as a common filter in most of the data processing pipelines of the ENCODE project. The second has multiple names but all of them have the same meaning: the signal is strong and inconsistent close to highly expressed genes and in the vicinity of promoters having R-loop (Teytelman et al. 2013, Park et al. 2013, Jain et al 2015, Wreczycka et al. 2019).

Finally, genomic variants located in the binding sites and the turnover of the proteins that bind to the same location can contribute to inconsistent ChIP-seq results across individuals and genomic locations (Lickwar et al., 2012, T. Furey 2012).

3.5.1 Blacklisted regions from ENCODE

High-throughput sequencing assays rely on the usage of annotated genomic regions. Regions where genome assembly results in an erroneous signal, have been analyzed and also annotated throughout the ENCODE project production phase and named as “Blacklisted”. These genomic positions often produce artefact signal in certain loci mainly because of excessive unstructured anomalous sequences. Reads mapping to them are uniquely mappable so simple mappability filters do not remove them. These regions are often found at specific types of repeats such as centromeres, telomeres and satellite repeats.

Blacklisted regions were initially manually annotated for *Homo sapiens* (human) genome assembly GRCh37 (referred to as hg19) to cover a large number of repeat elements in the genome, particularly rRNA, alpha satellites, and other simple repeats. Afterwards, their annotation has been extended to the mouse using mouse ENCODE (*Mus musculus:* mm9 and mm10), and modENCODE input ChIP-seq samples (*Caenorhabditis elegans:* ce10 and ce11, *Drosophila melanogaster:* dm3 and dm6). The removal of Blacklisted regions has become a standard post-processing quality check before the downstream analysis of next-generation sequencing data. The step of removing blacklisted regions is included in the ChIP-seq and ATAC-seq ENCODE pipelines provide.

3.5.2 Conservation of artefacts in ChIP-seq data

Another source of noise present only in ChIP-seq data are the recently discovered High Occupancy Target (HOT) regions. These regions were identified in several independent works in yeast (Park et al. 2013, Teytelman et al. 2013,). In one case (Park et al. 2013), it was shown that false-positive signals in ChIP-seq data are derived from high rates of transcription, are inherent to the ChIP procedure, and can be exacerbated by sequencing library construction procedures. The expression bias can be strong enough that a known transcriptional repressor like Tup1 was found to erroneously appear to be an activator.

In another work always in yeast (Teytelman et al. 2013), the silencing proteins Sir2, Sir3, and Sir4 showed 238 unexpected euchromatic loci exhibiting enrichment of all three. These 238 loci were renamed "hyper-ChIPable" and were located in the vicinity of highly expressed regions with strong polymerase II and polymerase III enrichment signals. Furthermore, the apparent enrichment of various proteins at hyper-ChIPable loci was not a consequence of artefacts associated with deep sequencing methods, as confirmed by ChIP-quantitative PCR but rather of a technical issue with the immunoprecipitations.

Afterwards, regions with similar characteristics of the aforementioned were confirmed in *Drosophila* (Dhawal et al. 2015). In this work were consistent results were obtained while investigating two ISWI-containing nucleosome remodelling factors, ACF and RSF. Employing several polyclonal and monoclonal antibodies directed against their signature subunits, ACF1 and RSF-1, robust profiles were obtained indicating that both remodelers co-occupied a large set of active promoters. Further validation included controls using chromatin of mutant embryos that do not express ACF1 or RSF-1. However, ChIP-seq profiles were unchanged, suggesting that they were not due to specific immunoprecipitation.

Finally, using data from the ENCODE project of ChIP-seq data from a collection of target proteins that were compared with data in which the protein target was removed (Wreczycka et al., 2019), the authors showed that as also in mouse and humans it is possible to observe such artefact signals. These regions were renamed high occupancy target (HOT) and shared common characteristics of “hyper-ChIPable” and the “phantom-peaks”. A peculiar characteristic of these promoters was the low methylation value, the high CG content and R-loop structures.

3.6 Identification of cell types using single-cell sequencing data

Next-generation sequencing technologies have allowed the study of cell populations for particular tissues or cell lines (referred to also as “bulk”) (GTEx consortium 2015). However, the limitation of such techniques is the impossibility to detect rare cell types that can be present in a subpopulation of cells belonging to the main population. Nowadays, single-cell sequencing technologies and corresponding data processing algorithms overcome the limitations of bulk sequencing assays for several different molecular layers of gene expression that are important for the study of genomic regulation. In fact, the regulation of gene expression can be studied at the single-cell level for gene expression by RNA quantification (scRNA-seq) (Wu et al. 2014), open chromatin accessible regions (scATAC-seq) (Buenrostro et al. 2018), histone modifications (scChIP-seq) (Grosselin 2019, A. Rotem 2015) and DNA methylation (scMet) (Luo et al. 2017, Mulqueen et al. 2018). Altogether, these technologies are allowing to discriminate sub-group of cells from a population that might be relevant to better understand clonal variability during development or cancer evolution. In fact, the application of these technologies is a unique opportunity to understand the interplay between intrinsic cellular processes and extrinsic stimuli such as the local environment or neighbouring cells in cell fate determination. Furthermore, single-cell studies are also of interest in the clinics to better understand how a small group of cells, within a population, may determine the outcome of an infection, drug or antibiotic resistance and cancer relapse (reviewed in Salibe et al. 2014).

Single-cell data, derived from next-generation sequencing technologies, comes with different data analysis challenges. The usual steps in the data pre-processing differ from bulk sequencing because of the different library preparation and protocols available. For every protocol, specific pre-processing steps are important to discriminate artefacts during the library preparation that results in the formation of duplicated cells (Wolock 2019) or duplicated barcodes (Lareau et al. 2019) that have to be considered before starting any downstream analysis. Other challenges are in the post-processing data analysis that usually involves the correction for technical noise given by batch effects or the cell cycle of the different cells, the imputation of missing data and dimensionality reduction to keep the most informative features to cluster or classify the cells in informative groups (Fig. 7).

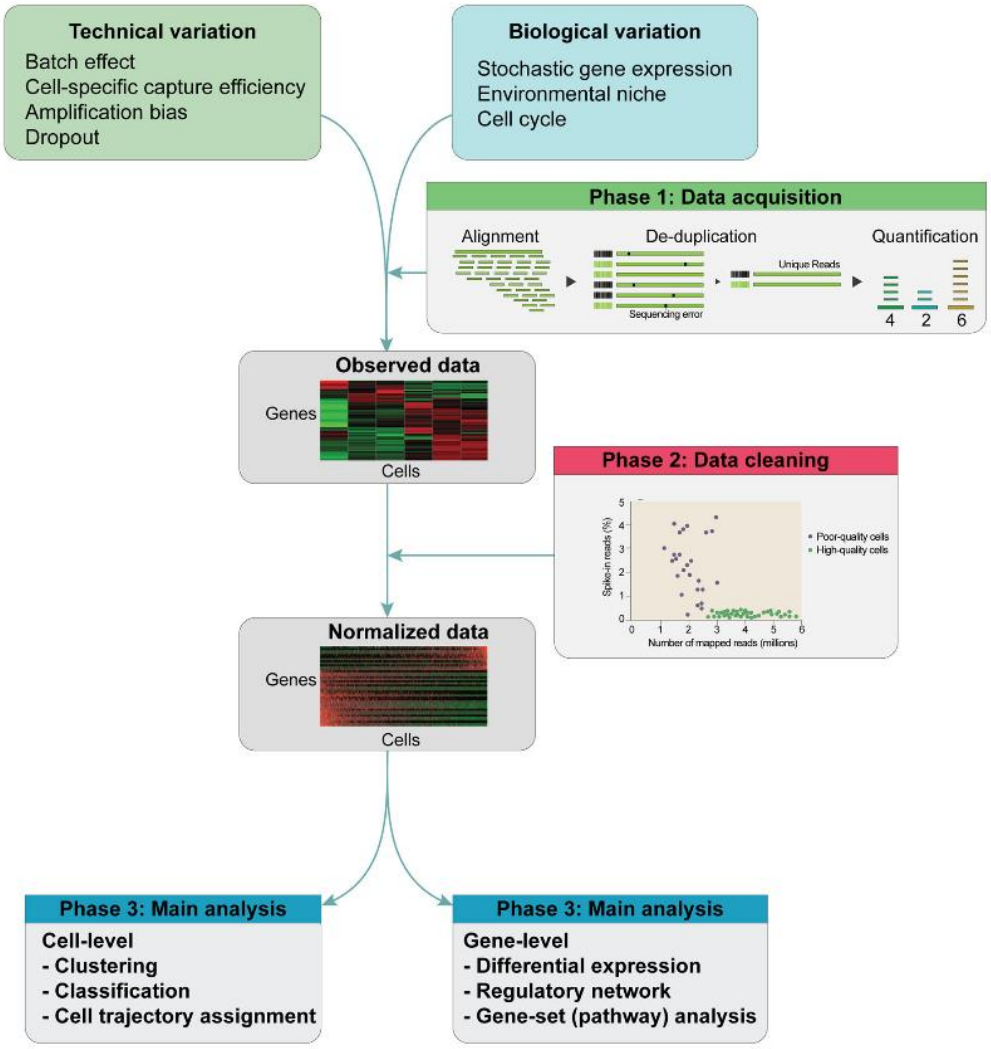


Fig. 7. A schematic overview of scRNA-seq analysis pipelines that can be applied to other types of single-cell data such as scATAC-seq, scMethylation and scChIP-seq. The data are inherently noisy with confounding factors, such as technical and biological variables. After sequencing, alignment and de-duplication are performed to quantify an initial gene expression profile matrix and remove duplicate cells and barcodes. Next, normalization is performed with raw expression data using various statistical methods. Additional QC can be performed when using spike-ins by inspecting the mapping ratio to discard low-quality cells. Finally, the normalized matrix is

then subjected to the main analysis through clustering of cells to identify subtypes. Cell trajectories can be inferred based on these data and by detecting differentially expressed genes, differentially accessible regions or differentially methylated regions between clusters (Hwang et al. 2018).

Single-cell datasets are defined as having the labels (or ground truth) if are obtained by sorting the cells using a set of specific cell surface markers that can be used to sort the cells with the Fluorescence-Activated Cell Sorting (FACS), or can be un-labelled if they belong from a heterogeneous un-sorted population of cells. The identification of a group of cells can be obtained using a different set of machine learning algorithms such as i) “classification algorithms” in case the cells are annotated in an automatic way using supervised machine learning techniques or ii) with “unsupervised clustering techniques” (Kiselev et al. 2019) in the case the cells are annotated using marker genes or marker regions.

3.6.1 Clustering approaches applied to single-cells

Among the mature sequencing technologies that can be applied to single cells we can find scRNA-seq (Tang et al. 2009, Hashimshony et al. 2012, Picelli et al. 2013, Macosko et al. 2015, as also the commercial platforms Fluidigm C1, Wafergen ICELL8 and the 10X Genomics Chromium), scATAC-seq (Fluidigm C1 array (Buenrostro et al. 2018), the 10X Genomics droplet-based scATAC platform and a recently optimized split-pool protocol (Cusanovich et al 2018)), scMeth (Luo et al. 2017, RM Mulqueen 2018) and scChIP-seq (Grosselin et al. 2019, Rotem et al. 2015). The datasets produced by these sequencing techniques have allowed the development of computational methods to identify groups of cell types through unsupervised clustering techniques. All the methods and respective algorithms used for the clustering of the single cells have common main steps: i) normalization of the reads after quantification and removal of bias given by the different library size among the cells, ii) identification of the most informative features with the highest variance (genes for scRNA-seq, DNA regions for scATAC-seq or scChIP-seq and methylated regions for scMet), iii) dimensionality reduction techniques such as principal component analysis (PCA) to projects data into a lower-dimensional space and iv) clustering followed by visualization of the group of cells. At the moment of writing this thesis, only scRNA-seq, scATAC-seq and scMeth technologies have a reasonable amount of methods to process and analyze their data. Thus, I will give attention only to these three technologies and skip scChIP-seq.

For scRNA-seq, many tools use variants of these standard procedures aforementioned: SC3 method (Kiselev et al. 2017) uses a small subset of principal components and pcaReduce (Žurauskiene et al. 2016) applies PCA iteratively to select the most informative principal component for downstream analysis. Subsequently, the distances among the cells are calculated in the lower dimensional space by using only the highly variable selected genes. Among the different choices available to compute the distances among the cells we can find: Euclidean distance, cosine similarity, Pearson’s correlation

and Spearman's correlation. The main advantage of the three latter measures is their scale invariance, that is, they consider relative differences in values, making them more robust to the library or cell size differences (Kiselev et al. 2019). Popular clustering methods such as k-means and hierarchical clustering can be applied on the scaled data. Different methods apply different combinations of dimensionality reduction approach and clustering techniques. Methods such as scanpy, Seurat and PhenoGraph use PCA combined with a graph-based method to cluster the cells; SC3 uses PCA+k-means; SIMLR uses a data-driven dimensionality reduction+k-means); CIDR uses PCA+hierarchical clustering; GiniClust uses the Gini Index to define the imbalance of among the group of cells and the density-based spatial clustering of applications with noise (DBSCAN), TSCAN uses PCA+Gaussian mixture model, mpath and SINCERA use Hierarchical clustering, BackSpin uses Biclustering (hierarchical); RaceID, RaceID2 and RaceID3 uses k-means; SNN-Cliq uses a Graph-based approaches for clustering (reviewed in Kiselev et al. 2019).

While for scRNA-seq the features to cluster the single cells are fixed (the genes) for scATAC-seq the definition of the features and the steps that proceed the clustering analysis are different according to the method being used. Thus, for scATAC-seq, there is one additional step respect to scRNA-seq called feature definition or, as described in Fig. 8, "Define regions" (Chen et al 2019).

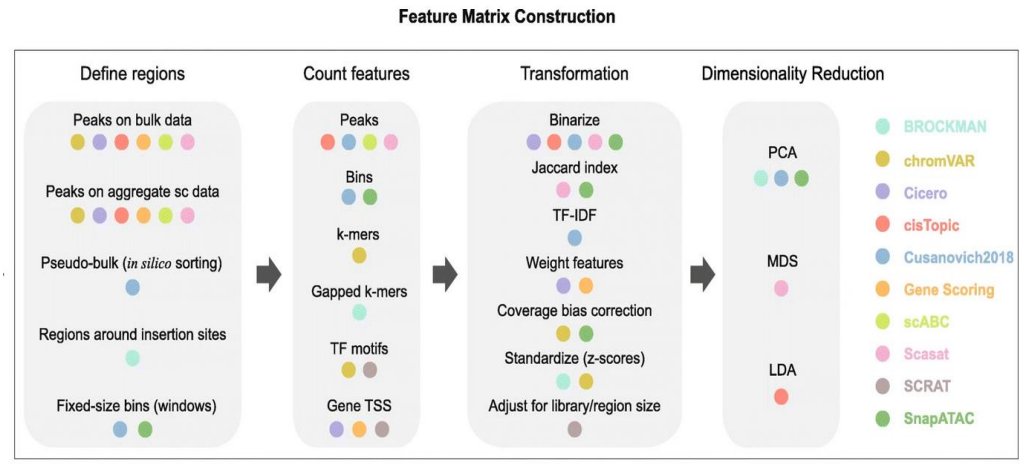


Fig. 8. The feature matrix construction techniques used by each method grouped into four broad categories: define regions, count features, transformation, and dimensionality reduction. A coloured dot under a technique indicates that the method (signified by the respective colour in the legend on the right) uses that technique (H. Chen et al 2019).

Each method differs in the main steps that proceed with the clustering analysis. In fact, BROCKMAN (deBoer 2018) represents genomic sequences by gapped k-mers (short DNA sequences of length *k*) within transposon integration sites and infers the variation in k-mer occupancy using principal component analysis (PCA). chromVAR (Schep et al. 2017) estimates the dispersion of chromatin accessibility within peaks sharing the same feature,

e.g., motifs or k-mers. Cicero (Pliner et al. 2018) calculates a gene activity score based on accessibility at a promoter region and the regulatory potential of peaks nearby. cisTopic (Bravo González-Blas et al. 2019) applies latent Dirichlet allocation (LDA) (a Bayesian topic modelling approach commonly used in natural language processing) to identify cell states from topic-cell distribution and explore *cis*-regulatory regions from region-topic distribution. Previous approaches that utilize latent semantic indexing (LSI) (termed here as *Cusanovich2018* (Cusanovich et al. 2018, Bravo Gonzalez-Blas et al. 2019, Cusanovich et al. 2018) first partition the genome into windows, normalize reads within windows using the term frequency-inverse document frequency transformation (TF-IDF), reduce dimensionality using singular value decomposition (SVD), and perform a first-round of clustering (referred to as "in silico cell sorting") to generate clades and call peaks within them. Finally, the clusters are refined with a second-round of clustering after TF-IDF and SVD based on read counts in peaks. The Gene Scoring method (Lareau et al. 2019) assigns each gene an accessibility score by summarizing peaks near its transcription start site (TSS) and weighting them by an exponential decay function based on their distances to the TSS. scABC (Zamanighomi et al. 2018) first calculates a global weight for each cell by taking into account the number of distinct reads in the regions flanking peaks (to estimate the expected background). Based on these weights, it then uses weighted k-medoids to cluster cells based on the reads in peaks. Scasat (Baker et al. 2019) binarizes peak accessibility and uses multidimensional scaling (MDS) based on the Jaccard distance to reduce dimensionality before clustering. SCRAT (Ji et al. 2017) summarizes read counts on different regulatory features (e.g., transcription factor binding motifs, gene TSS regions). SnapATAC (Fang et al. 2019) segments the genome into uniformly sized bins and adjusts for differences in library size between cells using a regression-based normalization method; finally, PCA is performed to select the most significant components for clustering analysis. After feature matrix construction, three commonly used clustering approaches such as K-means, Louvain, and hierarchical clustering (Kiselev et al. 2019) can be used to cluster the single cells in groups.

Finally, for scMeth datasets, there are few available methods that can be used to analyse such data until performing the clustering. Among these, EpiScanpy (Danese et al. 2019) makes the many existing RNA-seq workflows from scanpy available to large-scale single-cell data from other omics modalities such as scMeth. The method implements multiple feature space constructions for epigenetic data and shows the feasibility of common clustering, dimension reduction and trajectory learning techniques. Another method named Epiclomal (De Souza et al. 2018) uses a probabilistic clustering method arising from a hierarchical mixture model to simultaneously cluster sparse single-cell DNA methylation data and infer their corresponding hidden methylation profiles. Lastly, (Pairwise Dissimilarity Clustering (PDclust) (Hui et al. 2018) first calculates a pairwise dissimilarity measure of all the CpGs among the single cells and after uses euclidean distance on these values to perform clustering analysis. Other methods available for single-cell methylation, such

as DeepCpG (Angermueller et al. 2017) and Melissa (Kapourani et al. 2019) are intended to solve the sparsity of scMet data imputing methylation values at CpGs sites.

3.6.2 Annotation approaches applied to single-cells

Single-cell data analysis involves the identification of cell populations in a given dataset thanks to the implementation of specific machine learning algorithms. There are several unsupervised clustering techniques adopted for the identification of cell groups. These algorithms, as explained in the previous paragraph, compute the similarity of the cells followed by cell population annotation by assigning labels to each cluster. This approach is very valuable for the identification of novel cell populations and resulted in cellular maps of entire cell lineages, organs, and even whole organisms (Abdelaal et al. 2019).

While the previously described approaches are very useful, it can be time demanding to be implemented. To overcome the limitation above other solutions have been developed. These solutions are named as “annotation approaches” and can run in automatic. There are two main characteristics that distinguish the annotation approaches: i) the adoption of “a-priori” knowledge of known marker genes (or DNA and methylated regions) to annotate the cell in specific group and ii) the adoption of the “rejection option” that consists in a negative test that uses marker genes of a specific tissue to annotate the cells in a totally different tissue. While for scRNA-seq there are several methods available, for other techniques such as scATAC-seq, scChIP-seq and scMeth there are no methods for such task. Thus, I will limit the description of such algorithms only for scRNA-seq datasets that might be potentially adopted also for scATAC-seq and scMeth.

There are 22 available methods to automatically classify cells using scRNA-seq data. These methods differ from each other according to the classifier adopted and whether they use the rejection option. Methods such as Garnett (Pliner et al. 2019) uses a generalized linear model together with a prior knowledge and the rejection option to classify the single cells; Moana (Wagner et al. 2018) uses a support vector machine (SVM) with linear kernel to classify the single cells with a prior knowledge; DigitalCellSorter (Domanskyi et al. 2019) uses a voting based on cell type markers as classifier and a prior knowledge; SCIINA (Zhang et al. 2019) uses a bimodal distribution fitting to marker genes as classifier and a prior knowledge; scVI (Lopez et al. 2018) uses Neural network classifier; Cell-BLAST (Cao et al. 2019) uses Cell-to-cell similarity and a rejection option; ACTIIN (Ma et al. 2019) uses Neural network as classifier; LAmbDA (Johnson 2019) uses Random forest as classifier; scmapcluster and scmap (Kiselev et al. 2018) uses Nearest median classifier and rejection option; scPred (Alquicira-Hernandez et al. 2018) uses support vector machine (SVM) with radial kernel as classifier with rejection option; CHETAH (Kanter et al. 2019) uses a correlation to training set as classifier with rejection option; CaSTLe (Lieberman et al. 2018) uses Random forest as classifier; SingleR (Aran et al. 2019) uses correlation training as classifier;

scID (Boufea et al. 2019) uses linear discriminant analysis (LDA) as classifier with rejection option; singleCellNet (Tan et al. 2018) uses a random forest classifier.

3.7. Aim of the thesis

The last decade has seen a rapid increase in the number of high-throughput sequencing technologies, which has enabled the investigation of the basic principle of gene expression regulation. Thus, quantitative methods of analyzing and interpreting the data produced by such technologies are of fundamental importance. Therefore, for my thesis, I used data produced in the research groups in which I conducted my research together with datasets available in the genomic community to develop new and apply existing computational methods with the aim of addressing biological questions on a molecular level. More specifically**,** I investigated how cells use regulatory DNA elements and epigenetic modifications to promote or maintain a specific cellular identity, which can also be prevented when there is unwanted variation. Furthermore, given the importance of computational methods to analyze single-cell ATAC-seq datasets, I evaluated which methods outperform others in identifying cell types of various datasets (for example, mouse tissue, hematopoietic stem cells, or unlabeled PBMCs).

In the first project, I aimed to understand the role of GADD45 proteins in DNA demethylation in mESCs by analyzing WGBS data generated from wild type mESCs and mESCs which are knockout for *Gadd45* genes. I investigated whether GADD45-demethylated loci are located at TET-dependent demethylated sites that harbour 5hmC, 5fC, and 5caC oxidative products, indicating active DNA demethylation. Additionally, I determined if these regions are enriched for particular classes of regulatory elements. Finally, I asked whether the demethylated regions identified were significantly associated with 2C-like specific genes in order to understand if GADD45 proteins can promote the 2C-like state by regulating 5mC turn over in mESCs.

In the second project, I benchmarked several computational methods that can be implemented to classify cell types using single-cell ATAC-seq data. The aim of this work was to evaluate which methods are suitable to distinguish cells differentiating into the three main blood lineages (red blood cells, lymphocytes, and myeloid) as well as a population of cells called Peripheral Blood Mononuclear Cells (PBMCs). During this project, I addressed the problem of interpreting clustering results from unlabeled datasets, (like PBMCs), by developing an algorithm that leverages the information of marker genes combined with the Gini Coefficient.

In the third and final project, I developed a method to address a common problem in ChIP-seq data analysis related to the reliability of the results. It is accepted nowadays that not all genomic regions have the same reproducibility in ChIP-seq results. My developed method is able to point out which regions tend to have a high rate of variability in several independent replicates and protein targets, and it does so in a cell-type-specific manner. Thus, the adoption of this method can improve the identification of cells and target-specific binding sites to improve a better interpretation of the results during downstream analysis.

4. Results

4.1 Preamble

This paper is the result of my stay in the laboratory DNA demethylation, DNA repair and Reprogramming headed by Prof. Dr Christof Niehrs at the IMB in Mainz. My contribution to the paper is indicated at the end of the manuscript and a repository is available to reproduce the entire analysis and figures at this link: https://github.com/tAndreani/DNA-Demethylation-Gadd45.

4.2 Chapter 1

GADD45 promotes locus-specific DNA demethylation and 2C cycling in embryonic stem cells (Schule KM, Leichsenring M, Andreani T et al, *Genes & Development* 33, 782-798, 2019)

Abstract

Mouse embryonic stem cell (ESC) cultures contain a rare cell population of “2C-like” cells resembling two-cell embryos, the key stage of zygotic genome activation (ZGA). Little is known about positive regulators of the 2C-like state and two-cell stage embryos. Here we show that GADD45 (growth arrest and DNA damage 45) proteins, regulators of TET (TET methylcytosine dioxygenase)-mediated DNA demethylation, promote both states. Methylome analysis of Gadd45a, b, g triple-knockout (TKO) ESCs reveal locus-specific DNA hypermethylation of ~7000 sites, which are enriched for enhancers and loci undergoing TET – TDG (thymine DNA glycosylase)-mediated demethylation. Gene expression is misregulated in TKOs, notably upon differentiation, and displays signatures of DNMT (DNA methyltransferase) and TET targets. TKOs manifest impaired transition into the 2C-like state and exhibit DNA hypermethylation and down-regulation of 2C-like state-specific genes. Gadd45a,b double-mutant mouse embryos display embryonic sublethality, deregulated ZGA gene expression, and developmental arrest. Our study reveals an unexpected role of GADD45 proteins in embryonic two-cell stage regulation.

Introduction

Mouse embryonic stem cells (ESCs) are a model for the inner cell mass around implantation stage. ESCs are heterogeneous and contain subpopulations with different properties (Hayashi et al. 2008; Macfarlan et al. 2012; Toyooka et al. 2008; Zalzman et al. 2010). One of these subpopulations (1-5%) is transcriptionally and epigenetically similar to the 2-cell stage embryo and hence referred to as '2C-like' (Macfarlan et al. 2012; Zalzman et al. 2010). The embryonic 2-cell stage is a key phase of mouse development during which the major wave of zygotic genome activation occurs (ZGA; reviewed in (Eckersley-Maslin et al. 2018; Jukam et al. 2017; Svoboda 2017). During this period, the bulk of the genome becomes transcriptional active, which is accompanied by extensive chromatin modification. 2C-like ESCs exhibit unique molecular features of totipotent cleavage-stage cells and in chimeras they can contribute to both embryonic and extraembryonic derivatives, including trophoblast (Choi et al. 2017; De Iaco et al. 2017; Ishiuchi et al. 2015; Macfarlan et al. 2012; Rodriguez-Terrones et al. 2018; Whiddon et al. 2017). ESCs cycle in and out of this transient 2C-like state at least once within nine passages (Zalzman et al. 2010). Characteristic markers for the 2C-like state are murine endogenous retrovirus with leucine tRNA primer (MERVL) retrotransposon and Zinc finger and SCAN domain-containing protein 4 (Zscan4; Macfarlan et al. 2012; Zalzman et al. 2010). Thus, 2C-like ESCs model essential aspects of the 2-cell stage embryo and ZGA (reviewed in (Eckersley-Maslin et al. 2018; Ishiuchi and Torres-Padilla 2013)). Few positive regulators of the 2C-like state or ZGA are known, including the transcriptional regulators ZSCAN4 (Amano et al. 2013; Falco et al. 2007; Hirata et al. 2012; Zalzman et al. 2010), DUX (Hendrickson et al. 2017; De Iaco et al. 2017), STELLA (Huang et al. 2017), TBX3 (Dan et al. 2013), as well as DPPA2 and DPPA4 (Eckersley-Maslin et al. 2019).

Here we report a role for the small gene family Gadd45 (Growth arrest and DNA damage protein 45) a,-b and -g, in regulation of the 2C-like state. GADD45α is a stress response protein, which interacts with the key enzymes of the DNA demethylation machinery, TET1 (TET methylcytosine dioxygenase 1) and TDG Thymine DNA glycosylase; G. Barreto et al. 2007; S. Cortellino et al. 2011; S. Kienhöfer et al. 2015; Z. Li et al. 2015). TET enzymes convert 5-methylcytosine (5mC) sequentially to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC; Guo et al. 2011; He et al. 2011; Ito et al. 2011; Kriaucionis and Heintz 2009; Tahiliani et al. 2009). DNA repair via TDG removes 5fC and 5caC to restore unmethylated cytosine (Cortázar et al. 2011; Cortellino et al. 2011; Shen et al. 2013). GADD45α is an adapter protein that tethers TET/TDG to sites of DNA demethylation, which functions in locus-specific DNA demethylation (Arab et al. 2014; Barreto et al. 2007; Cortellino et al. 2011; Li et al. 2010; Sabag et al. 2014; Zhang, et al. 2011a). GADD45α recruits TET/TDG to specific sites in the genome via additional cofactors (Arab et al. 2014; Arab et al. 2019; Schäfer et al. 2013; Schäfer et al. 2018).

Since not only Gadd45a but also Gadd45b and Gadd45g promote DNA demethylation (Gavin et al. 2015; Jarome et al. 2015; Ma et al. 2009; Rai et

al. 2008; Sen et al. 2010) and since single mouse mutants are viable (Hollander et al. 1999; Lu et al. 2001; Lu et al. 2004), this raises the question of whether the genes have overlapping roles in development and differentiation. To address this question we have generated and characterized Gadd45a,b,g triple–knockout mouse embryonic stem cells (TKO-ESCs). We find that GADD45 proteins are dispensable for maintaining pluripotency and self-renewal. However, methylome analysis indicates that GADD45 proteins are required for DNA demethylation of specific loci and normal gene expression. Moreover, GADD45 proteins promote the 2C-like state and Gadd45a,b double mutant mouse embryos show partial deregulation of ZGA genes at the 2-cell stage and developmental arrest. Collectively, the results indicate that GADD45 proteins act redundantly to promote locus-specific demethylation as well as embryonic 2-cell stage.

Results

Gadd45 TKO ESCs are pluripotent and self-renew

We generated homozygous deletions in Gadd45a, Gadd45b and Gadd45g in ESCs using the CRISPR/Cas9 system (Jinek et al. 2012; Le Cong et al. 2013; Mali et al. 2013). Six gRNAs were cotransfected, two for each Gadd45 gene, to create 300-700 bp deletions between the 5'-UTR and the second intron, covering the start codon (Fig. 1A). Out of 276 colonies obtained after selection, three independent Gadd45 triple-knockout (TKO) ESC clones were obtained (Fig. S1A). Sequencing confirmed deletion of the respective genomic regions in the TKO-ESCs (Fig. S1B) and Western blot and mass-spectrometry analysis showed that both GADD45α and GADD45□ were undetectable in TKO-ESCs (Fig. S1C-D). GADD45ɣ was undetectable in both wild type and mutant ESCs (Fig. S1E) and even if truncated GADD45 ɣ protein was expressed, it would be non-functional since deletion of exons 1 and 2 include the dimerization domains (aa43-86) required for GADD45 ɣ function (Zhang, et al. 2011b). To generate three independent wild type ESC control (Co) clones, ESCs were transfected with Cas9 and the selection marker, but without specific gRNAs.

The Gadd45 TKO-ESCs showed no apparent loss of pluripotency, Oct4, Nanog and Sox2 expression was not reduced (Fig. 1B), and their morphology as well as growth rate was normal (Fig. S2A-C). In teratoma assays, TKO-ESCs gave rise to derivatives of all three germ layers (Fig. 1C).

Given the previously described functions of GADD45 proteins in active DNA demethylation, we analyzed global 5mC, 5hmC, 5fC, and 5caC levels by quantitative mass spectrometry. Global levels of 5mC and its oxidative derivatives were only mildly affected in Gadd45 TKO-ESCs (Fig. 1D-G). There was a slight increase of 5mC with concomitant slight decrease of oxidized cytosine derivatives in TKO-ESCs compared to Co-ESCs. This result is consistent with GADD45 proteins acting not in global- but in locus-specific demethylation.

Mouse ESCs are a model for the inner cell mass (ICM) around implantation stage with relatively high methylation at promoters, enhancers and bivalent loci (Habibi et al. 2013). However, ESCs can be reverted to a hypomethylated ground state more similar to pre-implantation embryos using small-molecule MEK and GSK3β inhibitors ('2i'; Ficz et al. 2013; Leitch et al. 2013), as well by vitamin C (Blaschke et al. 2013). Hence, we induced global DNA demethylation via vitamin C or 2i treatment, however global levels of cytosine modifications in Gadd45 TKO-ESCs were changing similarly to Co-ESCs (Fig. S2D – S2E). We conclude that Gadd45 genes are dispensable for ESC maintenance, as is the case for Tet and Dnmt genes (Dawlaty et al. 2011; Dawlaty et al. 2014; Lei et al. 1996; Okano et al. 1999; Tsumura et al. 2006).

Loci undergoing TET-dependent oxidation are hypermethylated in TKO-ESCs

To unravel DNA methylation changes we performed whole-genome bisulfite sequencing (WGBS) of Co-and TKO-ESCs to obtain base-pair-resolution methylomes. Co-ESCs showed the characteristic bimodal distribution of CpG methylation, but the distribution in TKO-ESCs was skewed towards higher methylation (Fig. 2A). To call differentially methylated CpG-regions (DMRs) in TKO-ESCs, we used a stringent cutoff of 5% FDR and > 30% methylation difference on at least two CpGs. WGBS does not discriminate 5hmC from 5mC (Booth et al. 2012), hence we may underestimate the number of GADD45- dependent TET target sites. DMRs were broadly distributed on all 19 autosomes (data not shown). The DMR analysis identified 6,904 hypermethylated, but only 34 hypomethylated regions in TKO-ESCs (Fig. S3A). Thus, although global 5mC levels were only slightly increased by LC-MS/MS, the skewed bimodal methylation pattern and the ~200-fold bias towards hypermethylated DMRs indicate a locus-specific DNA hypermethylation in TKO-ESCs.

We therefore focused on the hypermethylated DMRs (hyper-DMRs). The majority overlapped with intronic and intergenic regions (Fig. S3B). There was around 2.5-fold enrichment for enhancers and coding exons (Fig. 2B). Moreover, hyper-DMRs were enriched at sites marked by 5fC or 5caC after Tdg knockdown (Shen et al. 2013)and to a lesser extent at sites marked by 5hmC (Kong et al. 2016; Shen et al. 2013). Hyper-DMRs were also enriched for TET-dependent hyper-DMRs (Lu et al. 2014). Positional correlation analysis centered on hyper-DMRs revealed prominent overlap with sites marked by 5hmC (Fig. 2C). Hyper-DMRs also overlapped with 5fC/5caC peaks accumulating in Tdg knockdown ESCs (Shen et al. 2013), indicating that hyper-DMRs are targets of TET/TDG-mediated cytosine oxidation and excision in ESCs. In contrast, there was little overlap with 5mC sites, indicating that the association of hyper-DMRs occurred specifically with oxidized cytosines.

In another positional correlation analysis, we plotted the average levels of methylation change between TKO- and Co-ESCs against the center of genomic features derived from a wide panel of published genome-wide mapping data sets in ESCs, including 5mC oxidative derivatives, DNA-binding factors, and major histone modifications. Moreover, we divided the analysis between proximal and distal elements with regard to gene transcription start sites (TSS). This analysis corroborated that the main co-occurrence of hypermethylation in TKO-ESCs was with sites marked by 5fC and 5caC, both in proximal and distal sites (Fig. 2D, top three rows). Hypermethylation accumulated also at the center of 5hmC peaks, but to a lower level and restricted to distal loci.

Conversely, CpG islands (CGIs), which typically occur at proximal sites, showed the lowest levels of methylation difference, consistent with the fact

that they are constitutively unmethylated (Deaton and Bird 2011). Other genomic features correlated with promoter CGIs, such as Pol2, TBP, and transcription elongation factors (Nelfa, Spt5) followed this trend. In general, this methylation signature parallels the signature in Tet1,2,3 TKO-ESCs (Lu et al. 2014), whereby TET mediated DNA demethylation i) mainly occurs at distal regulatory elements and ii) affects sites marked in control cells by 5fC and 5caC more pronounced than those marked by 5hmC. We conclude that the hypermethylation signature in TKO-ESCs closely correlates with loci processed by TET/TDG, consistent with GADD45 proteins acting in locus-specific DNA demethylation.

We segmented hyper-DMRs and carried out transcription factor (TF) binding motif analysis using HOMER (Fig. 3A). The most prominent hit in all DMRs was Klf5, a possible reader of methylated DNA (Liu et al. 2014; Spruijt et al. 2013), which promotes the pluripotent ESC state and is required for trophectoderm development (Ema et al. 2008; Parisi et al. 2008). Other prominent hits were Ets-like TF binding elements (Ehf, Etv1, Fli1, Elk4/1). Interestingly, hyper-DMRs overlapping with enhancers harboring 5hmC were enriched for motifs of Zscan4, a key regulator of the 2C-like state.

Methylation regulated genes are downregulated in Gadd45 TKO-ESCs

To identify genes differentially expressed upon Gadd45 deficiency RNA-seq analysis was carried out under three culture conditions: normal serum culture, and two conditions inducing global demethylation, vitamin C and 2i treatment (Blaschke et al. 2013; Ficz et al. 2013; Leitch et al. 2013). A total of 135 genes were differentially expressed in Gadd45 TKO-ESCs versus Co-ESCs during normal serum culture (FDR 10%). This number decreased sharply upon vitamin C- or 2i treatment (Fig. 3B), supporting that gene deregulation in TKO-ESCs is due, directly or indirectly, to DNA hypermethylation. Around 25% of deregulated genes overlapped with hyper-DMR associated genes (Fig. S3C) further indicating that only a moderate fraction of GADD45-dependent genes are methylation-regulated. This is in line with generally modest correlation between gene expression and DNA methylation in ESCs (Karimi et al. 2011; Lu et al. 2014) Genes downregulated in Gadd45 TKO-ESCs overlapped significantly with genes upregulated in Dnmt1 -/- /Dnmt3a -/- /Dnmt3b -/- TKO-ESCs (Fig. 3C; Karimi et al. 2011). Consequently, localization of genes downregulated in Gadd45 TKO-ESCs is enriched on the X-chromosome (Benjamini p= 2.2x10 -6), as has been observed for upregulated genes in Dnmt TKO-ESCs (Fouse et al. 2008). These genes tend to be involved in germ cell regulation (Wang et al. 2001). Moreover, genes deregulated in Gadd45 TKO-ESCs correlated with genes deregulated upon Tet1 knockdown in ESCs (Fig. 3D; Huang et al. 2014).

Previous studies showed that DNA methylation and demethylation play a more important role during differentiation than during pluripotency (Dawlaty et al. 2014; Jackson et al. 2004; Lei et al. 1996; Okano et al. 1999; Sakaue et al. 2010). We therefore subjected TKO-ESCs to three differentiation protocols

and analyzed transcriptome changes by RNA-seq. First, ESCs were differentiated for 8 days as embryoid bodies (EBs). Second, we differentiated ESCs for 6 days in serum-free monolayer culture. Third, ESCs were differentiated for 4 days as EBs and then treated for 4 more days with retinoic acid (RA). While the latter two protocols favor neuronal differentiation, the unguided EB culture allows differentiation into all three germ layers (Bibel et al. 2007; Ying et al. 2003). Gadd45a, b and g were all expressed at varying levels under these differentiation regimes (Fig. S3D).

The number of differentially expressed genes in TKO cells versus Co cells was 907 upon monolayer differentiation and 659 in EBs (FDR 10%, Fig. 3E), and thus gene deregulation in Gadd45 TKO cells is indeed increased upon differentiation (compare Fig. 3B, E). Only 22 genes were differentially expressed in RA-treated EBs, suggesting that the particular neural lineages induced by RA are less sensitive to Gadd45 deficiency. Only 48 genes were commonly deregulated in TKO cells between EB and monolayer differentiation (not shown), indicating that the function of the Gadd45 genes is highly context dependent. Interestingly, genes differentially expressed in Gadd45 TKO EBs were more than twice as likely to be marked by 5fC in their promoter regions compared to unaffected genes (Fig. S3E), supporting that GADD45-dependent genes are prone to undergo active DNA demethylation during EB differentiation.

GO term analysis in TKO EBs showed that downregulated genes were highly enriched for developmental terms such as system development, cell fate determination, cell migration, and axon guidance (Fig. S4A). Less pronounced GO term enrichment was found for genes upregulated in TKOs (Fig. S4B). For genes downregulated in monolayer differentiated TKO cells, GO term enrichment was also found for cell motility related terms (Fig. S4C), whereas genes upregulated were enriched for developmental and neuronal functions (Fig. S4D). Despite sharing similar GO ontologies, the genes affected in Gadd45 TKO EBs were largely distinct from those affected in Gadd45 TKO monolayer cells (not shown). We conclude that in differentiating ESCs, GADD45 proteins regulate genes related to developmental, neuronal, and cell motility function, consistent with the involvment of GADD45 proteins in the regulation of neural development (Huang et al. 2010; Kaufmann and Niehrs 2011; Ma et al. 2009).

We validated selected genes commonly deregulated in Gadd45 TKO-, Dnmt TKO- and Tet1 knockdown ESCs by qPCR. Commonly deregulated genes did not significantly cluster in terms of gene ontology (GO) enrichment (not shown), but included various germline specific (e.g. Rhox2a and Asz1) and pluripotency related (e.g. Pramel6 and Pramel7) genes (Fig. 3F and Fig. S5A).

To analyze the Gadd45 gene redundancy in differential gene expression, we conducted rescue experiments. Transient combined overexpression of Gadd45a, Gadd45b and Gadd45g rescued downregulation of a panel of misregulated genes in Gadd45 TKO-ESCs and further increased their

expression levels in Co-ESCs (Fig. S5A). Not only combined, but also individual Gadd45a, Gadd45b or Gadd45g overexpression was effective in these rescue experiments, indicating that Gadd45 genes in ESCs function redundantly (Fig. S5B-C). Genes downregulated in Gadd45 TKO-ESCs were also induced by 5'-deoxyazacytidine treatment (Fig. 3G), further supporting that gene downregulation in TKO-ESCs involved DNA hypermethylation.

To test directly for DNA hypermethylation in TKO-ESCs, we analyzed the methylation status of regulatory elements in the vicinity of selected GADD45-dependent genes, which are shared with DNMT- and TET1-dependent genes. The majority of CpG dinucleotides in the Rhox2a promoter (Fig. 3H), the Pramel6 promoter (Fig. S6A), and the Gm364 promoter (Fig. S6B) were hypermethylated in Gadd45 TKO-ESCs. In contrast, in two control gene promoters displaying high and low methylation levels, respectively (Sry, Rerg), no methylation changes were observed in TKO-ESCs (Fig. S6C-D).

We conclude that GADD45 proteins act redundantly to maintain normal expression and methylation levels of selected genes in ESCs, and that this list of genes overlaps with TET1 and DNMT target genes.

Gadd45 TKO-ESCs show impaired 2C-like state

We used the ESCAPE database (Xu et al. 2014), which integrates high-content data from embryonic stem cells, to conduct enrichment analysis of genes downregulated in undifferentiated Gadd45 TKO-ESCs (Fig. 4A). The top hit returned was a set of genes reported upregulated upon Gadd45a overexpression (Nishiyama et al. 2009), corroborating the validity of the ESCAPE analysis. Among the other hits with similar significance were genes upregulated upon misexpression of Zscan4 (zinc finger and SCAN domain-containing protein 4; (Nishiyama et al. 2009), a transcription factor whose recognition motif was enriched in hyper-DMRs (Fig. 3A). Zscan4 is a marker and regulator of the 2-cell embryo (Falco et al. 2007). Consistent with the overlap between GADD45- and ZSCAN4-regulated genes, we found that genes upregulated in 2C-like cells tend to be downregulated in Gadd45 TKO-ESCs, suggesting a requirement of Gadd45 genes in regulating the 2C-like state (Fig. 4B). Indeed, although only 97 genes in Gadd45 TKO-ESCs were downregulated (> 0.5-fold, 10% FDR), these were enriched in genes specifically expressed in 2C-like cells (Fig. S7A; Macfarlan et al. 2012). In contrast, not a single gene up-regulated in Gadd45 TKO-ESCs overlapped with the 2C gene set. 2C-specific genes were only modestly enriched in the vicinity of hyper-DMRs (Fig. 4C), which is expectable since the 2C-like cells only represent a small fraction of the ESC population. Of these 178 hyper-DMR- and 2C-like associated genes, six were also downregulated in the Gadd45 TKO-ESCs (Igfbp2, Inpp4b, Pramel6, Pramel7, Snhg11 and Tmem92). Other hypermethylated 2C genes may be downregulated only upon 2C cycling and hence escape detection.

To confirm these results, we monitored the expression of prominent 2C-associated genes in Co- and TKO-ESCs after combined overexpression of Gadd45a, -b and -g or GFP (Fig. 4D). Overexpression of Gadd45 genes not only rescued downregulation of the majority of tested 2C-associated genes in TKO-ESCs, but also increased expression levels even in Co-ESCs. In contrast, expression of retroviral elements unrelated to the 2C status (intracisternal A-particle, IAP) was unchanged.

To investigate the role of GADD45 in regulating the 2C-like state, we stably introduced a 2C-reporter, Zscan4c::eGFP (Zalzman et al. 2010), in Co- and TKO-ESCs (Fig. 4E). Flow cytometry revealed a significant reduction of Zscan4 + cells in TKO-ESCs (Fig. 4F). Combined overexpression of the Gadd45 genes rescued the reduction of Zscan4 + cells in TKO-ESCs to control levels (Fig. 4G). Likewise, individual overexpression of Gadd45a, -b or -g increased the percentage of Zscan4 + cells , as well as of cells harboring the mERVL 2C-reporter (Macfarlan et al. 2012), suggesting redundant function of the GADD45 proteins in 2C regulation (Fig. 4H and Fig S7B). Interestingly, long term 2i treatment (seven passages), which induces hypomethylation (Ficz et al. 2013; Leitch et al. 2013), abolished the observed difference in 2C-like population (Fig. S7C). In contrast, overexpression of Tet1, Tet2 or Tdg did not affect the frequency of Zscan4 + cells in TKO- nor in Co-ESCs (Fig. S7D-F).

Zscan4 + sorted cells show a global demethylation relative to unsorted ESCs (Eckersley-Maslin et al. 2016). We confirmed by mass-spectrometry that Co Zscan4 + cells show reduced 5mC levels (compare Fig. 4I Co ‘unsorted’ to ‘Zscan4 + ’) while Zscan4 + TKO cells showed hypermethylation. No difference was found for 5hmC (Fig. 4J).

Gene expression analysis of TKO-ESCs (Fig. 5A) revealed reduced RNA levels of 2C-specific genes not only in unsorted- but also in Zscan4 + TKO-ESCs (e.g. Zscan4, Sp110, Tcstv1). This reduction was not due to deficient maintenance of 2C-like state but reflected deficient entry, since cycling-out of the Zscan4 + state occurred with the same kinetics in Co- and TKO-ESCs (Fig. 5B).

Among the downregulated 2C genes was Dux, a key regulator of ESCs cycling into the 2C-like state (Hendrickson et al. 2017; Iaco et al. 2017). This suggested that GADD45 may act upstream of DUX to promote the 2C-like state. Concordantly, overexpression of Dux using a doxycycline-inducible plasmid restored the reduced number of 2C-like cells in TKO-ESCs (Fig. 5C). The expression level of repressors of the 2C-like state (Trim28, Lsd1, G9a, Chaf1a) was not altered in unsorted, Zscan4 + or Zscan4 - TKO-ESCs (Fig. S7G).

Finally, we explored the consequences of 2C misregulation in ESC transdifferentiation. While ESCs normally do not give rise to trophoblast, they do sporadically transdifferentiate into this lineage, which can be further

enhanced by BMP4 treatment (Beddington and Robertson 1989; Hayashi et al. 2010). Trophoblast transdifferentiation involves 2C cycling, since 2C-like cells are not lineage restricted (Macfarlan et al. 2012) and cycling through the 2C-state is critical to restore the full developmental capacity in ESCs (Amano et al. 2013). We treated ESCs with BMP4, which induced 2C-associated genes (Fig. 5D) as well as the trophoblast stem cells markers Cdx2 and Elf5 (Fig. 5E) supporting that ESCs employ the 2C-like state during transdifferentiation. Induction of both 2C- and trophoblast markers was greatly reduced in TKO-ESCs. Moreover, expression of placental markers induced by BMP4 (e.g. Serpin, Psg and Prl gene families) was almost abolished in TKOs (Fig. 5F). The reduced transdifferentiation potential of Gadd45 TKO-ESCs is consistent with impaired 2C-like cycling. Interestingly, Dnmt1 deficiency yields the opposite phenotype to Gadd45 deficiency, i.e. activation of trophoblast lineage markers (Cambuli et al. 2014). Our data are therefore in line with the notion that DNA methylation is a barrier to ESC- to trophoblast transdifferentiation (Ng et al. 2008).

(Gadd45a/Gadd45b) -/- mice are sublethal and show partially impaired ZGA gene expression

Interrogating a database of early mouse transcriptome (Park et al. 2015) revealed that Gadd45a and Gadd45b belong to a “2-Cell Transient” cluster, showing a peak of expression specifically in 2-cell embryos (Fig. 6A). Gadd45g belongs to the “Major ZGA cluster” but shows low expression during cleavage stages. This raised the possibility that the impairment of 2C cycling in ESCs, in fact, reflects a role for GADD45 in the embryonic 2-cell stage, coinciding with the major phase of zygotic genome activation (Eckersley-Maslin et al. 2018). Single Gadd45a, -b, or -g mutants are viable and fertile (Hollander et al. 1999; Lu et al. 2001; Lu et al. 2004) but our results in ESCs indicate that they may compensate for each other in 2C-state regulation. As generation of triple mutants is challenging, we generated Gadd45a,b double mutant (DKO) mice by intercrossing double heterozygous animals. All nine possible genotypes were obtained at expected Mendelian ratios, except for the homozygous Gadd45a,b double mutants, whose frequency was 50% reduced (Fig. 6B, arrow). Surviving DKO mice showed normal body size, but displayed phenotypic abnormalities characteristic of neural tube closure defects (NTDs) such as curly tail and spina bifida. Litter size of DKO intercrossing was 50% reduced compared to double heterozygous crosses (Fig. 6C). At embryonic stage E13.5, 50% (n=24) of DKO embryos resulting from double heterozygous breeding and 80% (n=37) from homozygous DKO breeding, showed, beyond curly tail and spina bifida, also defects like exencephaly and cranial hemorrhage (Fig. 6D-E, S7H). Sublethality, exencephaly, and cranial hemorrhaging are also observed in Tet1, 2 DKO mice (Dawlaty et al. 2013). The increased phenotypical abnormalities of DKO embryos resulting from breeding homozygous DKOs compared to double heterozygous mice hints at a requirement of GADD45 in the germ line, as is the case for TET1 (Yamaguchi et al. 2013).

To investigate the impact of Gadd45a,b deficiency on gene expression we performed transcriptome analysis of 2-cell stage embryos from Gadd45a,b DKO crosses. Transposable elements (TE) were hardly affected in DKO embryos. There were no downregulated- and few upregulated TE transcripts, including LINE1 elements (L1MC, Lx2B; Fig. 6F). Also the number of deregulated non-repetitive genes in DKOs was limited (n= 104, 10% FDR, Fig. 6G), in accord with subviability of DKO mice. However, misregulated genes showed a clear signature of impaired 2-cell stage entry: First, genes downregulated in DKOs corresponded to genes upregulated in 2-cell stage embryos compared to oocytes (Macfarlan et al. 2012), while genes upregulated in DKOs overlapped with genes upregulated in oocytes (Fig. 6H-I). Second, ZGA genes associated specifically with the 2C-like state (Macfarlan et al. 2012) overlapped with none of the up- but seven downregulated genes in DKOs (Fig. 6J). The enrichment of maternal- and depletion of zygotic transcripts in DKO embryos indicates that GADD45α,□ promote the embryonic 2-cell stage. Hence, we monitored the development of wild type (WT) and DKO preimplantation embryos in vitro. No differences were observed for development until the 2-cell stage between WT and DKO embryos (data not shown). However, only ~40% of the DKO embryos reached 8-cell stage versus ~80% of the wild type embryos (Fig. 6K). The affected DKO embryos remained at 2- or 4-cell stage, or died. Incomplete penetrance of preimplantation defects are also observed in other mutant mice (Narducci et al. 2002). However, the impaired in vitro pre-implantation development supports our findings in mouse embryonic stem cells showing that GADD45 proteins are involved in 2-cell stage regulation.

Discussions

GADD45α,□,ɣ are adaptors for TET/TDG mediated DNA demethylation but in which physiological processes and at which genomic loci they mediate demethylation remains poorly understood. In addition, they show overlapping expression and hence may act functionally redundant, complicating their analysis. Here we present analyses of triple-knockout ESCs and Gadd45a,b DKO mice showing that i) GADD45α,□, Eine der aktuellen, großen Fragen der Biologie ist es zu verstehen, welche genetischen und epigenetischen Faktoren an der Regulation der Genexpression beteiligt sind und in welchen Fällen ihre Deregulierung zur Entwicklung von abnormalen Phänotypen oder Krankheiten beitragen kann. Innovationen bei Genomsequenzierungstechniken und entsprechenden Datenverarbeitungsalgorithmen ermöglichten objektive Analysen der verschiedenen genomischen und epigenomischen Komponenten der Transkription in der Auflösung von einzelnen Nukleotiden. Daher ist es jetzt möglich, verschiedene Daten sowohl für Bulk- als auch für Einzelzellproben zu integrieren und die molekularen Komponenten der Genexpressionsregulation zu verstehen, mithilfe einer reproduzierbaren rechnerischen ad-hoc-Analyse.

Als interdisziplinäres Feld nutzt die Bioinformatik verschiedene quantitative Disziplinen wie Statistik und maschinelles Lernen. Dies ermöglicht die Implementierung von detaillierten Analysen zur Unterstützung und Aufklärung spezifischer, fundamentaler Entdeckungen, sowie zur Prüfung unerwarteter Vorhersagen, die sich aus der explorativen Datenanalyse ergeben. Insbesondere ist die Bioinformatik notwendig für die Untersuchung der genomischen Grundlagen der Genregulation angesichts der Komplexität der erzeugten Daten. Die Anwendung bestehender und die Entwicklung neuer bioinformatischer Methoden verbessern die Interpretation neuer Daten zu, indem verschiedene Datentypen aus mehreren Quellen integriert werden.

In dieser Dissertation habe ich bioinformatische Methoden angewendet und entwickelt, um grundlegende biologische Fragen in der genomischen Erforschung der epigenetischen Genregulation zu untersuchen: i) habe ich eine Pipeline für die Datenanalyse der Bisulfitsequenzierung im gesamten Genom erstellt, um zu verstehen, wie Gene und DNA-Sequenzen von Gadd45-Proteinen demethyliert werden und wie dies mit einem der wichtigsten Entwicklungsstadien in embryonalen Maus-Stammzellen (mESCs) zusammenhängt, ii) entwickelte ich eine Metrik auf der Grundlage des Gini-Index, um die Ergebnisse von unüberwachten Clusterings von verschiedenen Berechnungsmethoden zu bewerten, die zur Trennung von periphere mononukleäre Blutzellen (PBMCs) von einzelzell-ATAC-seq-Proben angewendet wurden, von denen die Markierungen der Zellen nicht vorhanden waren, und iii) habe ich einen Algorithmus entwickelt, mit dem variable Regionen in ChIP-seq-Daten extrahiert werden können, um die Identifizierung Protein-spezifischer Bindungsstellen in mehreren Zelllinien des ENCODE-Projekts zu verbessern. Zusammen sind diese drei Studien ein

signifikanter Beitrag durch die Bioinfromatik zur Verbesserung der Interpretation von Genomdaten in Hinblick auf die Untersuchung der epigenetischen Genregulation.

are not required to maintain pluripotency and self-renewal in ESCs. Yet, ii) GADD45 proteins are required for locus-specific demethylation of ~7,000 sites, notably on enhancers and at sites harboring oxidized 5mC; iii) Gadd45-mutant ESCs display methylation-related gene misexpression; iv) GADD45 proteins promote the 2C-like ESC state and 2-cell embryo stage, regulating a subset of ZGA-specific genes.

GADD45 proteins are not required for pluripotency but for differentiation of ESCs

We find that Gadd45 TKO-ESCs remain pluripotent and self-renew, as has been found for other DNA demethylation deficient ESCs, such as Tet1,2,3 and Tdg knockout ESCs (Cortázar et al. 2011; Dawlaty et al. 2014). A previous study (Li et al. 2015) reported that also Gadd45a,b double knockout ESCs remain pluripotent. In agreement with this study, overall oxidized cytosine levels were mostly unchanged in Gadd45 TKO-ESCs. This is expected, because GADD45α functions in locus-specific, rather than global demethylation (Arab et al. 2014; Arab et al. 2019; Schäfer et al. 2013; Schmitz et al. 2009). Our rescue experiments indicate that all three GADD45 proteins can compensate for the loss of all three genes, supporting their functional redundancy.

While GADD45 proteins were dispensable for overall ESC maintenance, a subset of genes was downregulated and hypermethylated in TKO-ESCs. Moreover, DNA-hypomethylating 2i, vitamin C, and 5-azadeoxycytidine treatment rescued this downregulation, suggesting that it was the direct or indirect consequence of DNA hypermethylation. Similarly, previous reduced representation bisulfite sequencing identified 68 hypermethylated but no hypomethylated loci in Gadd45a,b DKO-ESCs (Li et al. 2015). Most of these sites overlapped with 5hmC- and 5fC-enriched regions, corroborating a role in DNA demethylation. Taken together with the overlap between the genes misregulated in Gadd45 and Tet mutants, this supports the conclusion that these two protein families cooperate in enzymatic DNA demethylation (Arab et al. 2014; Arab et al. 2019; Kienhöfer et al. 2015; Li et al. 2015).

Even though GADD45 proteins act locally and DNMTs globally, there was a significant overlap between genes misregulated in Gadd45 TKO-ESCs and Dnmt TKO-ESCs. However, despite millions of genomic sites that become unmethylated in Dnmt TKO-ESCs, including at least 6,100 promoters, only a few hundred genes are actually de-repressed (Karimi et al. 2011). These ‘hotspot’ genes misregulated in Dnmt TKO-ESCs are enriched for a role in germ cell development and localize on the X-chromosome. Likewise, genes in the MageA and Rhox clusters that are prominently reactivated in the DNMT TKO-ESCs, are downregulated in Gadd45 TKO-ESCs. Thus, even though

DNMTs act globally, respective gene expression changes in ESCs only occur on hotspots, which are prone to react to DNA methylation changes, such as germ cell-specific genes. Indeed, TET1 is a prominent regulator of gene expression in germ cells (Hill et al. 2018; Yamaguchi et al. 2012), suggesting that ESCs recapitulate aspects of germ cell gene regulation via DNMT-TET mediated methylation-demethylation.

In contrast to a limited role of GADD45 proteins during ESC pluripotency, gene deregulation in TKOs was greatly increased upon EB- and monolayer differentiation. Neuronal genes were particularly affected, corroborating a role for Gadd45 genes during neural differentiation (Huang et al. 2010; Kaufmann and Niehrs 2011; Ma et al. 2009). Indeed, while Gadd45a,b DKO mice were viable, they were sublethal, with embryos showing gross abnormalities including defects in neural tube closure and brain hemorrhage. This phenocopies Tet1,2 DKOs, which are also sublethal, with affected embryos showing exencephaly and cranial hemorrhaging, and where the defects were attributed to reduced 5hmC, increased 5mC, and aberrant imprinting (Dawlaty et al. 2013).

TET/TDG processed sites are main targets of GADD45 mediated DNA demethylation

Our methylome analysis of TKO-ESCs revealed ~7,000 hyper-DMRs, the greatest number of locus-specific methylation changes so far reported to be associated with GADD45 function. Since only 34 hypo-DMRs were detected (arguing against generalized methylation-misregulation), and since the most significant association of the hyper-DMRs was with oxidized 5mC, we conclude that the hyper-DMRs arose as the result of impaired enzymatic DNA demethylation. This is consistent with direct interaction of GADD45α with both TET1 and TDG (Barreto et al. 2007; Cortellino et al. 2011; Kienhöfer et al. 2015; Li et al. 2015). Moreover, GADD45-dependent hyper-DMRs are enriched at TET-dependent hyper-DMRs (Lu et al. 2014) with the caveat that the number of called DMRs is not directly comparable, due to differences in ESC lines and in DMR-calling algorithms used. Consistently, we reach similar conclusions for Gadd45 TKO-ESCs as have been reported for Tet1,2,3 TKO-ESCs (Lu et al. 2014). First, the most important concordance of hyper-DMRs was with sites marked by oxidized 5mC derivatives. Enrichment of hyper-DMRs at TDG 5fC- and 5caC target sites concurs with the conclusion of a dual role of GADD45α, to stimulate TET1 chemical processivity for iterative 5mC oxidation, and to enhance TDG processing of 5fC/5caC (Kienhöfer et al. 2015; Li et al. 2015). Second, enhancers are the most enriched genomic target of both GADD45 and TET-mediated DNA demethylation. Enhancers in ESCs tend to be hypomethylated (Kieffer-Kwon et al. 2013; Stadler et al. 2011; Ziller et al. 2013) and thus GADD45 plays a significant role in this phenomenon. Enhancers are also the main target of GADD45α mediated DNA demethylation in mouse embryonic fibroblasts (Schäfer et al. 2018).

Both in Tet- and Gadd45-TKO-ESCs, enhancer hypermethylation affected expression of relatively few genes, consistent with the minor role of 5mC in gene repression of early embryos and ESCs (Bogdanovic et al. 2011; Fouse et al. 2008). The regulatory role of DNA demethylation for Gadd45 may manifest more upon later cell differentiation, as e.g. mutation of Dnmt and Tet in ESCs leads to differentiation defects (Dawlaty et al. 2014; Jackson et al. 2004) and to lethality in whole embryos (Dai et al. 2016; Okano et al. 1999). Hence, the dynamics of DNA methylation in ESCs may reflect a role as molecular memory for subsequent enhancer regulation in differentiating cells, rather than in regulation of acute gene expression (Kim et al. 2018).

GADD45 proteins promote a 2C-like state and ZGA-specific gene expression

A main finding of this study is that Gadd45 TKO-ESCs displayed impaired cycling into the 2C-like state and Gadd45a,b DKOs showed partial gene misregulation at the 2-cell stage, as well as embryonic sublethality with developmental stage arrest. This suggests that 2C-cycling defects in TKO-ESCs mirror a role of GADD45 during 2-cell stage development and ZGA. The 2C-like state of ESCs recapitulates key aspects of the 2-cell stage mouse embryo, both phenotypically and molecularly (reviewed in (Eckersley-Maslin et al. 2018; Ishiuchi and Torres-Padilla 2013)). Concordantly, Gadd45 expression peaks in 2-cell stage mouse embryos and while Gadd45a,b DKOs display only moderate gene misregulation, our results may underestimate their role, as Gadd45g could partially compensate for Gadd45a,b deficiency in DKOs (Bogdanovic et al. 2011).

Entry into the 2C-like state is accompanied by genome-wide DNA demethylation (Dan et al. 2017; Eckersley-Maslin et al. 2016). However, ESCs cultured in serum are in a distinct, potentially artefactual epigenetic state: The de novo DNA methylation machinery, only transiently active in embryos during implantation, is instead continuously operating in ESCs (Brandeis et al. 1994; Lienert et al. 2011). Hence, downregulation of 2C genes in TKO-ESCs might rather reflect a role of GADD45 proteins during early 2-cell stage in protecting against de novo methylation. This is because the genome of 2-cell embryos is already hypomethylated due to global demethylation occurring in zygotes (reviewed in (Eckersley-Maslin et al. 2018; Lee et al. 2014; Messerschmidt et al. 2014)) and requires TET3 to protect against de novo DNA methylation (Amouroux et al. 2016). It is this protection against de novo DNA methylation that may require GADD45.

In apparent contradiction to these findings, transition to the 2C-like state in Tet TKO-ESCs is actually enhanced, as TET proteins repress type III ERVs and 2C-specific genes (Lu et al. 2014). Moreover, Tet TKO mouse embryos develop past implantation stage to gastrulae, albeit with altered expression of a few hundred genes at the blastocyst stage (Dai et al. 2016). If GADD45 act via TETs to affect 2C-like state and 2-cell embryo development, how can this discrepancy be explained? First, Tet TKO mouse embryos were generated by

crossing mice with TET-deficient germ cells (Dai et al. 2016), and hence with zygotes that had adapted to a state completely devoid of TET enzymatic activity. Furthermore, close inspection of Tet1,3 DKO mice showed unexpectedly variable expression of ~150 genes in 8-cell stage embryos, among them the Zscan4 cluster (Kang et al. 2015). In fact, acute Tet1,2,3 knockdown in oocytes leads to developmental arrest at 2-cell stage with severe ZGA gene misregulation (M. Wossidlo, pers. commun.). Second, TETs have a dual function, involving both their catalytic- as well as non-catalytic, gene repressive function, the latter being the dominant gene-regulatory mode of TET1 in ESCs (Williams et al. 2011). The repressive role involves recruitment of KAP1/TRIM28 by TETs, a negative 2C-like state regulator (Lu et al. 2014). Consistent with a dual role, Tet-deficient ESCs and epiblast-derived stem cells can display either up- or downregulation of Zscan4 cluster expression, depending on culture conditions (Khoueiry et al. 2017; Yang et al. 2016). Similarly, for LINE1 regulation TETs play a dual role in ESCs, activating and silencing expression via LINE1 promoter demethylation and via recruiting the SIN3A repressor, respectively (La Rica et al. 2016).

In analogy, TET enzymes may play a dual role in the regulation of the key developmental transition at the 2-cell stage: First, a repressive one, whereby possibly TET represses LINE1 elements, whose transcripts silence Dux expression and thereby the 2-cell state (Percharde et al. 2018). Second, a promoting one, as supported by our study, whereby GADD45 targets TETs to promote the 2-cell stage via demethylation and protection against de-novo methylation.

Material & Methods

Statistic

Unless otherwise indicated, statistical significance was tested with unpaired, two-tailed Student's t-test with three biological replicates. Unless otherwise indicated, data are presented as mean ± standard deviation from three independent clones. *p<0.05, **p<0.005, ***p<0.0005.

Gene ontology (GO) enrichment analysis for the ontology "biological process" was carried out using GOrilla (Eden et al. 2009), using the parameter "Two unranked lists of genes" and the whole mouse transcriptome as a background list.

Commonly deregulated genes were identified using publicly available datasets. Overlaps were identified using ENSEMBL or RefSeq IDs if available, and gene symbols if not. Statistical significance of overlaps was determined using hypergeometric test. Only genes detectable by both platforms employed were considered. For scatter plots, only genes which were significantly deregulated in both respective studies are shown.

Quantitative PCR

Total RNA was isolated using a Qiagen RNeasy mini kit with on-column DNase digest (Qiagen). First strand cDNA was generated using SuperScript II reverse transcriptase (Invitrogen). Real-time PCR was performed in technical duplicates using Roche LightCycler480 probes master and primers in combination with predesigned monocolor hydrolysis probes of the Roche Universal Probe Library (UPL). For quantification, Roche LC480 quantification software module was used. Expression levels were normalized to Gapdh or Tbp expression first, followed by normalization to control conditions specific to the individual experiment, as indicated.

Quantitative mass spectrometry of DNA modifications

Genomic DNA sample preparation and quantification of 5mC and its oxidative derivatives was carried out as described before (Schomacher et al. 2016).

ESC culture, treatment and differentiation

ESCs were cultured on gelatinized cell culture vessels in LIF-conditioned DMEM supplemented with 15% ESC-grade FBS, 2mM L-Glutamine, 50U/ml Penicillin/Streptomycin, 1xNEAA, 1mM sodium pyruvate, 100µM β-mercaptoethanol, at 37°C, 5% CO 2 and 21% O 2 , with daily medium changes. Cells were passaged every 2 days once reaching approximately 60% confluency with a ratio of 1:8. For ESC culture on mouse embryonic fibroblast (MEF) feeder, ESCs were plated in a 12-well preplated with CF-1 MEF feeder (ATCC).

For growth curve analysis cells were counted in duplicates using a TC10 automated cell counter (Biorad). For EB differentiation, 3.5x10 6 mESCs were plated on non-adherent 10cm bacterial dishes (Greiner) in 15ml CA medium (Bibel et al. 2007). CA medium was changed every other day. For retinoic acid induced EB differentiation, CA medium was supplemented with retinoic acid (final concentration 5µM) at days 4 and 6 of differentiation. EBs were harvested after 8 days of differentiation.

For monolayer differentiation, approximately 5.000 ESCs per cm 2 were plated on gelatinized cell culture plates in regular ESC medium on the evening before differentiation. On the next morning, cells were washed twice with PBS and medium was changed to N2B27 medium (50% Advanced DMEM/F12, 50% Neurobasal, 2mM L-Glutamine, 50U/ml Penicillin/Streptomycin, 50µg/ml BSA, 0.5x N2-supplement (Invitrogen), 0.5x B27-supplement (Invitrogen).

Medium was changed on days 3 and 5 of differentiation. Cells were harvested after 6 days of differentiation.

For 2i treatment, 230.000 ESCs were plated per 6-well in regular ESC medium. On the next day, cells were washed twice with PBS and cultivated up to 72h in 2i medium (N2B27 medium, 1µM PD0325901, 3µM CHIR99021, 4% LIF-supernatant).

For vitamin C treatment, 230.000 ESCs per well were plated in 6-well-plates in regular ESC medium. The next day, vitamin C (L-ascorbic acid 2-phosphate sesqui-magnesium salt) was added to a final concentration of 100µg/ml. Cells were incubated in vitamin C containing medium for up to 72h and passaged once during that time.

For 5-azacytidine treatment, ESCs were incubated in 10µM 5-deoxyazacytidine in ESC medium for 48h. For transdifferentiation, ESCs were resuspended in DMEM supplemented with 15% Knockout Serum Replacement (Invitrogen), 2mM L-Glutamine, 50U/ml Penicillin/Streptomycin, 1xNEAA, 1mM Na-pyruvate, 100µM β-mercaptethanol medium and 10ng/ml recombinant hBMP-4 (R&D Systems) after passaging and plated at a density of 10 4 cells per cm2 on gelatinized cell culture vessels. Medium was changed on days 2, 4, 6 and 7 of transdifferentiation. Cells were harvested after 8 days.

For transient overexpression, preplated ESCs were transfected with Lipofectamine 2000 (Thermo-Fischer) according to the manufacturer´s recommendations. Medium was renewed on the next day and cells were harvested 48h post transfection.

CRISPR/Cas9 mediated knockout and introduction of stable expressing Zscan4c::eGFP

1.45x10 6 ESCs were seeded on 10cm dishes each and transfected on the next day with either i) 13.5µg empty px330 vector and 1.5µg pPuro, ii) 2.25µg of each of the six guide RNA containing plasmids (see Supplemental Material and Methods) and 1.5µg pPuro using 45µl Lipofectamine 2000 (Thermo-Fischer) in a volume of 1.5ml OptiMEM, but otherwise as decribed above. ESCs were passaged from 1x10cm to 2x15cm dishes the following day. Cells were selected with 2µg/ml puromycin from 48h post-transfection and kept under selection until the day of freezing. Forming ESC colonies were transferred to gelatinized 48-well plates 6 days after passaging. On the next day, colonies were washed once with 500µl PBS, dissociated with 100µl 0.25% trypsin for 90s at 37°C, quenched with 800µl of ESC medium and plated on a new 48-well plate. Three days later, cells were passaged and partially used for genotyping PCR (Primers see Supplemental Material and Methods). For transfection of Zscan4c::eGFP 1x10 6 ESCs were seeded on gelatinized 10cm dishes and transfected 4 hours after seeding using Lipofectamine 2000 (Thermo-Fischer). Cells were selected with 5µg/ml blasticidin 48h post-transfection for eleven days, expanded and frozen for

further analysis. For transfection of 2C::tdTomato reporter (Addgene #40281) 1x10 6 ESCs were seeded on gelatinized 10cm dishes and transfected 4 hours after seeding using Lipofectamine 2000 (Thermo-Fischer). Cells were selected with 150µg/ml hygromycin 48h post-transfection for seven days, expanded and frozen for further analysis.

Plasmids

pZscan4-Emerald was a kind gift from M. Ku (Zalzman et al. 2010). pCW57.1_Luciferase and pCW57.1_mDuxCA-3xHA were a kind gift from B.Cairns (Hendrickson et al. 2017). The expression constructs in this study were pCS2+FLAG_hTDG (Schomacher et al. 2016), pCS2+-GFP (Barreto et al. 2007), pCS2+myc-MmGadd45a (M. Gierl, unpubl.), pCS2+myc-MmGadd45b(M. Gierl, unpubl.), pCS2+myc-MmGadd45g (Gierl et al. 2012), pRKW2-mTet2 (Ko et al. 2010). The catalytic domain of mouse Tet1 was inserted into pCS2+ as N-terminal HA-tag expression construct (A. Ernst, unpubl.).

Teratoma assays

ESCs were sent cryo-conserved to EPO Berlin GmbH, where they were thawed and passaged twice before transplantation. Cells were resuspended in PBS, mixed with Matrigel and transplanted into the flanks of three NSG mice per mESC clone. Tumor weight and size were measured twice per week. Animals injected with the same ESC clone were sacrificed once the average tumor size reach approximately 1.0 cm 3 . Tumors were excised, weighed and cut. 1/3 rd was shock frozen and 2/3 rds were fixed in formalin. Formalin fixed tumors were paraffin embedded, sectioned, and stained with hematoxylin and eosin.

Flow cytometry analysis and fluorescence activated cell sorting

Cells were detached using 0.25% trypsin and resuspended in PBS containing 2.5% ESC grade and 5mM EDTA pH 8. Suspended cells were then analysed by the BD LSRFortessaSORP flow cytometry system using DiVa software. For downstream analysis on sorted ESCs, cells were sorted according to the fluorescence intensity of eGFP into PBS using BD FACSAria III SORP with a 85 µm nozzle. Data analysis was performed with FlowJo software (version 10.5.3).

Whole-genome bisulfite sequencing

Genomic DNA from Control and Gadd45 TKO-ESCs was purified using a QIAamp DNA Mini kit (Qiagen) according to the manufacturer´s recommendations. An additional RNAseA treatment (10mg/ml, Qiagen) was done after cell lysis. WGBS library preparation was carried out using the TruSeq PCR-Free Library Prep Kit (LT) and the Epitect Kit (Qiagen) for bisulfite conversion. Sequencing was performed on Illumina HiSeq X in paired-end mode (PE150).

RNA sequencing

ESC samples

Total RNA was isolated using the RNeasy mini kit (Qiagen) with on-column DNase digest. NGS library preparation was performed using Illumina's TruSeq Stranded mRNA HT Sample Prep Kit with dual-indexing following the standard protocol (Illumina Part # 15031047 Rev. D). Libraries were profiled with a DNA 1000 chip on an Agilent 2100 Bioanalyzer and quantified using the Qubit dsDNA HS Assay Kit on a Qubit 2.0 Fluorometer (Life Technologies). All 36 samples were pooled in equimolar ratio and sequenced on 8 HiSeq 2000 lanes for 35 cycles plus additional 16 cycles for the i7 and i5 index reads.

Isolation and culture of preimplantation embryos

For RNA-sequencing, 2-cell stage embryos were collected from 3-week old wild type (n=4) and (Gadd45a/Gadd45b) -/- (n=4) superovulated females 20 h after the appearance of the vaginal plug. 2-cell stage embryos coming from the same litter were pooled and considered as one biological sample. Embryos were collected in M2 medium (Sigma) supplemented with 0.3 mg/ml hyaluronidase (Sigma), washed twice with PBS and directly transferred into lysis buffer (Smart Seq v4 Ultra Low Input RNA Kit for Sequencing, Takara). Sample preparation was done according to the manufacturer's recommendations. cDNA was amplified using 11 PCR cycles. NGS library preparation was performed using NuGEN´s Ovation Ultralow System V2 1-96 (2014). Libraries were prepared with a starting amount of 2,36ng of fragmented cDNA and were amplified in 11 PCR cycles. Libraries were profiled in a High Sensitivity DNA on a 2100 Bioanalyzer (Agilent technologies) and quantified using the Qubit dsDNA HS Assay Kit, in a Qubit 2.0 Fluorometer (Life technologies). All 8 samples were pooled in equimolar ratio and sequenced on one NextSeq 500 Highoutput FC, SR for 85 cycles plus 7 cycles for the index read.

For the in vitro development assay, mouse preimplantation embryos were collected as described above and then cultured in EmbryoMax® Human Tubal Fluid medium (Millipore) at 37 °C, 5% CO 2 for 48 h. The experiment was performed in biological triplicates using three independent breeding for wild type and (Gadd45a/ Gadd45b) -/- animals.

Animal experiments

Gadd45a and Gadd45b knockout mice were kindly provided by M. C. Hollander (Gupta et al. 2005; Hollander et al. 1999). Both strains were backcrossed several generations into the C57BL/6N background and interbred to obtain Gadd45a +/- / Gadd45b +/- (DHet) mice, which were further intercrossed to generate wild type, Gadd45a +/- / Gadd45b +/- and Gadd45a

-/- / Gadd45b -/- (DKO) animals from homogeneous genetic background. Mice were housed under 12:12 light/dark cycles and provided with ad libitum food and water, in accordance with national and European guidelines. For embryo isolation at stage E13.5, timed matings were set up between DHet mice and DKO animals. All procedures were performed with the approval of the ethical committees on animal care and use of the federal states of Rheinland-Pfalz, Germany.

Data availability

All NGS data have been deposited in the NCBI's Gene Expression Omnibus (GEO) as private data under superseries accession number GSE127720 (temporary token: orixkisqnhstpyd; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127720).

Acknowledgements

We thank C. Christopoulou for assistance with the rescue experiments. We thank M. C. Hollander for Gadd45a and Gadd45b mutant mice. B. Cairns, M. Ku, A. Rao and H. Richly kindly provided reagents. We thank C. Scholz for technical support. Contributions by the IMB Core Facilities Flow Cytometry, Genomics, Proteomics, Microscopy and DKFZ High Throughput Sequencing Unit are gratefully acknowledged. M.U.M was supported by Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (NSERC-PDF). This work was supported by an ERC advanced grant ("Demethylase").

Author contributions

K.M.S. conceived and conducted most 2C related experiments. M.L. generated and characterized the TKO-ESCs. V.V. carried out all knock-out mouse analyses. T.A., M.M. and E.K. carried out bioinformatics analyses. M.U.M. performed LC-MS/MS measurements. All authors analyzed and discussed the data. C.N. conceived and coordinated the study and wrote the paper with contribution from K.M.S and M.L.

References

Amano T, Hirata T, Falco G, Monti M, Sharova LV, Amano M, Sheer S, Hoang HG, Piao Y, Stagg CA, et al. 2013. Zscan4 restores the developmental potency of embryonic stem cells, Nat Commun 4: 1966.

Amouroux R, Nashun B, Shirane K, Nakagawa S, Hill PW, D'Souza Z, Nakayama M, Matsuda M, Turp A, Ndjetehe E, et al. 2016. De novo DNA methylation drives 5hmC accumulation in mouse zygotes, Nat Cell Biol 18: 225–233.

Arab K, Karaulanov E, Musheev M, Trnka P, Schäfer A, Grummt I, Niehrs C. 2019. GADD45A binds R-loops and recruits TET1 to CpG island promoters, Nat Genet.

Arab K, Park YJ, Lindroth AM, Schäfer A, Oakes C, Weichenhan D, Lukanova A, Lundin E, Risch A, Meister M, et al. 2014. Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A, Mol Cell 55: 604–614.

Barreto G, Schäfer A, Marhold J, Stach D, Swaminathan SK, Handa V, Döderlein G, Maltry N, Wu W, Lyko F, et al. 2007. Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation, Nature 445: 671–675.

Beddington RS, Robertson EJ. 1989. An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo, Development 105: 733–737.

Bibel M, Richter J, Lacroix E, Barde Y-A. 2007. Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells, Nat Protoc 2: 1034–1043.

Blaschke K, Ebata KT, Karimi MM, Zepeda-Martínez JA, Goyal P, Mahapatra S, Tam A, Laird DJ, Hirst M, Rao A, et al. 2013. Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells, Nature 500: 222–226.

Bogdanovic O, Long SW, van Heeringen SJ, Brinkman AB, Gómez-Skarmeta JL, Stunnenberg HG, Jones PL, Veenstra GJC. 2011. Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis, Genome Res 21: 1313–1327.

Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S. 2012. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution, Science 336:934–937.

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H. 1994. Sp1 elements protect a CpG island from de novo methylation, Nature 371: 435–438.

Cambuli F, Murray A, Dean W, Dudzinska D, Krueger F, Andrews S, Senner CE, Cook SJ, Hemberger M. 2014. Epigenetic memory of the first cell fate decision prevents complete ES cell reprogramming into trophoblast, Nat Commun 5: 5538.

Choi YJ, Lin C-P, Risso D, Chen S, Kim TA, Tan MH, Li JB, Wu Y, Chen C, Xuan Z, et al. 2017. Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells, Science 355.

Cortázar D, Kunz C, Selfridge J, Lettieri T, Saito Y, MacDougall E, Wirz A, Schuermann D, Jacobs AL, Siegrist F, et al. 2011. Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability, Nature 470: 419–423.

Cortellino S, Xu J, Sannai M, Moore R, Caretti E, Cigliano A, Le Coz M, Devarajan K, Wessels A, Soprano D, et al. 2011. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair, Cell 146: 67–79.

Dai H-Q, Wang B-A, Yang L, Chen J-J, Zhu G-C, Sun M-L, Ge H, Wang R, Chapman DL, Tang F, et al. 2016. TET-mediated DNA demethylation controls gastrulation by regulating Lefty-Nodal signalling, Nature 538: 528–532.

Dan J, Li M, Yang J, Li J, Okuka M, Ye X, Liu L. 2013. Roles for Tbx3 in regulation of two-cell state and telomere elongation in mouse ES cells, Sci Rep 3: 3492.

Dan J, Rousseau P, Hardikar S, Veland N, Wong J, Autexier C, Chen T. 2017. Zscan4 Inhibits Maintenance DNA Methylation to Facilitate Telomere Elongation in Mouse Embryonic Stem Cells, Cell Rep 20:1936–1949.

Dawlaty MM, Breiling A, Le T, Barrasa MI, Raddatz G, Gao Q, Powell BE, Cheng AW, Faull KF, Lyko F, et al. 2014. Loss of Tet enzymes compromises proper differentiation of embryonic stem cells, Dev Cell 29:102–111.

Dawlaty MM, Breiling A, Le T, Raddatz G, Barrasa MI, Cheng AW, Gao Q, Powell BE, Li Z, Xu M, et al. 2013. Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development, Dev Cell 24: 310–323.

Dawlaty MM, Ganz K, Powell BE, Hu Y-C, Markoulaki S, Cheng AW, Gao Q, Kim J, Choi S-W, Page DC, et al.2011. Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development, Cell Stem Cell 9: 166–175.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription, Genes Dev 25: 1010–1022.

Eckersley-Maslin M, Alda-Catalinas C, Blotenburg M, Kreibich E, Krueger C, Reik W. 2019. Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program, Genes Dev.

Eckersley-Maslin MA, Alda-Catalinas C, Reik W. 2018. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition, Nat Rev Mol Cell Biol.

Eckersley-Maslin MA, Svensson V, Krueger C, Stubbs TM, Giehr P, Krueger F, Miragaia RJ, Kyriakopoulos C, Berrens RV, Milagre I, et al. 2016. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs, Cell Rep 17: 179–192.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla. A tool for discovery and visualization of enriched GO terms in ranked gene lists, BMC Bioinformatics 10: 48.

Ema M, Mori D, Niwa H, Hasegawa Y, Yamanaka Y, Hitoshi S, Mimura J, Kawabe Y-i, Hosoya T, Morita M, et al. 2008. Krüppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs, Cell Stem Cell 3: 555–567.

Falco G, Lee S-L, Stanghellini I, Bassey UC, Hamatani T, Ko MSH. 2007. Zscan4. A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells, Dev Biol 307: 539–550.

Ficz G, Hore TA, Santos F, Lee HJ, Dean W, Arand J, Krueger F, Oxley D, Paul Y-L, Walter J, et al. 2013. FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency, Cell Stem Cell 13: 351–359.

Fouse SD, Shen Y, Pellegrini M, Cole S, Meissner A, van Neste L, Jaenisch R, Fan G. 2008. Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation, Cell Stem Cell 2: 160–169.

Gavin DP, Kusumo H, Sharma RP, Guizzetti M, Guidotti A, Pandey SC. 2015. Gadd45b and N-methyl-D-aspartate induced DNA demethylation in postmitotic neurons, Epigenomics 7: 567–579.

Gierl MS, Gruhn WH, Seggern A von, Maltry N, Niehrs C. 2012. GADD45G functions in male sex determination by promoting p38 signaling and Sry expression, Dev Cell 23: 1032–1042.

Guo JU, Su Y, Zhong C, Ming G-l, Song H. 2011. Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond, Cell Cycle 10: 2662–2668.

Gupta M, Gupta SK, Balliet AG, Hollander MC, Fornace AJ, Hoffman B, Liebermann DA. 2005. Hematopoietic cells from Gadd45a- and Gadd45b-deficient mice are sensitized to genotoxic-stress-induced apoptosis, Oncogene 24: 7170–7179.

Habibi E, Brinkman AB, Arand J, Kroeze LI, Kerstens HHD, Matarese F, Lepikhov K, Gut M, Brun-Heath I,Hubner NC, et al. 2013. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells, Cell Stem Cell 13: 360–369.

Hayashi K, Sousa Lopes SMC de, Tang F, Lao K, Surani MA. 2008. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states, Cell Stem Cell 3: 391–401.

Hayashi Y, Furue MK, Tanaka S, Hirose M, Wakisaka N, Danno H, Ohnuma K, Oeda S, Aihara Y, Shiota K, et al. 2010. BMP4 induction of trophoblast from mouse embryonic stem cells in defined culture conditions on laminin, In Vitro Cell Dev Biol Anim 46: 416–430.

He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, et al. 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA, Science 333: 1303–1307.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol Cell 38: 576–589.

Hendrickson PG, Doráis JA, Grow EJ, Whiddon JL, Lim J-W, Wike CL, Weaver BD, Pflueger C, Emery BR, Wilcox AL, et al. 2017. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons, Nat Genet 49: 925–934.

Hill PWS, Leitch HG, Requena CE, Sun Z, Amouroux R, Roman-Trufero M, Borkowska M, Terragni J, Vaisvila R, Linnett S, et al. 2018. Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte, Nature 555: 392–396.

Hirata T, Amano T, Nakatake Y, Amano M, Piao Y, Hoang HG, Ko MSH. 2012. Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells, Sci Rep 2:208.

Hollander MC, Sheikh MS, Bulavin DV, Lundgren K, Augeri-Henmueller L, Shehee R, Molinaro TA, Kim KE, Tolosa E, Ashwell JD, et al. 1999. Genomic instability in Gadd45a-deficient mice, Nat Genet 23: 176–184.

Huang HS, Kubish GM, Redmond TM, Turner DL, Thompson RC, Murphy GG, Uhler MD. 2010. Direct transcriptional induction of Gadd45 gamma by Ascl1 during neuronal differentiation, Mol Cell Neurosci 44: 282–296.

Huang Y, Chavez L, Chang X, Wang X, Pastor WA, Kang J, Zepeda-Martínez JA, Pape UJ, Jacobsen SE, Peters B, et al. 2014. Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells, Proc Natl Acad Sci U S A 111: 1361–1366.

Huang Y, Kim JK, Do DV, Lee C, Penfold CA, Zylicz JJ, Marioni JC, Hackett JA, Surani MA. 2017. Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition, Elife 6.

Iaco A de, Planet E, Coluccio A, Verp S, Duc J, Trono D. 2017. DUX-family transcription factors regulate zygotic genome activation in placental mammals, Nat Genet 49: 941–945.

Ishiuchi T, Enriquez-Gasca R, Mizutani E, Bošković A, Ziegler-Birling C, Rodriguez-Terrones D, Wakayama T, Vaquerizas JM, Torres-Padilla M-E. 2015. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly, Nat Struct Mol Biol 22: 662–671.

Ishiuchi T, Torres-Padilla M-E. 2013. Towards an understanding of the regulatory mechanisms of totipotency, Curr Opin Genet Dev 23: 512–518.

Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y. 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine, Science 333: 1300–1303.

Jackson M, Krassowska A, Gilbert N, Chevassut T, Forrester L, Ansell J, Ramsahoye B. 2004. Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells, Mol Cell Biol 24: 8862–8871.

Jarome TJ, Butler AA, Nichols JN, Pacheco NL, Lubin FD. 2015. NF-κB mediates Gadd45β expression and DNA demethylation in the hippocampus during fear memory formation, Front. Mol. Neurosci. 8: 539.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity, Science 337: 816–821.

Jukam D, Shariati SAM, Skotheim JM. 2017. Zygotic Genome Activation in Vertebrates, Dev Cell 42: 316–332.

Kang J, Lienhard M, Pastor WA, Chawla A, Novotny M, Tsagaratou A, Lasken RS, Thompson EC, Surani MA, Koralov SB, et al. 2015. Simultaneous deletion of the methylcytosine oxidases Tet1 and Tet3 increases

transcriptome variability in early embryogenesis, Proc Natl Acad Sci U S A 112: E4236-45.

Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, et al. 2011. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs, Cell Stem Cell 8: 676–687.

Kaufmann LT, Niehrs C. 2011. Gadd45a and Gadd45g regulate neural development and exit from pluripotency in Xenopus, Mech Dev 128: 401–411.

Khoueiry R, Sohni A, Thienpont B, Luo X, Velde JV, Bartoccetti M, Boeckx B, Zwijsen A, Rao A, Lambrechts D, et al. 2017. Lineage-specific functions of TET1 in the postimplantation mouse embryo, Nat Genet 49: 1061–1072.

Kieffer-Kwon K-R, Tang Z, Mathe E, Qian J, Sung M-H, Li G, Resch W, Baek S, Pruett N, Grøntved L, et al. 2013. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation, Cell 155: 1507–1520.

Kienhöfer S, Musheev MU, Stapf U, Helm M, Schomacher L, Niehrs C, Schäfer A. 2015. GADD45a physically and functionally interacts with TET1, Differentiation 90: 59–68.

Kim HS, Tan Y, Ma W, Merkurjev D, Destici E, Ma Q, Suter T, Ohgi K, Friedman M, Skowronska-Krawczyk D, et al. 2018. Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells, Nature 556: 510–514.

Ko M, Huang Y, Jankowska AM, Pape UJ, Tahiliani M, Bandukwala HS, An J, Lamperti ED, Koh KP, Ganetzky R, et al. 2010. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2, Nature 468: 839–843.

Kong L, Tan L, Lv R, Shi Z, Xiong L, Wu F, Rabidou K, Smith M, He C, Zhang L, et al. 2016. A primary role of TET proteins in establishment and maintenance of De Novo bivalency at CpG islands, Nucleic Acids Res 44: 8682–8692.

Kriaucionis S, Heintz N. 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain, Science 324: 929–930.

La Rica L de, Deniz Ö, Cheng KCL, Todd CD, Cruz C, Houseley J, Branco MR. 2016. TET-dependent regulation of retrotransposable elements in mouse embryonic stem cells, Genome Biol 17: 234.

Le Cong, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems, Science 339: 819–823.

Lee HJ, Hore TA, Reik W. 2014. Reprogramming the methylome: erasing memory and creating diversity, Cell Stem Cell 14: 710–719.

Lei H, Oh SP, Okano M, Jüttermann R, Goss KA, Jaenisch R, Li E. 1996. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells, Development 122: 3195–3205.

Leitch HG, McEwen KR, Turp A, Encheva V, Carroll T, Grabole N, Mansfield W, Nashun B, Knezovich JG, Smith A, et al. 2013. Naive pluripotency is associated with global DNA hypomethylation, Nat Struct Mol Biol 20: 311–316.

Li Y, Zhao M, Yin H, Gao F, Wu X, Luo Y, Zhao S, Zhang X, Su Y, Hu N, et al. 2010. Overexpression of the growth arrest and DNA damage-induced 45alpha gene contributes to autoimmunity by promoting

DNA demethylation in lupus T cells, Arthritis Rheum 62: 1438–1447. Li Z, Gu T-P, Weber AR, Shen J-Z, Li B-Z, Xie Z-G, Yin R, Guo F, Liu X, Tang F, et al. 2015. Gadd45a promotes DNA demethylation through TDG, Nucleic Acids Res 43: 3986–3997.

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D. 2011. Identification of genetic elements that autonomously determine DNA methylation states, Nat Genet 43: 1091–1097.

Liu Y, Olanrewaju YO, Zheng Y, Hashimoto H, Blumenthal RM, Zhang X, Cheng X. 2014. Structural basis for Klf4 recognition of methylated DNA, Nucleic Acids Res 42: 4859–4867.

Lu B, Ferrandino AF, Flavell RA. 2004. Gadd45beta is important for perpetuating cognate and inflammatory signals in T cells, Nat Immunol 5: 38–44.

Lu B, Yu H, Chow C, Li B, Zheng W, Davis RJ, Flavell RA. 2001. GADD45 gamma mediates the activation of the p38 and JNK MAP kinase pathways and cytokine production in effector TH1 cells, Immunity 14: 583–590.

Lu F, Liu Y, Jiang L, Yamaguchi S, Zhang Y. 2014. Role of Tet proteins in enhancer activity and telomere elongation, Genes Dev 28: 2103–2119.

Ma DK, Jang M-H, Guo JU, Kitabatake Y, Chang M-l, Pow-anpongkul N, Flavell RA, Lu B, Ming G-l, Song H. 2009. Neuronal Activity-Induced Gadd45b Promotes Epigenetic DNA Demethylation and Adult Neurogenesis, Science 323: 1074–1077.

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity, Nature 487:57–63.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9, Science 339: 823–826.

Messerschmidt DM, Knowles BB, Solter D. 2014. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos, Genes Dev 28: 812–828.

Narducci MG, Fiorenza MT, Kang S-M, Bevilacqua A, Di Giacomo M, Remotti D, Picchio MC, Fidanza V, Cooper MD, Croce CM, et al. 2002. TCL1 participates in early embryonic development and is overexpressed in human seminomas, Proc Natl Acad Sci U S A 99: 11712–11717.

Ng RK, Dean W, Dawson C, Lucifero D, Madeja Z, Reik W, Hemberger M. 2008. Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5, Nat Cell Biol 10: 1280–1290.

Nishiyama A, Xin L, Sharov AA, Thomas M, Mowrer G, Meyers E, Piao Y, Mehta S, Yee S, Nakatake Y, et al. 2009. Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors, Cell Stem Cell 5: 420–433.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development, Cell 99: 247–257.

Parisi S, Passaro F, Aloia L, Manabe I, Nagai R, Pastore L, Russo T. 2008. Klf5 is involved in self-renewal of mouse embryonic stem cells, J Cell Sci 121: 2629–2634.

Park S-J, Shirahige K, Ohsugi M, Nakai K. 2015. DBTMEE. A database of transcriptome in mouse early embryos, Nucleic Acids Res 43: D771-6.

Percharde M, Lin C-J, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M. 2018. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity, Cell 174: 391-405.e19.

Rai K, Huggins IJ, James SR, Karpf AR, Jones DA, Cairns BR. 2008. DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45, Cell 135: 1201–1212.

Rodriguez-Terrones D, Gaume X, Ishiuchi T, Weiss A, Kopp A, Kruse K, Penning A, Vaquerizas JM, Brino L, Torres-Padilla M-E. 2018. A molecular roadmap for the emergence of early-embryonic-like cells in culture, Nat Genet 50: 106–119.

Sabag O, Zamir A, Keshet I, Hecht M, Ludwig G, Tabib A, Moss J, Cedar H. 2014. Establishment of methylation patterns in ES cells, Nat Struct Mol Biol 21: 110–112.

Sakaue M, Ohta H, Kumaki Y, Oda M, Sakaide Y, Matsuoka C, Yamagiwa A, Niwa H, Wakayama T, Okano M. 2010. DNA methylation is dispensable for the growth and survival of the extraembryonic lineages, Curr Biol 20: 1452–1457.

Schäfer A, Karaulanov E, Stapf U, Döderlein G, Niehrs C. 2013. Ing1 functions in DNA demethylation by directing Gadd45a to H3K4me3, Genes Dev 27: 261–273.

Schäfer A, Mekker B, Mallick M, Vastolo V, Karaulanov E, Sebastian D, Lippen C von der, Epe B, Downes DJ, Scholz C, et al. 2018. Impaired DNA demethylation of C/EBP sites causes premature aging, Genes Dev 32: 742–762.

Schmitz K-M, Schmitt N, Hoffmann-Rohrer U, Schäfer A, Grummt I, Mayer C. 2009. TAF12 Recruits Gadd45a and the Nucleotide Excision Repair Complex to the Promoter of rRNA Genes Leading to Active DNA Demethylation, Mol Cell 33: 344–353.

Schomacher L, Han D, Musheev MU, Arab K, Kienhöfer S, Seggern A von, Niehrs C. 2016. Neil DNA glycosylases promote substrate turnover by Tdg during DNA demethylation, Nat Struct Mol Biol 23: 116–124.

Sen GL, Reuter JA, Webster DE, Zhu L, Khavari PA. 2010. DNMT1 maintains progenitor function in self-renewing somatic tissue, Nature 463: 563–567.

Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung H-L, Zhang K, Zhang Y. 2013. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics, Cell 153: 692–706.

Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PWTC, Bauer C, Münzel M, Wagner M, Müller M, Khan F, et al. 2013. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives, Cell 152: 1146–1159.

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions, Nature 480: 490–495.

Svoboda P. 2017. Mammalian zygotic genome activation, Semin Cell Dev Biol. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1, Science 324: 930–935.

Toyooka Y, Shimosato D, Murakami K, Takahashi K, Niwa H. 2008. Identification and characterization of subpopulations in undifferentiated ES cell culture, Development 135: 909–918.

Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S-i, Sakaue M, Matsuoka C, Shimotohno K, Ishikawa F, Li E, Ueda HR, et al. 2006. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b, Genes Cells 11: 805–814.

Wang PJ, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia, Nat Genet 27: 422–426.

Whiddon JL, Langford AT, Wong C-J, Zhong JW, Tapscott SJ. 2017. Conservation and innovation in the DUX4-family gene network, Nat Genet 49: 935–940.

Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PAC, Rappsilber J, Helin K. 2011. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity, Nature 473: 343–348.

Xu H, Ang Y-S, Sevilla A, Lemischka IR, Ma'ayan A. 2014. Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells, PLoS Comput Biol 10:e1003777.

Yamaguchi S, Hong K, Liu R, Shen L, Inoue A, Diep D, Zhang K, Zhang Y. 2012. Tet1 controls meiosis by regulating meiotic gene expression, Nature 492: 443–447.

Yamaguchi S, Shen L, Liu Y, Sendler D, Zhang Y. 2013. Role of Tet1 in erasure of genomic imprinting, Nature 504: 460–464.

Yang J, Guo R, Wang H, Ye X, Zhou Z, Dan J, Wang H, Gong P, Deng W, Yin Y, et al. 2016. Tet Enzymes Regulate Telomere Maintenance and Chromosomal Stability of Mouse ESCs, Cell Rep 15: 1809–1821.

Ying Q-L, Stavridis M, Griffiths D, Li M, Smith A. 2003. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture, Nat Biotechnol 21: 183–186.

Zalzman M, Falco G, Sharova LV, Nishiyama A, Thomas M, Lee S-L, Stagg CA, Hoang HG, Yang H-T, Indig FE, et al. 2010. Zscan4 regulates telomere elongation and genomic stability in ES cells, Nature 464:858–863.

Zhang R-p, Shao J-z, Xiang L-x. 2011. GADD45A protein plays an essential role in active DNA demethylation during terminal osteogenic differentiation of adipose-derived mesenchymal stem cells, J Biol Chem 286: 41083–41094.

Zhang W, Fu S, Liu X, Zhao X, Zhang W, Peng W, Wu C, Li Y, Li X, Bartlam M, et al. 2011. Crystal structure of human Gadd45γ corrected reveals an active dimer, Protein Cell 2: 814–826.

Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, Jager PL de, Rosen ED, Bennett DA, Bernstein BE, et al. 2013. Charting a dynamic DNA methylation landscape of the human genome, Nature 500: 477–481.

Figures

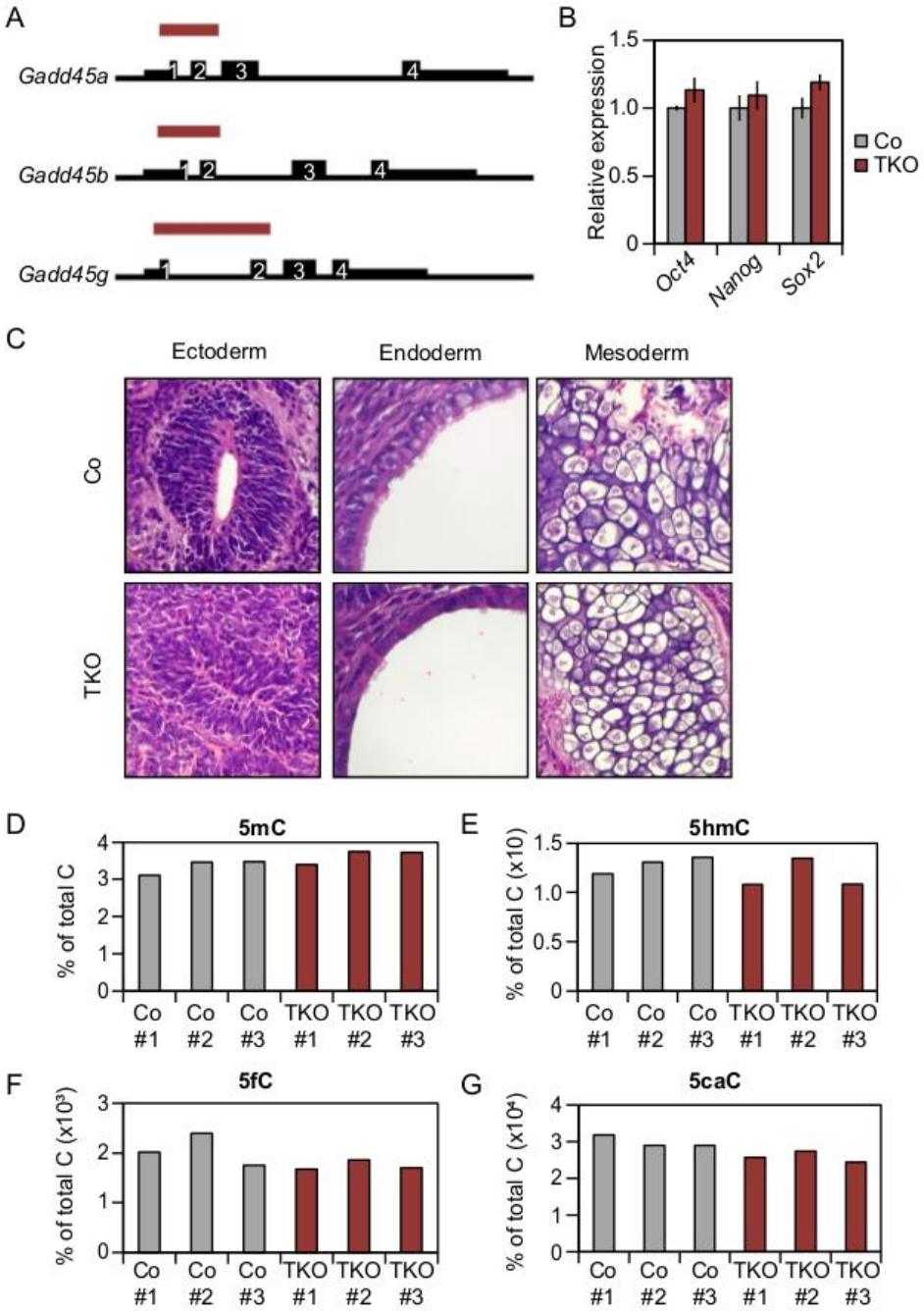

A
Gadd45a
Gadd45b
Gadd45g
1
2
3
4
B
Relative expression
1.5
1.0
0.5
0
Oct4
Nanog
Sox2
Co
TKO
C
Ectoderm
Endoderm
Mesoderm
Co
TKO
D
5mC
% of total C
4
3
2
1
0
Co #1
Co #2
Co #3
TKO #1
TKO #2
TKO #3
E
5hmC
% of total C (x10)
1.5
1.0
0.5
0
F
5fC
% of total C (x10³)
3
2
1
0
G
5caC
% of total C (x10⁴)
4
3
2
1
0

Figure 1: Gadd45 TKO-ESCs are pluripotent and show normal global levels of DNA modifications (A) Scheme of the CRISPR/Cas9-mediated Gadd45 knockout strategy. Numbers, exons; red bars, location of deletion. (B) Relative expression of representative pluripotency markers in Control- (Co) and Gadd45 TKO-ESCs measured by qPCR. Expression values are relative to the average expression in Co-ESCs. (C) Hematoxylin/Eosin staining of paraffin Co and Gadd45 TKO teratoma sections. Representative examples for ectoderm, endoderm and mesoderm derivatives are shown. (D-G) 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) levels in individual Co- and Gadd45 TKO-ESC clones determined by LC-MS/MS. Values are given as % of total cytosine (C).

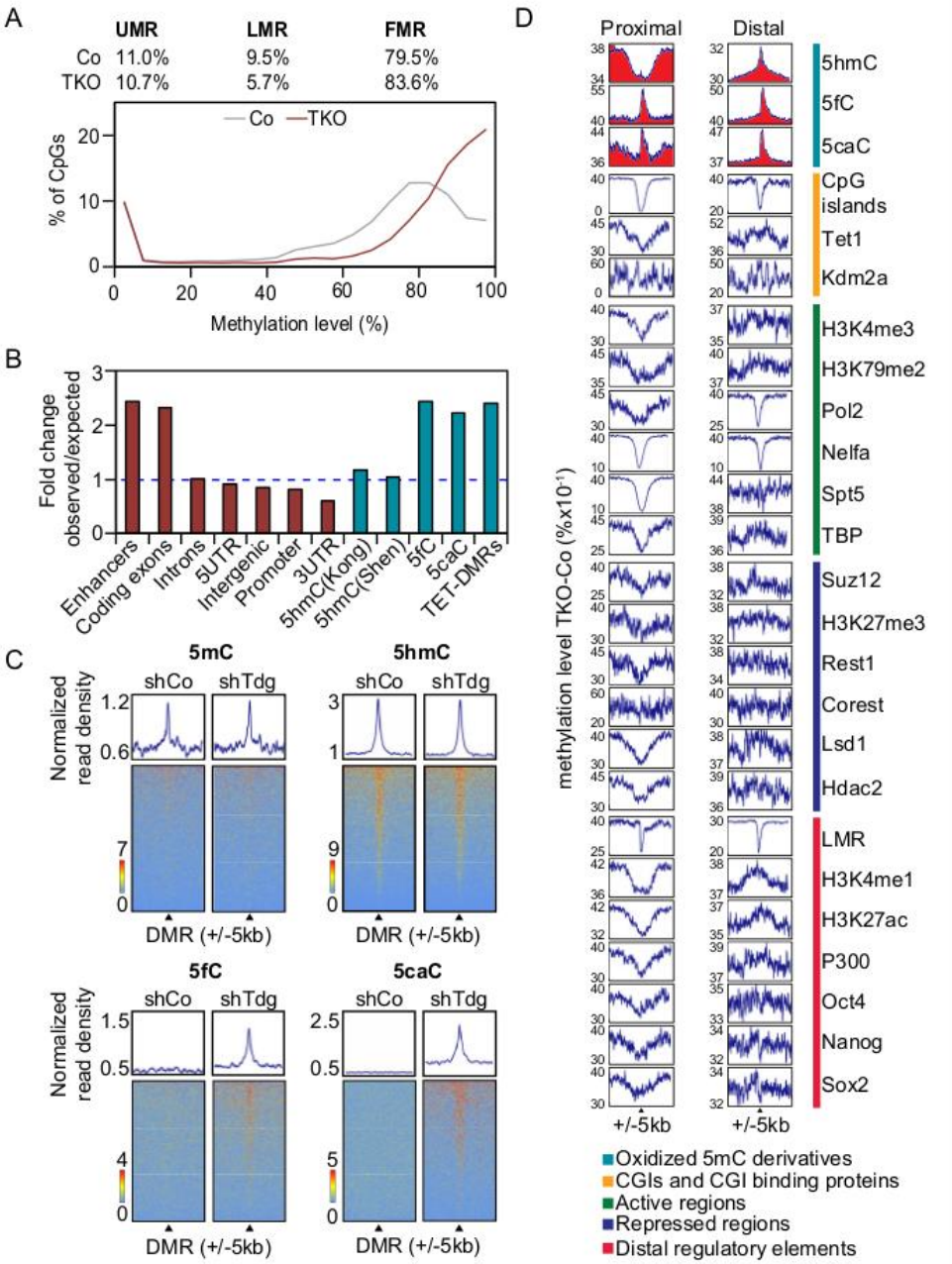


Figure 2. Loci undergoing TET-dependent oxidation are hypermethylated in Gadd45 TKO ESCs. (A) The bimodal methylation pattern of individual CpG sites in ESCs is skewed to a higher methylation level in Gadd45 TKO ESCs. Methylation levels are shown as the average of two biological replicates. (UMR) Unmethylated regions; (LMR) lowly methylated regions; (FMR) fully methylated regions as defined by Stadler et al. (2011). (B) Relative enrichment of Gadd45 TKO hypermethylated DMRs (hyper-DMRs) at various genomic elements (red), oxidative 5mC derivatives, and Tet-TKO hyperDMRs (blue). (C) Heat maps of depicted DNA modifications (Shen et al. 2013) centered (±5 kb) on Gadd45 TKO hyper-DMRs (black triangles) in untreated and sh Tdg-treated ESCs. (D) Average methylation differences between Gadd45 TKO and control ESCs around centers (±5 kb) of annotated genomic features. Methylation differences are shown for proximal (within 1 kb of a gene transcription start site) and distal features. Red areas highlight oxidized 5mC derivatives.

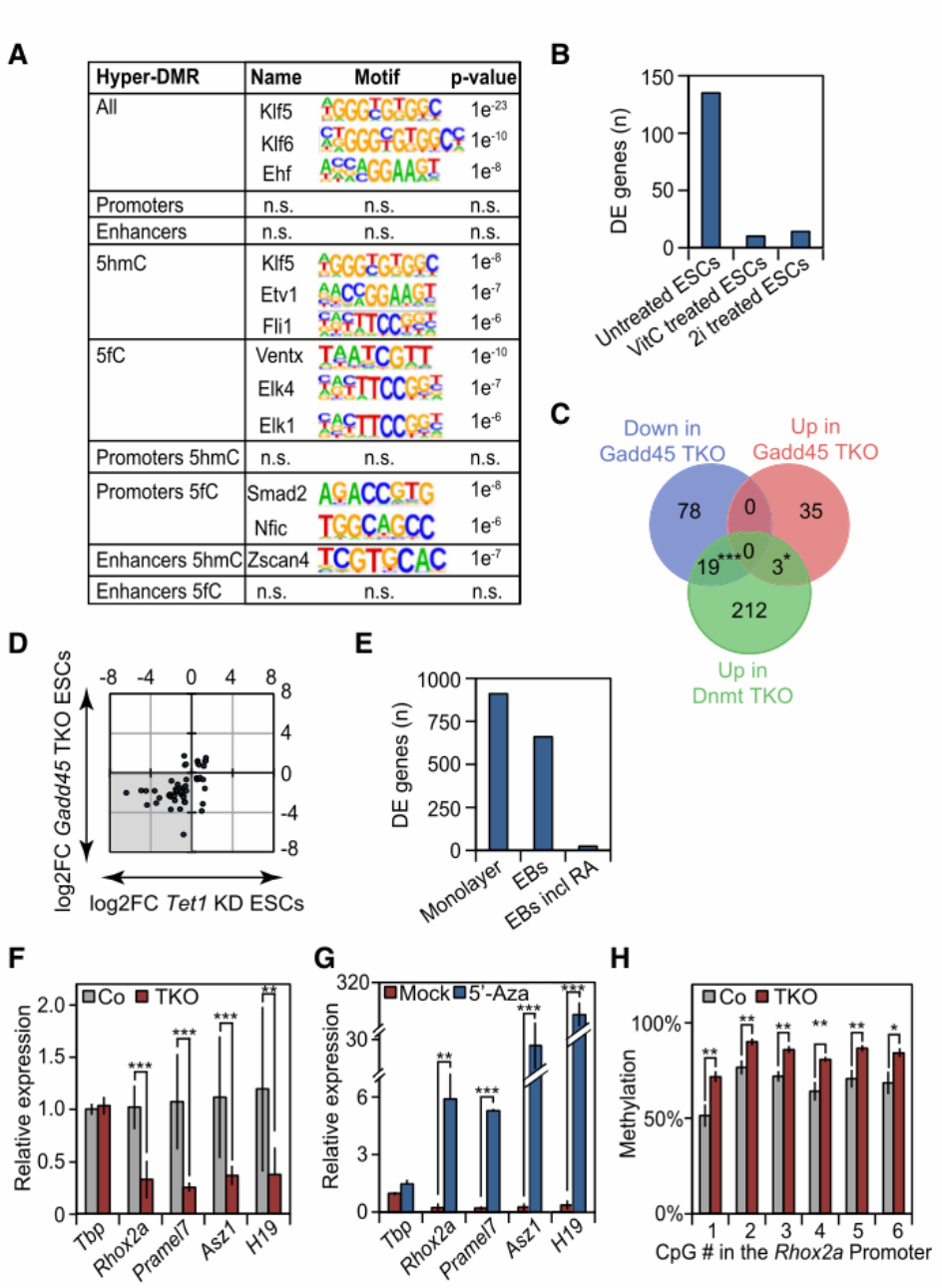


Figure 3: Methylation-regulated genes are downregulated in Gadd45 TKO-ESCs (A) Motif analysis of hyper-DMRs in Gadd45 TKO-ESCs using HOMER (Heinz et al. 2010). (B) Differentially expressed (DE) genes (FDR 10%) identified by RNA-Seq in Gadd45 TKO-ESCs versus Control- (Co) ESCs, which were untreated, 72h vitamin C treated (VitC), or 72h 2i treated. (C) Overlap of genes down- and upregulated in untreated Gadd45 TKO-ESCs and genes upregulated in Dnmt1,2,3 TKO-ESCs (Karimi et al. 2011). (D) Scatterplot of the common deregulated genes (grey) in Gadd45 TKO and Tet1 knockdown ESCs (Log2FC, Log2 of Fold Change versus control ESCs;

Huang et al. 2014). (E) Differentially expressed (DE) genes (FDR 10%) identified by RNA-Seq in Gadd45 TKO cells versus Co cells upon monolayer differentiation (Monolayer), embryoid body differentiation (EBs) or retinoic acid (RA) stimulation during EB differentiation. (F) Expression of selected GADD45-dependent genes in Co- and Gadd45 TKO-ESCs measured by qPCR. Expression is relative to Co-ESCs. Data are presented as mean ± SD from n=3 independent clones and n=3 independent experiments. (G) GADD45-regulated genes are DNA methylation-sensitive. Relative expression levels of selected GADD45-dependent genes in Gadd45 TKO-ESCs upon 48h of DMSO (Mock) or 5'-azadeoxycytidine (5'-Aza) treatment measured by qPCR. Expression is relative to DMSO treated Co-ESCs. (H) Hypermethylation of GADD45-dependent genes. DNA methylation of indicated CpGs in the Rhox2a promoter in Co- and Gadd45 TKO-ESCs monitored by site-specific bisulfite sequencing.

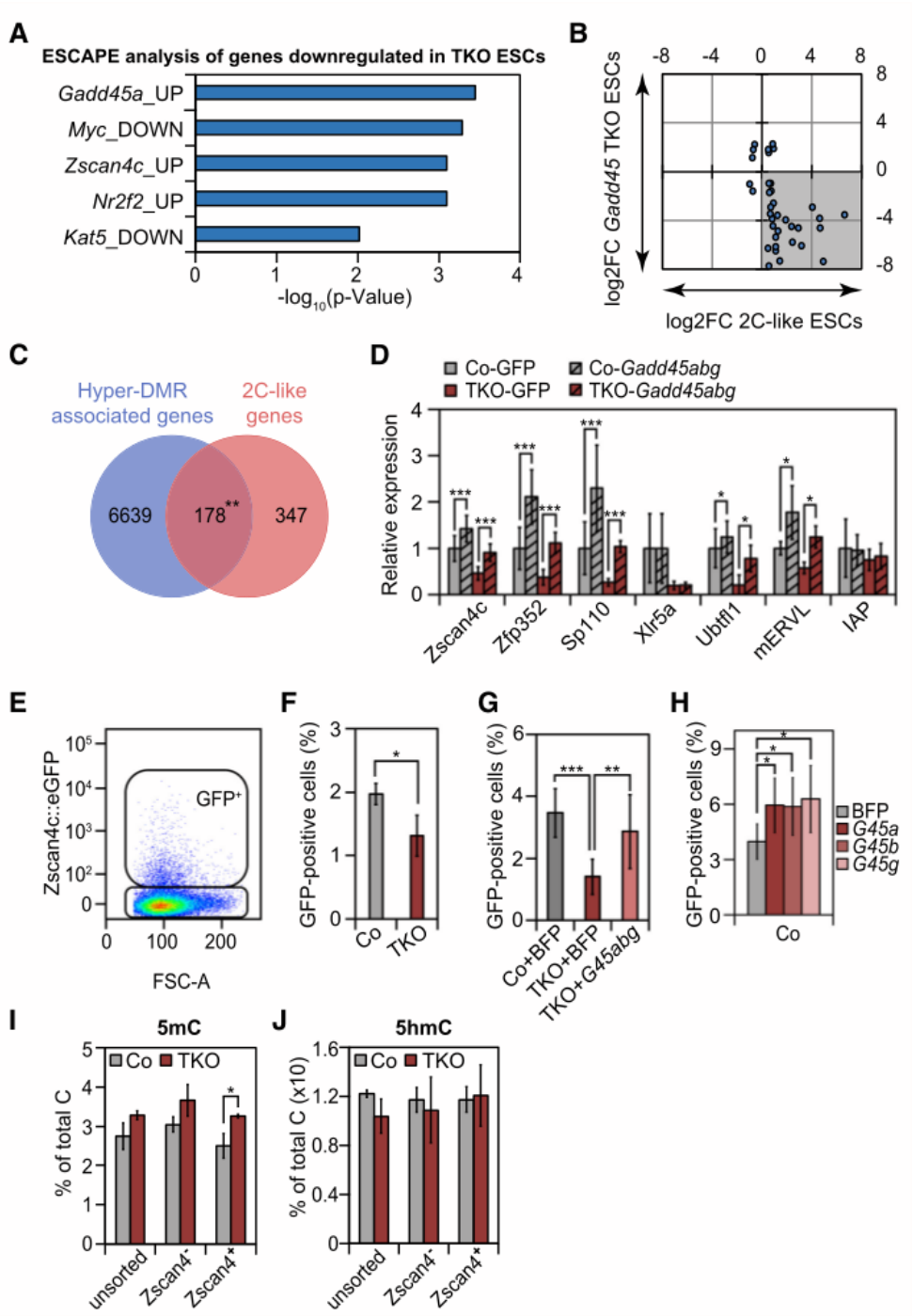


Figure 4: (A) Enrichment analysis of genes downregulated in Gadd45 TKO-ESCs using ESCAPE database (Xu et al. 2014). Enrichment analysis shows a significant overlap with genes deregulated (UP or DOWN) upon overexpression of e.g. Gadd45a or Zscan4c. (B) Scatterplot of the common deregulated genes (grey) in Gadd45 TKO-ESCs and upregulated in the 2C-like state (Log2FC, Log2 of Fold Change versus control ESCs; Macfarlan et al. 2012) (C) Overlap between 2C- (Macfarlan et al. 2012) and hyper-DMR-associated genes. (D) qPCR expression analysis of selected 2C-associated genes and retroviral elements in Control- (Co) and Gadd45 TKO-ESCs, 48h after transfection with the indicated genes. Expression is relative to GFP

transfected Co-ESCs. Data are presented as mean ± SD from n = 3 independent clones and n=2 independent experiments. Statistical significance was tested with two-tailed, paired Student's t-test. (E) Scatterplot showing Zscan4c::eGFP positive cells in bulk Co-ESCs measured by flow cytometry analysis. Representative gates used for bulk analysis are boxed. (F) Flow cytometry analysis of Zscan4c::eGFP positive cells in Co- and Gadd45 TKO-ESCs. (G) Flow cytometry analysis of Zscan4c::eGFP positive cells in Co- or Gadd45-TKO-ESC, 48h after transfection with the indicated genes. Data are presented as means ± SD from n = 3 independent clones and n=2 independent experiments. Statistical significance was tested with two-tailed, unpaired (TKO versus Co) or paired (BFP versus Gadd45 overexpression) Student's t-test. (H) Flow cytometry analysis of Zscan4::eGFP positive cells in Co-ESCs, 48h after transfection with the indicated genes. Statistical significance was tested with two-tailed, paired Student's t-test. (I-J) 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) levels in unsorted, Zscan4 - or Zscan4 + sorted Co- and Gadd45 TKO-ESC clones determined by LC-MS/MS. Values are % of total cytosine (C).

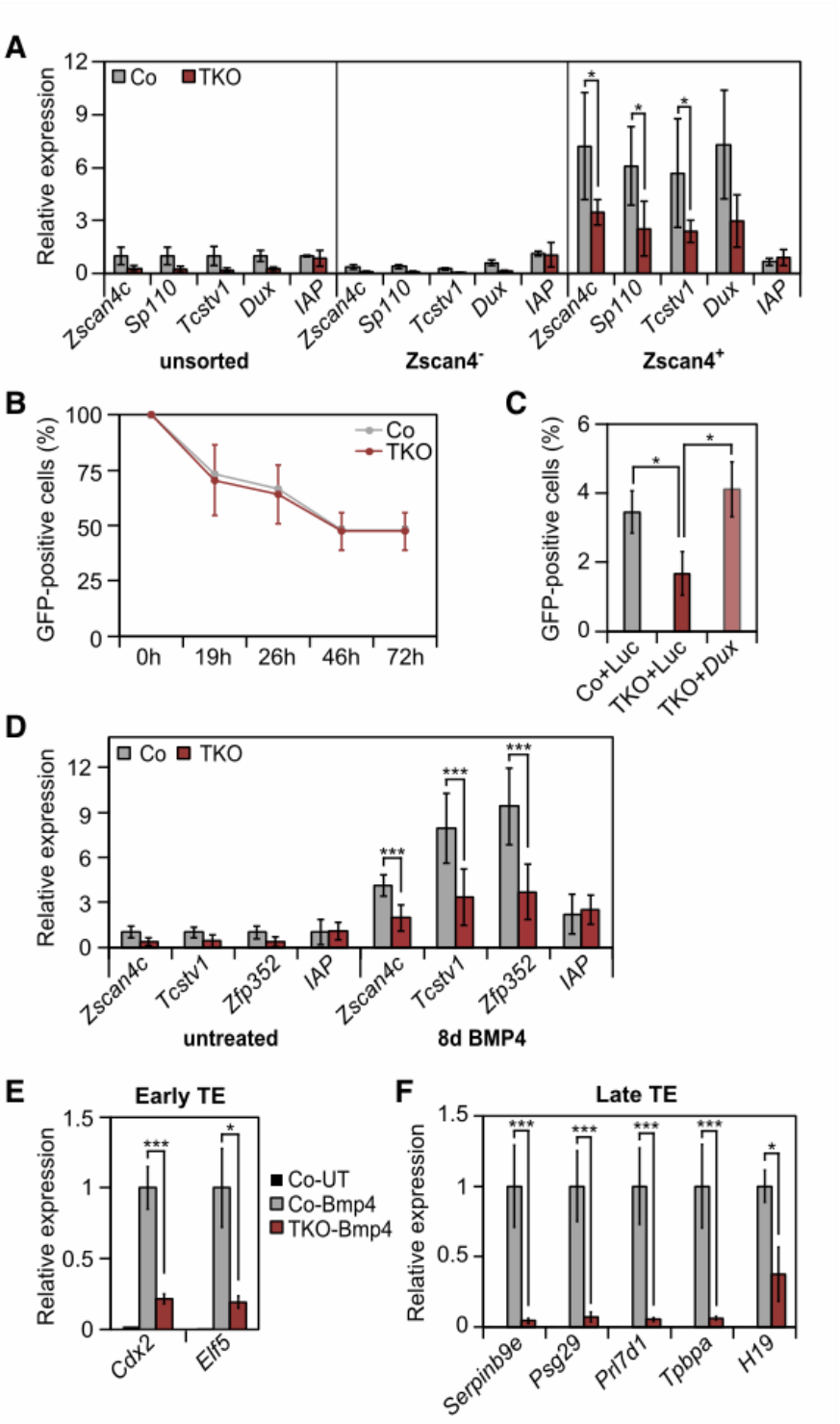


Figure 5: (A) qPCR expression analysis of selected 2C-associated genes in Control- (Co) and Gadd45 TKO-ESC in unsorted, Zscan4 - or Zscan4 + FACS-sorted cells. Expression is relative to the average expression in unsorted Co-ESCs. IAP, intracisternal A-particle (B) Reanalysis of Zcan4 + sorted Co- and TKO-ESCs at the indicated time points using flow cytometry. (C) Dux overexpression restores the reduced number of 2C-like cells in Gadd45 TKO-ESCs. Flow cytometry analysis of Zscan4c::eGFP positive cells in Co- and Gadd45 TKO-ESC, 48h after transfection with the indicated genes

and 24 hours after doxycycline addition. Statistical significance was tested with two-tailed, unpaired (TKO versus control) or paired (Luciferase versus Dux overexpression) Student's t-test. (D) qPCR expression analysis of selected 2C-associated genes in untreated or BMP4 treated (8 days) Co- and Gadd45 TKO-ESC. Expression is relative to untreated Co-ESCs. Data are presented as mean ± SD from n=3 independent clones and n=3 independent experiments. (E-F) Impairment of trophectoderm transdifferentiation in Gadd45 TKO-ESCs. Gadd45 TKO and Co-ESCs were treated for 8 days with BMP4. Induction of early (E) and late (F) trophectoderm (TE) marker genes is relative to BMP4-treated Co-ESCs. UT, untreated

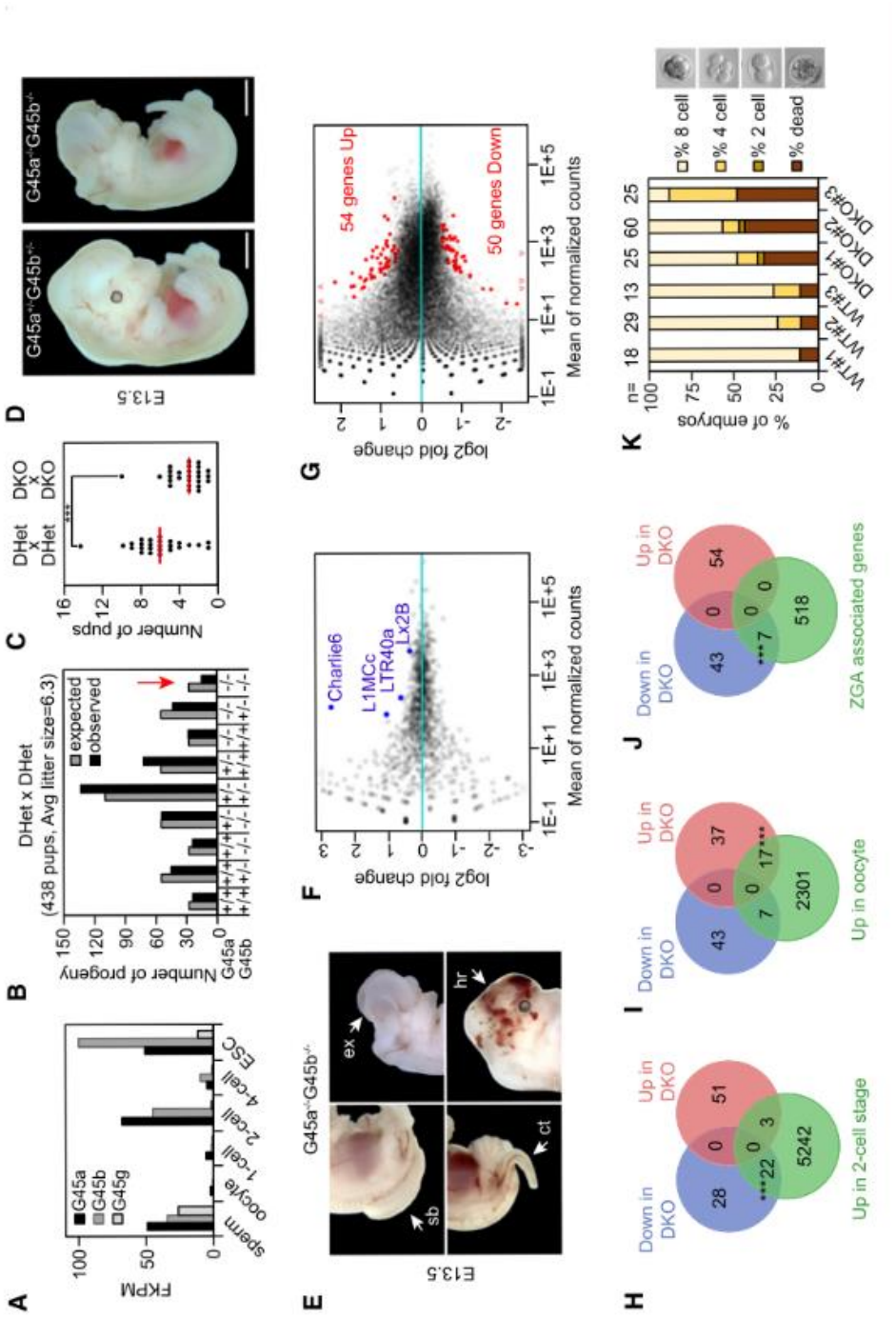


Figure 6. (A) Expression analysis of Gadd45a, -b, -g in preimplantation embryos and ESCs from DBTMEE database (Park et al. 2015). FPKM, Fragments Per

Kilobase of exon model per Million reads mapped. (B) Genotypic analysis of progenies from Gadd45a,b double heterozygous (DHet) mice showing expected and observed Mendelian ratios. (C) (Gadd45a/Gadd45b) -/- (DKO) intercrossing show reduced liter size compared to intercrossed double heterozygous animals. Data points indicate the number of pups per litter. Red lines indicate average litter size. Statistical significance was tested with two-tailed, paired Student's t-test. (D) Images of E13.5 heterozygous and homozygous Gadd45a/Gadd45b mouse embryos. Scale bar, 2 mm. (E) Developmental abnormalities in DKO embryos showing curly tail (ct), exencephaly (ex), cranial hemorrhage (he), and spina bifida (sb). (F-G) RNA-seq differential expression analysis of repeats (F) and genes (G) in 2-cell stage DKO embryos. Significantly deregulated (FDR 10%) repeats and genes are highlighted in blue and red respectively. (H-J) Overlap of genes up- and downregulated in Gadd45a,b DKO 2-cell stage embryos and genes normally upregulated in (H) 2-cell stage embryos, (I) oocytes, or (J) during ZGA (Macfarlan et al. 2012). (K) In vitro development of wild type (WT) and Gadd45a,b DKO preimplantation embryos isolated from three independent breeding. Development was scored 24 hours after isolation. Representative microscopic images of the scored embryonic stages are shown.

Supplemental information

For simplicity I include the supplementary info related to my part of the paper.

Genome-wide methylome analysis

WGBS data sets contained an average of 490 and 479 million paired end reads (151 bp) for two replicates (= two independent clones) of Gadd45 TKO and Co mouse ESCs respectively. Low-quality bases and adapter-containing reads were trimmed from raw data using Trim Galore v0.4.4 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with default parameters. Reads were aligned over the reference mouse genome (NCBI38/mm10) using Bismark v.0.18.0 (Krueger and Andrews 2011) with parameters, '-n 1 -I 0 -X 1000 --score_min L,0,-0.6' resulting in unique mapping rate of 72% on average. Methylation calling for individual cytosines was performed using bismark_methylation_extractor with parameters '-p --ignore 5 --ignore_r2 5 --ample_memory – bedGraph --counts'. Differentially methylated regions (DMRs) between TKO- and Co- ESCs were identified using methylKit 1.3.3 (Akalin et al. 2012) with parameters: (1) Minimum read coverage: 10x per strand, (2) Minimum methylation difference: 30%, (3) Minimum number of consecutive CpGs affected: 2, (4) using a fixed window of 100 bp with (5) false discovery rate <= 5%. Consecutive DMRs were merged. The TKO vs Co scatterplot was generated using methylKit. Gene regulatory regions (i.e. exons, introns, 5UTR, promoter, intergenic, 3UTR) were downloaded from UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=690092087_JliNVUb0ssiztYi4Ees7TTAZA3PF). Fold change enrichment at each regulatory feature was computed as the ratio between the number of hyper-DMRs and the number of random genomic regions. The random set of genomic regions was generated using 'shuffleBed –incl' from BEDTools (Quinlan and Hall 2010). Heatmaps and frequency plots were generated using deepTools v.3.0.1 (Ramírez et al. 2014) with the command 'computeMatrix reference-point -a 5000 -b 5000 --missingDataAsZero –skipZeros'. Methylation differences were calculated by subtracting the average methylation level in 100bp bins between TKO- and Co-ESCs from the center of each feature in +/- 5kb. In this, features within +/- 1 kb of transcription start sites (TSS) were considered as proximal elements and rest as distal elements. We used the following published data sets: 5hmC (Kong et al. 2016), 5hmC, 5fC and 5caC (Shen et al. 2013); TET-dependent DMRs (Lu et al. 2014); Tet1 (Wu et al. 2011); Kdm2aA (Blackledge et al. 2010); Oct4, Sox2, Nanog, p300, Lsd1, RNAPol2, H3K4me1, H4K4me3, H3K79me2, H3K27ac, Hdac2, Rest1, Corest, H3K27me3, Suz12 (Whyte et al. 2012), Nelfa, Spt5 (Rahl et al. 2010); TBP (Kagey et al. 2010); LMR (Stadler et al. 2011). mm9 genome features were converted to mm10 using Batch Coordinate Conversion (liftOver) tool from UCSC Genome Browser Utilities (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Motif analysis was performed using HOMER v3.12 (Heinz et al. 2010) by segmenting the hyper-

DMRs at enhancers, promoters and 5mC oxidized products using parameters '-size 25 –len 8'. Hyper-DMRs were associated with near-by genes using GREAT_v0.3.0 (McLean et al. 2010) with default parameters. All genes that were expressed in at least one sample of ESCs (RNA-seq, this manuscript) were used as background list.

References

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012.

methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles, Genome Biol 13: R87.

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data, Bioinformatics 31: 166–169.

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ, Klose RJ. 2010. CpG islands recruit a histone H3 lysine 36 demethylase, Mol Cell 38: 179–190.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat Biotechnol 26: 1367–1372.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013.

STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29: 15–21. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010.

Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol Cell 38: 576–589.

Ignatiadis N, Klaus B, Zaugg JB, Huber W. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing, Nat Methods 13: 577–580.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture, Nature 467: 430–435.

Kong L, Tan L, Lv R, Shi Z, Xiong L, Wu F, Rabidou K, Smith M, He C, Zhang L, et al. 2016. A primary role of TET proteins in establishment and maintenance of De Novo bivalency at CpG islands, Nucleic Acids Res 44: 8682–8692.

Krueger F, Andrews SR. 2011. Bismark. A flexible aligner and methylation caller for Bisulfite-Seq applications, Bioinformatics 27: 1571–1572.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol 10: R25.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools, Bioinformatics 25: 2078–2079.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics 30: 923–930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 15: 550.

Lu F, Liu Y, Jiang L, Yamaguchi S, Zhang Y. 2014. Role of Tet proteins in enhancer activity and telomere elongation, Genes Dev 28: 2103–2119.

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity, Nature 487: 57–63.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions, Nat Biotechnol 28: 495–501.

Quinlan AR, Hall IM. 2010. BEDTools. A flexible suite of utilities for comparing genomic features, Bioinformatics 26: 841–842.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc regulates transcriptional pause release, Cell 141: 432–445.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data, Nucleic Acids Res 42: W187-91.

4.3 Preamble

This paper is the result of my visit to the Molecular Pathology research unit of the Massachusetts General Hospital at Harvard Medical School in Boston under the supervision of Prof. Luca Pinello. My contribution to the paper is indicated at the end of the manuscript and a repository to reproduce the entire analysis and figures is available at this link: https://github.com/tAndreani/scATAC-benchmarking.

4.4 Chapter 2

Assessment of computational methods for the analysis of single-cell ATAC-seq data (Chen H, Lareau C, Andreani T et al, *Genome Biology* Volume 20, Pages 241, 2019).

Abstract

Recent innovations in single-cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq) enable profiling of the epigenetic landscape of thousands of individual cells. scATAC-seq data analysis presents unique methodological challenges. scATAC-seq experiments sample DNA, which, due to low copy numbers (diploid in humans), lead to inherent data sparsity (1–10% of peaks detected per cell) compared to transcriptomic (scRNA-seq) data (10 – 45% of expressed genes detected per cell). Such challenges in data generation emphasize the need for informative features to assess cell heterogeneity at the chromatin level. For this, We present a benchmarking framework that is applied to 10 computational methods for scATAC-seq on 13 synthetic and real datasets from different assays, profiling cell types from diverse tissues and organisms. Methods for processing and featurizing scATAC-seq data were compared by their ability to discriminate cell types when combined with common unsupervised clustering approaches. We rank evaluated methods and discuss computational challenges associated with scATAC-seq analysis including inherently sparse data, determination of features, peak calling, the effects of sequencing coverage and noise, and clustering performance. Running times and memory requirements are also discussed. As results, This reference summary of scATAC-seq methods offers recommendations for best practices with consideration for both the non-expert user and the methods developer. Despite variation across methods and datasets, SnapATAC, Cusanovich2018, and cisTopic outperform other methods in separating cell populations of different coverages and noise levels in both synthetic and real datasets. Notably, SnapATAC is the only method able to analyze a large dataset (> 80,000 cells).

Introduction

Individual cell types within heterogeneous tissues coordinate to perform complex biological functions, many of which are not fully understood. Recent technological advances in single-cell methodologies have resulted in an increased capacity to study cell-to-cell heterogeneity and the underlying molecular regulatory programs that drive such variation.

To date, most single-cell profiling efforts have been performed via quantification of RNA by sequencing (scRNA-seq). While this provides snapshots of inter-and intra-cellular variability in gene expression, investigation of the epigenomic landscape in single cells holds great promise for uncovering an important component of the regulatory logic of gene expression programs. Enabled by advances in array-based technologies, droplet microfluidics, and combinatorial indexing through split-pooling (Cusanovich DA et al. 2018) (Fig.1a), single-cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq) has recently overcome previous limitations of technology and scale to generate chromatin accessibility data for thousands of single cells in a relatively easy and cost-effective manner.

However, the analysis of scATAC-seq data presents methodological challenges distinct from those of single-cell transcriptomic (scRNA-seq) data. The primary difficulty arises from a difference in the number of RNA vs DNA molecules available for profiling in single cells. While for an expressed gene several RNA molecules are present in a single cell, scATAC-seq assays profile DNA, a molecule which is present in only few copies per cell (two in a diploid organism). The low copy number results in an inherent per-cell data sparsity, where only 1-10% of expected accessible peaks are detected in single cells from scATAC-seq data, compared to 10 –45% of expressed genes detected in single cells from scRNA-seq data (Mereu et al. 2019, Ding J et al. 2019). This emphasizes the need to recover informative features from sparse data to assess variability between cells in scATAC-seq analyses. Further, determination of which features best define cell state is currently unclear.

The difference in readout (gene expression vs chromatin accessibility) has also motivated a variety of approaches to selecting informative features in scATAC-seq methods. While most processing pipelines share common upstream processing steps (i.e., alignment, peak calling, and counting; Fig. 1b), existing computational approaches differ in the way they obtain a feature matrix for downstream analyses. For example, some methods select features based on the sequence content of accessible regions (e.g., k-mer frequencies (Schep et al. 2017, de Boer at al. 2018), or transcription factor (TF) motifs (de Boer CG et al. 2018), whereas other methods select features based on the genomic coordinates of the accessible regions (e.g., extended promoter regions to determine chromatin activity surrounding genes (Schep et al. 2017,

Ji Z. et al. 2017). Finally, the potential feature set in scATAC-seq, which includes genome-wide regions of accessible chromatin (Fig. 1c), is typically 10 –20× the size of the feature set in scRNA-seq experiments (which is defined and limited by the number of genes expressed). This larger feature set could be valuable in distinguishing a wider variety of cell populations and inferring the dynamics underlying cell organization into complex tissues (Corces ML. et al. 2017). However, the novelty and assay-specific challenges associated with these large-scale scATAC-seq datasets and the lack of analysis guidelines have resulted in diverging computational strategies to aggregate data across such an immense feature space with no clear indication as to which strategy or strategies are most advantageous.

Here, we provide the first benchmark assessment of computational methods for the analysis of scATAC-seq data. We discuss the impact of feature matrix construction strategies (e.g., sequence content-based vs genomic coordinates) on common downstream analysis, with a focus on clustering and visualization. This comprehensive survey of current available methods provides user-specific recommendations for best practices that aim to maximize inference capability for current and future scATAC-seq workflows. Importantly, we provide more than 100 well-documented Jupyter Notebooks (https://github.com/pinellolab/scATAC-benchmarking/) to easily reproduce our analyses. We anticipate that this will be a valuable resource for future scATAC-seq benchmark studies.

Results

Benchmark framework

For this benchmarking study, we created an unbiased framework to qualitatively and quantitatively survey the ability of available scATAC-seq methods to featurize chromatin accessibility data. Using this framework we evaluated several datasets of divergent size and profiling technologies. Using widely accepted quantitative metrics, we explored how differences in feature matrix construction influence outcomes in exploratory visualization and clustering, two common downstream analyses. The general overview of our framework is presented in Fig.2. For this study, we collected public data from three published studies (aligned files in BAM format) and generated ten simulated datasets with various coverages and noise levels (see the "Methods" section). To calculate feature matrices for downstream analysis, for each method, we followed the guidelines provided in the documentation in the original study or as suggested by the respective authors. After feature matrix construction, we used three commonly used clustering approaches (K-means, Louvain, and hierarchical clustering) (VY Kiselev et al 2019) and UMAP (L. Mclnnes et al. 2018) projection to find putative subpopulations and visualize cell-to-cell similarities for each method. Next, the quality of the clustering solutions was evaluated by adjusted Rand index (ARI), adjusted mutual information (AMI), and homogeneity (H) when FACS-sorting labels or tissues were available (gold standard) or by a proposed Gini-index-based metric called Residual Average Gini Index (RAGI) when only known marker genes were available (silver standard). Finally, based on these metrics, the methods were ranked by the quality of their clustering solutions across datasets.

Methods overview and featurization of chromatin accessibility data

Several computational methods have been developed to address the inherent sparsity and high dimensionality of single-cell ATAC-seq data, including BROCKMAN (CG de Boer et al. 2018), chromVAR (AN Schep et al. 2017), Cicero (HA Pliner et al. 2018), cisTopic (C Bravo-Gonzalez-Blas et al. 2019), *Cusanovich2018* (Da Cusanovich et al. 2018, Da Cusanovich et al. 2015, Da Cusanovich et al. 2018), Gene Scoring (CA Lareau et al. 2019), scABC (M Zamanighomi et al. 2018), Scasat (SM Baker et al. 2019), SCRAT (Z. Ji et al. 2017), and SnapATAC (R. Fang et al. 2019). Based on the proposed workflow of each method, we computed different feature matrices defined as a features-by-cells matrix (e.g., read counts for each cell (columns) in a given open chromatin peak *feature* (rows)) that could then be readily used for downstream analyses such as clustering. Starting from single-cell BAM files, the feature matrix construction can be roughly summarized into four different common modules: *define regions, count features, transformation,* and *dimensionality reduction* as illustrated in Fig. 2. Not every method uses all steps; therefore, we provide below, a short summary of the strategies adopted by each method

and a *per module* discussion to highlight key similarities and differences (for a more detailed description of each strategy, see the “Methods”section).

Briefly, BROCKMAN (CG de Boer et al. 2018) represents genomic sequences by gapped k-mers (short DNA sequences of length k) within transposon integration sites and infers the variation in k-mer occupancy using principal component analysis (PCA). chromVAR (AN Schep et al. 2017) estimates the dispersion of chromatin accessibility within peaks sharing the same feature, e.g., motifs or k-mers. Cicero (HA Pliner et al. 2018) calculates a gene activity score based on accessibility at a promoter region and the regulatory potential of peaks nearby. cisTopic (C Bravo-Gonzalez-Blas et al. 2019) applies latent Dirichlet allocation (LDA) (a Bayesian topic modeling approach commonly used in natural language processing) to identify cell states from topic-cell distribution and explore cis-regulatory regions from region-topic distribution. Previous approaches that utilize latent semantic indexing (LSI) (termed here as *Cusanovich2018* (Da Cusanovich et al. 2018, Da Cusanovich et al. 2015, Da Cusanovich et al. 2018)) first partition the genome into windows, normalize reads within windows using the term frequency-inverse document frequency transformation (TF-IDF), reduce dimensionality using singular value decomposition (SVD), and perform a first-round of clustering (referred to as “in silico cell sorting”) to generate clades and call peaks within them. Finally, the clusters are refined with a second-round of clustering after TF-IDF and SVD based on read counts in peaks. The Gene Scoring method (CA Lareau et al. 2019) assigns each gene an accessibility score by summarizing peaks near its transcription start site (TSS) and weighting them by an exponential decay function based on their distances to the TSS. scABC (M Zamanighomi et al. 2018) first calculates a global weight for each cell by taking into account the number of distinct reads in the regions flanking peaks (to estimate the expected background). Based on these weights, it then uses weighted k-medoids to cluster cells based on the reads in peaks. Scasat (SM Baker et al. 2019) binarizes peak accessibility and uses multidimensional scaling (MDS) based on the Jaccard distance to reduce dimensionality before clustering. SCRAT (Z. Ji et al. 2017) summarizes read counts on different regulatory features (e.g., transcription factor binding motifs, gene TSS regions). SnapATAC (R. Fang et al. 2019) segments the genome into uniformly sized bins and adjusts for differences in library size between cells using a regression based normalization method; finally, PCA is performed to select the most significant components for clustering analysis.

Define regions

An essential aspect of feature matrix construction is the selection of a set of regions to describe the data (e.g., putative regulatory elements such as peaks and promoters). Most methods described above, including chromVAR,

Cicero, cisTopic, Gene Scoring, scABC, and Scasat, define regions based on peak calling from either a reference bulk ATAC-seq profile or an aggregated single-cell ATAC-seq profile. *Cusanovich2018*, as briefly mentioned above, instead of aggregating single cell to call peaks, first creates pseudo-bulk clades by performing hierarchical clustering on the TF-IDF and SVD transformed matrix using the top frequently accessible windows. Then, peaks are called by aggregating cells within each pseudo-bulk clade. In addition to relying on peaks, some methods have proposed different strategies. BROCKMAN uses the union of regions around transposon integration sites. *Cusanovich2018* (before in silico sorting) and SnapATAC segment the genomes into fixed-size bins (windows) and count features within each bin.

Count features

Once feature regions are defined, raw features within these regions are counted. Note that some methods (e.g., chromVAR) may support the counting of multiple features. For cisTopic, *Cusanovich2018*, scABC, and Scasat, reads overlapping peaks are counted. For *Cusanovich2018* (before the in silico sorting step) and SnapATAC, reads overlapping bins are counted. k-mers are counted under peaks for chromVAR while gapped k-mers are counted for BROCKMAN around transposase cut sites. Similarly, transcription factor motifs (e.g., from the JASPAR database (A. Mathelier et al. 2015)) can be used as features by counting reads overlapping their binding sites in peaks (chromVAR) or genome-wide (SCRAT). If pre-defined genomic annotations such as coding genes are given, Gene Scoring, Cicero, and SCRAT use gene TSSs as anchor points to calculate gene enrichment scores based on reads nearby or just within peaks nearby.

Transformation

After building the initial raw feature matrix using the counting step, different transformation methods can be performed. Binarization of read counts is used by five out of the ten evaluated methods: Cicero, cisTopic, *Cusanovich2018*, Scasat, and SnapATAC (Fig. 2). This step is based on the assumption that each site is present at most twice (for diploid genomes) and that the count matrix is inherently sparse. Binarization is advantageous in alleviating challenges arising from sequencing depth or PCR amplification artifacts. SnapATAC and Scasat convert the binary count matrix into a cell-pairwise Jaccard index similarity matrix. *Cusanovich2018* normalizes the binary count matrix using the TF-IDF transformation. Cicero weights feature sites by their co-accessibility, while Gene Scoring weights sites by a decaying function based on its distance to a gene TSS. Both chromVAR and SnapATAC perform a read coverage bias correction to account for the influence of sample depth. chromVAR creates “background” peaks consisting of an equal number of peaks matched for both average accessibility and GC content to calculate bias-corrected deviation while SnapATAC uses a regression-based method to mitigate the coverage difference between cells. scABC also implements a similar step by calculating a weight for each cell, but these weights are not

used to transform the matrix; instead they are used later in the clustering procedure. Both BROCKMAN and chromVAR compute z-scores to measure the gain or loss of chromatin accessibility across cells. SCRAT adjusts for both library size and region length.

Dimensionality reduction

In the final step before downstream analysis, several methods apply different dimensionality reduction techniques to project the cells into a space of fewer dimensions. This step can refine the feature space mitigating redundant features and potential artifacts and potentially reducing the computation time of downstream analysis (Fig. 2). PCA is the most commonly used method (used by BROCKMAN, SnapATAC, and Cusanovich2018). cis-Topic uses latent Dirichlet allocation (LDA) to generate two distributions including topic-cell distribution and region-topic distribution. Choosing the top topics based on the topic-cell distribution reduces the dimensionality. Scasat uses multidimensional scaling (MDS). When reviewing the different methods to include in our benchmark, we noticed that not all methods perform a dimensionality reduction step, which could skew the relative performance across methods. Therefore, for chromVAR, Cicero (gene activity score), Gene Scoring, scABC, and SCRAT, we considered, in addition to the original feature matrix, also a new feature matrix after PCA transformation, since this is a simple and commonly used technique for dimensionality reduction.

To better evaluate the effects of different modules including *define regions, count features, transformation, and dimensionality reduction,* we also considered a simple control method, referred to as Control-Naïve, by combining the most common and simple steps for building a feature matrix, i.e., counting reads within peaks to obtain a peaks-by-cells raw count matrix and then performing PCA on it (the number of top principal components was determined based on the elbow plot for all the methods). Since the feature matrix of scABC is also a peaks-by-cells raw count matrix, this matrix after PCA will correspond to the one obtained by the Control-Naïve method (to avoid redundancies, in our assessment, we refer to this matrix as Control-Naïve).

We also noticed that some methods might slightly diverge from the proposed four modules common framework. For example, Cicero calculates gene activity scores by first performing two transformations (binarize and weight features) and then performing the counting step around the annotated TSS. We believe the proposed modularization of the feature matrix construction can still serve as a useful framework to represent the core components of the different methods and provides an intuitive and informative summary of the diverse scATAC-seq methodologies.

Once dimensionality reduction is completed, the transformed feature matrix can be used for unbiased clustering, visualization, or other downstream analyses. Here, we have used the final feature matrices generated by each

scATAC- seq analysis method and evaluated their performance in uncovering different populations by unsupervised clustering.

Clustering approaches and metrics used for performance evaluation

This study employed three diverse types of commonly used unsupervised clustering methods for single-cell analysis (VY Kiselev et al. 2019): K-means clustering, hierarchical clustering, and the Louvain community detection algorithm (see the “Methods” section). Clustering results were evaluated by three commonly used metrics: adjusted Rand index (ARI), adjusted mutual information (AMI), and homogeneity, when a gold standard solution was available (known labels for the simulation data and FACS-sorted cell populations or known tissues for the real datasets). We propose a Gini-index-based metric called Residual Average Gini Index (RAGI), which was used to evaluate the clustering results when no ground truth was available and only a few marker genes were known by which populations could be discriminated (see the “Methods” section). For each metric, we defined the clustering score as the highest score among the three clustering methods, i.e., the score which corresponded to the clustering solution that maximized the metric. This framework allowed for benchmarking the ability of each strategy to featurize chromatin accessibility data and its impact on important downstream analyses such as clustering and visualization. The following sections present the results of this evaluation for all the above described synthetic and real scATAC-seq datasets.

Clustering performance on simulated datasets

We simulated 10 scATAC-seq datasets using available bulk ATAC-seq datasets with clear annotations from the bone marrow and erythropoiesis (MR Corces et al 2016, J.C. Ulirich et al. 2019, Ludwig S. Leif et al. 2019) using varying noise levels and read coverages. Briefly, to generate the peaks-by-cells matrices, we defined a noise parameter (between 0 and 1) as the proportion of reads occurring in a random peak from one of the sorted populations. The remaining proportion of reads was distributed as a function of the bulk sample (see the “Methods”section). A feature matrix with a noise level of 0 preserved perfectly the underlying cell type specificity of the reads within peaks. Conversely, a feature matrix with a noise level of 1 contained no information to discriminate cell types based on the reads within peaks. In our study, we considered three noise levels: no noise (0), moderate noise (0.2), and high noise (0.4). To better and more fairly evaluate the contribution of the core steps of each method (i.e., *count features, transformation, and dimensionality reduction*) regardless of the pre-processing steps usually excluded from these methods (reads filtering, alignment, peak calling, etc.),

we compared the performance of each method using a set of pre-defined peak regions from bulk ATAC-seq datasets.

We selected the top 80,000 peaks based on the number of cells in which peaks were observed (each peak that was present in at least one cell) for all methods and all synthetic datasets. Using the bulk ATAC-seq bone marrow dataset, we simulated five additional datasets to explore the effect of coverage on clustering performance (5000 fragments, 2500 fragments, 1000 fragments, 500 fragments, 250 fragments respectively per cell). Each method was used to analyze all synthetic datasets as suggested in the method documentation (see Additional file 1: Note S1 and Additional file 1: Figure S2).

Simulated bone marrow datasets

We generated chromatin accessibility profiles (2500 fragments per cell) based on six different FACS-sorted bulk cell populations: hematopoietic stem cells (HSCs), common myeloid progenitor cells (CMPs), erythroid cells (Ery), and other three lymphoid cell types: natural killer cells (NK), CD4, and CD8 T cells (see Fig. 3a). We used ARI, AMI, and homogeneity metrics to compare the clustering solutions with the known cell type labels (Fig. 3b, Additional file 1: Figure S3, Additional file 1: Table S1). The top three methods based on these simulation settings were cisTopic, Cusanovich2018 , and SnapATAC. They performed equally well with no noise and moderate noise (with clustering scores close to 1.0) (Additional file 1: Figure S3, Additional file 1: Table S2). At a noise level of 0.4, the methods showed more separation in performance accordingly to the three metrics (Fig. 3b, Additional file 1: Table S3). SnapATAC, Cusanovich2018, and cisTopic clearly outperformed the Control-Naïve method with consistently higher clustering scores across all metrics. Scasat performed slightly better than the Control-Naïve method, and the remaining methods underperformed relative to the Control-Naïve method. For scABC (i.e., peaks-by-cells raw count matrix), hierarchical clustering performs much better than the other two clustering methods. chromVAR performance using k-mers as features was superior to the approach using motifs. Another k-mer-based method, BROCKMAN, demonstrated similar performance to the k-mer-based chromVAR method. Motif-based SCRAT performed better than motif-based chromVAR. Both Cicero gene activity scores and Gene Scoring (which summarize the chromatin accessibility around coding annotations without a dimensionality reduction step) generally performed poorly. PCA boosted the performance of scABC, Cicero, and Gene Scoring. This step improved clustering performance regardless of the clustering method (also we noted again that scABC after PCA is equivalent to the Control-Naïve method), especially for the Louvain approach. PCA also slightly boosted the performance of the k-mer-based chromVAR but did not markedly improve the results of the motif-based chromVAR or SCRAT analyses. We next investigated qualitatively the obtained clustering solutions, using the respective feature matrices to project the cells onto a 2-D space using UMAP and colored them based on the obtained clustering solutions (Additional file 1: Figure S4) or based on the true population labels used to generate the data

(Fig. 3d). The top two clustering solutions based on the ARI (SnapATAC with k-means and SnapATAC with Louvain) are shown for ease of comparison (Fig. 3c). Cusanovich2018 and SnapATAC are the only two methods that clearly separated all six populations. cisTopic slightly mixed CD4 and CD8 T cells. Scasat and the Control-Naïve method failed to separate CD4 and CD8 T cell populations. BROCKMAN slightly mixed NK with CD4 and CD8 T cells and could not further separate CD4 and CD8 T cells. It also failed to clearly separate HSC and CMP. Both k-mer-based and motif-based chromVAR as well as SCRAT could only separate the Ery population while failing to separate HSC and CMP as well as CD4, CD8 T cells, and NK. The chromVAR k-mer-based method mixed HSC and CMP to a lesser extent compared to the motif-based method. There was no clear separation of cells using scABC (the peaks-by-cells raw count matrix), Cicero, or Gene Scoring. We observed that PCA clearly improved the separation of cell populations for Cicero and Gene Scoring. It also slightly improved the separation of CD4, CD8 T cells, and NK populations by k-mer-based chromVAR. No clear improvement was observed for the motif-based chromVAR or SCRAT methods. We further observed that a lack of visual separation of cell types in the UMAP plots (scABC, Cicero, and Gene Scoring), corresponded with substantial variation between the performances of the three clustering methods, showing better performance in the k-means clustering (Fig. 3b, d). All methods except for Cusanovich2018 and SnapATAC demonstrated declining performance with increased noise level (Additional file 1: Figures S3, S5a). Cusanovich2018 and SnapATAC were more robust to noise, showing no noticeable changes at increasing noise levels, while cisTopic was slightly more sensitive to noise; its performance dropped markedly when the noise level was increased to 0.4. Next, the effect of the coverage on clustering performance was investigated. We progressively decreased the number of fragments per cell from a high coverage of 5000 fragments, to a medium coverage of 2500 fragments and 1000 fragments, then to a low coverage of 500 fragments, and finally to 250 fragments. The performance of all methods declined as coverage was decreased (Additional file 1: Figure S5b, Additional file 1: Figure S6, Additional file 1: Tables S4-S8). Cusanovich2018, SnapATAC, Scasat, and Control-Naïve are relatively robust to low coverage and outperform other methods. cisTopic worked well with high coverage but, in contrast to the above-listed methods, was more sensitive to lower coverages (Additional file 1: Figure S6e).

Simulated erythropoiesis datasets

Following the simulation of discrete sorted cell populations, we simulated three scATAC-seq datasets aimed at mimicking the continuous developmental erythropoiesis process and encompassing the following 12 populations: hematopoietic stem cells (HSCs), common myeloid progenitors (CMPs), megakaryocyte-erythroid progenitors (MEPs), multipotent progenitors (MPPs), myeloid progenitors (MyP), colony-forming unit-erythroid (CFU-E), proerythroblasts (ProE1), proerythroblasts (ProE2), basophilic erythroblasts (BasoE), polychromatic erytrhoblasts (PolyE), orthochromatic

erythroblasts (OrthoE), and OrthoE and reticulocytes (Orth/Ret). These datasets were generated as before with three noise levels (0, 0.2, and 0.4) and with 2500 fragments per cell.

To first quantitatively evaluate the clustering solutions, we used ARI, AMI, and the homogeneity metrics (Additional file 1: Figure S7 and Additional file 1: Table S9). Without noise, SnapATAC, cisTopic BROCKMAN, Cusanovich2018, and Scasat consistently outperform the Control-Naïve across the three metrics (Additional file 1: Figure S7a). chromVAR, as before, performs better using k-mers as features than when using motifs. SCRAT and scABC work as well as k-mer-based chromVAR. Again, methods such as Cicero and Gene Scoring that only summarize chromatin accessibility around TSS perform poorly. For scABC, Cicero, and Gene Scoring, we also notice that there are significant discrepancies between the three clustering methods, but their performances become similar after PCA (scABC after PCA is equivalent to the Control-Naïve method). Again, we observe that PCA can significantly improve the clustering performance of Louvain for scABC, Cicero, and Gene Scoring but not for chromVAR and SCRAT.

As before, to qualitatively assess population separation, we inspected UMAP projections applied to the noise-free simulated dataset (Additional file 1: Figure S7a). In accordance with the quantitative comparison, cisTopic, Cusanovich2018, SnapATAC, and BROCKMAN demonstrate better performance in separating cell types compared to the Control-Naïve method and are able to further separate BasoE and PolyE. Moreover, SnapATAC can clearly distinguish CFU-E, ProE1, and ProE2 while cis-Topic, Cusanovich2018 , and BROCKMAN are only able to separate ProE2 out of these three populations. Scasat performs similarly to the Control-Naïve method. chrom-VAR with k-mers as features and SCRAT are able to isolate six major groups including HSCs-MPPs, CMP, MEP, MyP, CFU-E-ProE1-ProE2, and BasoE-PolyE-OrthoE-Orth/Ret. chromVAR with k-mers performs well in preserving the order of CFU-E-ProE1-ProE2 and BasoE-PolyE-OrthoE-Orth/Ret. SCRAT can further separate BasoE-PolyE from OrthoE-Orth/Ret while mixing up CFU-E-ProE1-ProE2. As before, we noticed that chrom-VAR using k-mers as features obtained a better separation of cell types than when using motifs. scABC is able to preserve well the order of major groups in a continuous way but fails to separate CFU-E-ProE1-ProE2 and OrthoE-Orth/Ret. Cicero gene activity score and Gene Scoring mixed different cell types, but after a simple PCA step, they clearly separate cells into three major groups. scABC did not perform well and produced small noisy clusters with different cell types mixed together.

As expected, we observed that increasing the level of noise resulted in clustering performance decrease and a decline of visual separation of cell types for all the methods (Additional file 1: Figure S5c, Additional file 1: Figure S7, Additional file 1: Table S10-S11). SnapATAC, cisTopic, and Cusanovich2018 performed reasonably well when increasing the noise level, with SnapATAC the most robust among the three.

Clustering performance on real datasets

Following the benchmark of the synthetic datasets, we assessed the performance of the methods on real datasets. These datasets were generated using different technologies: the Fluidigm C1 array (JD Buenrostro et al. 2018), the 10X Genomics droplet-based scATAC platform, and a recently optimized split-pool protocol (DA Cusanovich et al. 2018). Each real dataset used was fundamentally different in its cellular makeup as well as size and subpopulation organization. Notably, as “true positive” labels are not always available, in addition to the metrics used on the simulated datasets, here we introduced the RAGI,a simple metric based on the Gini index that can be adopted when marker genes for the expected populations are known (see the “Methods” section). In our assessment of Cusanovich2018, to make a fair comparison, we use first the same set of peaks used for other methods instead of the peaks called from its pseudo-bulk-based procedure.

However, since this strategy may be important for the final clustering performance, the pseudo-bulk-based peak calling strategy is tested and discussed in a subsequent section.

Buenrostro2018 dataset

The first and smallest dataset we used in our benchmarking contains single-cell ATAC-seq data from the human hematopoietic system (hereafter *Buenrostro2018,* (JD Buenrostro et al. 2018)). This dataset consists of 2034 hematopoietic cells that were profiled and FACS-sorted from 10 cell populations including hematopoietic stem cells (HSCs), multipotent progenitors (MPPs), lymphoid-primed multipotent progenitors (LMPPs), common myeloid progenitors (CMPs), and granulocyte-macrophage progenitors (GMPs), GMP-like cells, megakaryocyte-erythroid progenitors (MEPs), common lymphoid progenitors (CLPs), monocytes (mono) and plasmacytoid dendritic cells (pDCs). Figure 4a illustrates the roadmap of hematopoietic differentiation. For this dataset, the FACS-sorting labels are used as the gold standard. The analysis details for each method are documented in Additional file 1: Note S2. We started by evaluating the clustering solutions based on the feature matrices generated by the different methods. We used the same metrics used for the synthetic datasets: ARI, AMI, and homogeneity (Fig. 4b, Additional file 1: Table S12). cisTopic, *Cusanovich2018*, chromVAR, SnapATAC, and Scasat outperform the other methods across all three metrics. We also observed that chromVAR with k-mers or TF motifs and with or without PCA performs consistently well. As before, k-mer-based features work better than motif-based features. This can be also observed when comparing BROCKMAN, another k-mer-based method, with SCRAT, which is a motif-based method. TSS-based methods

including Cicero and Gene Scoring did not perform well. Cicero requires a pre-processing step to assess cell similarity; poor performance might be due to the internally incorrectly inferred coordinates (our assessment used the t-SNE procedure as suggested in their documentation). Implementing PCA consistently improves the performance of scABC (as mentioned before, scABC after PCA is equivalent to the Control-Naïve method) and Cicero but does not impact the performance of chromVAR, SCRAT, and Gene Scoring. We also observed that for this dataset, the Louvain algorithm works consistently well across different metrics and methods and performs better than hierarchical clustering and k-means in almost all the cases. We also qualitatively assessed the separation of different cell types by visualizing cells in UMAP projections based on the FACS-sorted labels (Fig. 4d) and clustering solutions (Additional file 1: Figure S8). Figure 4c shows the best two combinations based on ARI: cisTopic with Louvain and *Cusanovich2018* with Louvain (the complete ranking is presented in Additional file 1: Table S12). As Fig. 4d shows, in accordance with the clustering analyses, cisTopic, Cusanovich2018, Scasat, SnapATAC, and chromVAR can generally separate cell types and reasonably capture the expected hematopoietic hierarchy. cisTopic and SnapATAC show a clear and compact separation among groups, with SnapATAC recovering finer structure within each cell type cluster. chromVAR with k-mers or motifs corresponds to a more continuous progression of the different cell types. Control-Naïve and BROCKMAN perform comparably in distinguishing cell types and preserving the continuous hematopoietic differentiation. Cicero gene activity scores, SCRAT, and scABC show ambiguous patterns of distinct cell populations while Gene Scoring fails to separate different cell types. For Cicero gene activity score, after performing PCA, the separation of different cells is noticeably improved. For SCRAT, performing PCA does not show clear improvement.

10 X Peripheral blood mononuclear cells (10X PBMCs) dataset

Next, we investigated a recent dataset produced by 10X Genomics profiling peripheral blood mononuclear cells (PBMCs) from a single healthy donor. In this dataset, 5335 single nuclei were profiled (~ 42 k read pairs per cell); no cell annotations are provided. Based on recent studies (C. Bravo Gonzalez-Lopez et al. 2019, HA Pliner et al. 2019), we expected ~ 8 populations: CD34+, natural killer and dendritic cells, monocytes, lymphocyteB and lymphocyte T cells, together with terminally differentiated CD4 and CD8 cells. Therefore, we used 8 as the number of expected populations for the clustering procedures. The analysis details for each method are documented in Additional file 1: Note S3.

Several marker genes have been proposed to label the different populations or to annotate clustering solutions for PMBCs (C. Bravo Gonzalez-Blas et al. 2019, HA Pliner et al. 2019). To measure cluster relevance based on these marker genes, we can annotate the clusters (or alternatively any group of cells) according to the accessibility values at those marker genes. In addition, accessibility at marker genes should be more variable between clusters than

accessibility at housekeeping genes (since they should be, by definition, more equally expressed across different populations). Based on these ideas, we proposed and calculated the Residual Average Gini Index (RAGI) score (see the “Methods” section) contrasting marker and housekeeping genes (Fig. 5a, Additional file 1: Table S13). For reasonable clustering solutions, we expect that the accessibility of marker genes defines clear populations corresponding to one or few clusters, whereas the accessibility of the housekeeping genes is broadly distributed across all the clusters.

As expected, methods with the highest performance, such as SnapATAC and chromVAR, showed a higher average accessibility for just one cluster for the same marker gene, while lower performing methods such as SCRAT or Gene Scoring showed higher average accessibility in multiple clusters for the same marker gene, further motivating the use of the RAGI metric (Additional file 1: Figure S9). Figure 5b shows for the top two performing methods based on RAGI (SnapATAC and chromVAR with k-mers) the gene accessibility patterns for 3 genes (S100A12 monocyte-specific, MS4A1 B cell-specific, and GAPDH housekeeping.)

The same three genes are also shown in UMAP plots of the other methods (Additional file 1: Figure S10). Again, we observed that Louvain algorithm performed better than k-means and hierarchical clustering for almost all scATAC-seq methods. Importantly, negative RAGI score for a method (see for example the solutions obtained by the Gene Scoring in Fig. 5a, Additional file 1: Figure S10) may suggest that its clustering solutions are defined by housekeeping genes rather than informative marker genes. We also qualitatively evaluated the clustering solutions of the different methods using UMAP projections (Fig. 5c, Additional file 1: Figure S11). We observed two major groups for all methods except for scABC. Among these methods, the UMAP projections based on feature matrices obtained by Control-Naïve, cisTopic, Cusanovich2018, Scasat SnapATAC, BROCKMAN, and chromVAR showed additional smaller groups and finer structures. For Cicero gene activity scores, performing PCA helps to improve the separation of more putative cell types. Instead, for SCRAT and Gene Scoring, the PCA step did not improve the separation. Given that the ranking of methods in datasets with ground truth is similar to the ranking based on the RAGI metric, we believe this simple approach is a reasonable surrogate metric that can be useful for evaluating unannotated datasets, a common scenario in single-cell omics studies.

sci-ATAC-seq mouse dataset

The last dataset analyzed in our benchmark consists of sci-ATAC-seq data from 13 adult mouse tissues (bone marrow, cerebellum, heart, kidney, large intestine, liver, lung, pre-frontal cortex, small intestine, spleen, testes, thymus, and whole brain), of which 4 were analyzed in duplicate for a total of 17 samples and 81,173 single cells (DA Cusanovich et al. 2018). Each tissue can

be interpreted as a coarse ground truth, used later to evaluate clustering solutions (Fig. 6a).

The analysis details for each method are documented in Additional file 1: Note S4. Despite using a machine with 1 TB of RAM memory, almost all the methods failed to even load this dataset, owing to its size. The only method capable of processing this dataset in a reasonable time was SnapATAC (~ 700 min). The other methods failed to run due to memory requirements. To understand the causes of this failure, we did an in-depth analysis of their scalability looking at their source code (Additional file 1: Note S5). Briefly, we found that the majority of the methods try to load the entire dataset in the central memory while SnapATAC uses a custom file format (.snap) based on HDF5 (https://support.hdfgroup.org/HDF5/whatishdf5.html), allowing out of core computation by efficiently and progressively loading in the central memory only the data chunks required at any given moment of the analysis.

On this dataset, SnapATAC was able to correctly cluster cells of the following tissues: kidney, lung, heart, cerebellum, whole brain, and thymus. However, for the other tissues, including the bone marrow and small intestine, cells are distributed in groups of mixed cell types (Additional file 1: Figure S12), as reflected by the score of the three metrics used for the other datasets evaluation (Additional file 1: Table S14), i.e., ARI = (HC = 0.24, k-means = 0.34, Louvain = 0.39), AMI = (HC = 0.55, k-means = 0.55, Louvain = 0.62), and homogeneity = (HC = 0.52, k-means = 0.54, Louvain = 0.60). To gain insight on the performance of the other methods on this dataset, we randomly selected 15% of cells from each sample to construct a smaller sci-ATAC-seq dataset consisting of 12,178 cells. As Fig. 6b shows Cusanovich2018, k-mer-based chromVAR, cisTopic, SnapATAC, Scasat, and Control-Naïve perform comparably well and have noticeably better clustering scores than the other methods (Additional file 1: Table S15). Consistent with what we observed previously, peak- or bin-level methods generally work better. In this dataset, k-mer-based chromVAR and its combination with PCA transformation performs equally well as peak- or bin-level methods and better than the motif-based methods. Simply counting reads within peaks (scABC) and gene-level-featurization-based methods (Gene Scoring and Cicero) perform poorly overall. Adding a PCA step improves noticeably scABC (scABC after PCA is the same as Control-Naïve) and Gene Scoring. It also slightly improves Cicero but it does not affect chromVAR and SCRAT.

As before, all the clustering solutions of the different methods were visualized in UMAP plots (Additional file 1: Figure S13). The top two combinations, i.e., Cusanovich2018 and chromVAR k-mers with PCA, are visualized in Fig. 6c. To visually compare the separation of the different tissues across methods, we also inspected UMAP plots where cells are colored based on the tissue of origin. Similar to what we observed using the clustering analysis, cisTopic, Cusanovich2018, and SnapATAC are able to separate cells into the major tissues and also to capture finer discrete groups. The Control-Naïve method and Scasat are also able to distinguish the major tissues but show some

mixing within each discrete cell population. k-mer-based chromVAR can separate out liver, kidney, and heart tissues and present the other tissues within a continuous bulk population while preserving the structure of the distinct tissues. We observed that after running PCA, k-mer-based chromVAR can recover an additional group of cells within the lung tissue and also detect finer structure within the cells from the brain. Compared with k-mer-based features, motif-based chromVAR and its combination with PCA transformation distinguished fewer tissue groups while mixing more cells from different tissues. BROCKMAN recovered a continuous structure with the different tissues but does not distinguish them clearly. Similarly, Gene Scoring put cells from different tissues into a big bulk population with limited separation. PCA improved its ability to separate out a few tissues, including the liver, heart, and kidney. SCRAT and Cicero gene activity scores mixed most of the cells from different tissues and performed poorly on this dataset with or without PCA.

Clustering performance summary

To assess and compare the overall performance of scATAC-seq analysis methods, we ranked the methods based on each metric (ARI, AMI, homogeneity, RAGI) by taking the best clustering solution for the three real datasets (Buenrostro2018 dataset, 10X PBMCs dataset, and the downsampled sci-ATAC-seq mouse dataset) and two synthetic datasets (simulated bone marrow dataset and simulated erythropoiesis dataset with the moderate noise level of 0.2 and a medium coverage of 2500 fragments per cell). Then, for each dataset except for the 10X PBMC dataset, we calculated the average rank across ARI, AMI, and homogeneity. For the 10X PBMC dataset, RAGI is calculated instead (Additional file 1:Figure S14a). Lastly, we calculated the average rank across different datasets. According to the average ranking, SnapATAC, cisTopic, and Cusanovich2018 are the top three methods to create feature matrices that can be used to cluster single cells into biologically relevant sub-populations (Fig. 7a). SnapATAC consistently performed well across all datasets. Both cisTopic and Cusanovich2018 demonstrated satisfactory performance across all datasets except for the 10X PBMCs dataset. Generally, methods that implement a dimensionality reduction step work better (SnapATAC, cisTopic, Cusanovich2018, Scasat, Control-Naïve, and BROCKMAN) than those without it (SCRAT, scABC, Cicero, and GeneScoring). We also observed that chromVAR performs better in real datasets than in simulated datasets and that the k-mer-based version of chromVAR consistently outperforms motif-based chromVAR. For the methods that do not implement dimensionality reduction, the PCA step does not always improve the performance except for scABC and Cicero, in which the PCA transformation consistently boosts the results. Interestingly, we observed that regardless of the method, the PCA consistently improves the clustering solutions obtained by the Louvain algorithm.

Keeping the first PC vs removing the first PC

We noticed that in some cases, the first principal component (PC) may only capture variation in sequencing depth instead of biologically meaningful variability. To make a thorough assessment of how the first PC affects the clustering results, we compared the effect of keeping vs removing the first PC on the three real datasets (for this comparison, we consider both the methods that implemented PCA and the combination of PCA and the methods that did not implement a dimensionality reduction step) (Additional file 1: Figure S15). Across all three datasets, we observe that for Control-Naïve, BROCKMAN, SCRAT-PCA, and Gene Scoring-PCA, removing the first PC consistently helped in better separating the different populations in UMAP projections and improved clustering performance. In contrast, the performance of chromVAR-PCA with motifs as features consistently dropped after removing the first PC. Cusanovich2018 and SnapATAC performed similarly before and after removing the first PC across all datasets. For Cicero-PCA, removing the first PC did not clearly affect its performance in Buenrostro2018 and 10X PBMCs datasets but improved its performance in the downsampled sci-ATAC mouse dataset.

Generally, the methods that implement binarization (e.g., Cusanovich2018, SnapATAC) or that implement cell coverage bias correction (e.g., chromVAR, SnapATAC) tend to be less affected by the sample sequencing depths. Therefore, for these methods, we believe that the first PC does not capture the library size and removing it does not help to improve the clustering results. On the contrary, for methods that do not implement any specific step to correct for potential artifacts associated with sequencing depth, the first PC is more likely to capture biologically irrelevant factors and therefore may reduce biology-driven differences. However, this operation must be applied with caution, since removing the first component could also in some cases remove some biological variation (e.g., motif-based chromVAR).

Clustering performance when running methods as end-to-end pipelines

When designing this study, we reasoned that a benchmark procedure could be approached from two very different perspectives. The first is the end user perspective, i.e., a user that runs a method as a black box following the provided documentation with the goal to obtain a reasonable clustering solution without worrying too much about the internal design choices and procedures. In these settings, it is not trivial to systematically compare the methods and understand which part related to the featurization may influence the final clustering performance, especially if also the clustering algorithms used are different. The second perspective that was used instead in the rest of this benchmarking effort is the developer perspective, i.e., we tried to understand what are the key steps of each method that can boost clustering performance of common clustering approaches. Regardless, we reasoned that it is important to provide some insights on the user perspective, since some readers will use the tested methods as end-to-end pipelines. Therefore, we also compared the clustering solutions produced by running the complete analysis pipelines as outlined in tutorials for the methods that explicitly

implement a clustering step (see Additional file 1: Note S6). We evaluated the clustering results using ARI, AMI, and homogeneity for the Buenrostro2018 and sci-ATAC-seq mouse datasets and RAGI for the 10X PBMCs data-set (Additional file 1: Tables S16-S18). We observe the top three methods, i.e., Cusanovich2018, cisTopic, and SnapATAC, still outperform the other methods but with a slightly different ranking (Cusanovich2018 is ranked first followed by cisTopic and SnapATAC, Fig. 7b, Additional file 1: Figure S14b). Also, both scABC and Cicero performed better than Scasat in this analysis.

Interestingly, we observed that SnapATAC, cisTopic, Cusanovich2018, and Scasat have even better clustering solutions in our benchmarking framework compared to using their own clustering approach. On the other hand, scABC and Cicero had better clustering results when running their own clustering procedure. scABC uses an unsupervised clustering method tailored to single-cell epigenomic data (including scATAC-seq). Although it uses the naïve peaks-by-cells raw count as its feature matrix, it calculates cells' weights by considering their sequencing coverage and giving more weight to cells with higher number of reads. Also, it performs two steps of clustering by using weighted k-medoid algorithm based on Spearman rank correlation to find landmarks first and then assigns cells to the landmarks. These specific steps help improve its clustering performance. For the Cicero clustering workflow, we used the gene activity scores and, as proposed in their tutorial, functions from Monocle2, to (i) normalize the scores and (ii) reduce the dimensionality with t-SNE by using the top PCs before clustering cells. These extra steps helped in improving its clustering solutions. This suggests that appropriate normalization steps need to be properly performed to improve clustering analysis, in addition to simple transformations like binarizing counts and/or performing a PCA.

Taken together, based on these analyses, we recommend using SnapATAC, cisTopic, or Cusanovich2018 to cluster cells in meaningful subpopulations. This step can be followed by methods such as Cicero, Gene Scores, or with TF motifs (e.g., chromVar) to annotate clusters and to determine cell types in an integrative approach.

Important considerations in defining informative regions for scATAC-seq analyses

Feature sets of informative peaks for scATAC analyses may be computed from bulk samples available through large-scale consortia such as ENCODE (EP Consortium et al. 2012) and ROADMAP (BE Bernstein et al. 2010) or more precise tissue-specific cell types as in the murine ImmGen Project (H Yoshida et al. 2019). However, scATAC-seq analyses often require de novo

inference of dataset-specific accessibility peaks in order to resolve cell types and regulatory activity.

To date, there are three major methods for generating peak sets for scATAC experiments. The first strategy (pseudo-bulk from all single cells, PB-All) for inferring peaks is to call peaks on a pseudo-bulk sample composed of all the reads from all cells in the library. The second (pseudo-bulk from FACS, PB-FACS) is to call peaks in a priori-defined cell types isolated by FACS sorting. A consensus peak set can be defined by combining summits of individual peaks using an iterative algorithm (MR Corces et al 2016, JD Buenrostro et al 2018, R Stark et al. 2011). Finally, a third strategy (pseudo-bulk from clades, PB-Clades) uses a pre-clustering of cells to define initial populations (DA Cusanovich et al. 2018, DA Cusanovich et al. 2018). Subsequent peak calling is performed in each initial cluster. Aggregate peak sets can then be defined from synthesizing the summits of each cluster-specific peak set as described above.

Bulk ATAC-seq peaks vs aggregated scATAC-seq peaks

To evaluate the effect of using peaks obtained from bulk ATAC-seq data vs peaks obtained from aggregated single-cell profiles, we reanalyzed the Buenrostro2018 dataset in which both are available (Additional file 1: Figures S16-S17). Here, we considered only the methods that use peaks as input (i.e., SnapATAC, SCRAT, and BROCKMAN are excluded). For the aggregated scATAC-seq peaks, we merged cells of the same cell type based on the FACS sorting labels and performed peak calling within each cell type. Then, peaks defined within each cell type were merged. For most methods, we did not observe clear differences in performance between the two input peak strategies. For cisTopic, Cusanovich2018, and Cicero, aggregated scATAC-seq peaks overall perform better across all three metrics (Additional file 1: Figure S18a, Additional file 1: Table S19). We also tested the strategy of defining pseudo-bulk samples from clades when no sorting labels are provided.

Cusanovich2018 is the only method that provides a workflow to identify initial clades and call peaks within each clade. It counts reads within the fixed-size windows and pre-clusters cells using hierarchical clustering to define initial clades from which peaks are called. We applied this strategy to all three real datasets (Additional file 1: Figure S19). We observed that in all three datasets, Cusanovich2018 performs well in identifying the isolated major groups and the identified clades match well the labels provided, including FACS-sorted labels, cell-ranger clustering solutions, and known tissue labels. Overall, the Cusanovich2018 “pseudo bulk” strategy for defining de novo peaks is able to capture the heterogeneity within single-cell populations and can serve as a promising unsupervised way to define pseudo-bulk subpopulations and to perform peak calling.

The effect of excluding regions using the ENCODE blacklist annotation

Blacklisted regions are those features annotated by ENCODE as belonging to a subset of genomic regions, which harbor the potential to produce artifacts in down-stream analyses. In order to assess the potential contribution that blacklist regions could have on the overall variation and population separability, we calculated (1) the proportion of reads mapped to blacklisted regions and (2) the proportions of bins with at least one read overlapping a blacklisted region vs the proportion of bins containing reads that do not overlap blacklist regions. Such a ratio corresponds closely to the feature set used by several of the evaluated methods.

We observed that for 10X Genomics and sci-ATACseq the percentage of reads mappable to blacklisted regions is only ~ 1%, while for the Fluidigm C1 based assay used in Buenrostro2018 is much higher, ~ 50%. However, when considering bins, only ~ 0.01 – 0.02% of bins with at least one read correspond to blacklist regions for all three technologies (Additional file 1: Figure S20). This fraction of bins containing one or more reads in a blacklisted region is likely negligible, and we hypothesize that the variation in the signal from reads in blacklisted regions is similarly negligible. It is worth noting that cisTopic, Scasat, SCRAT, and SnapATAC employ a blacklist filtering step to remove features annotated by ENCODE as belonging to a subset of genomic regions, which harbor the potential to produce artifacts in downstream analysis steps (HM Amemiya et al. 2019). Our benchmarking pipeline makes use of the ENCODE ATAC-seq pre-processing pipeline to call peaks, therefore the peaks overlapping with regions on the blacklist annotation list are already removed before implementing scATAC-seq methods. SCRAT and SnapATAC do not use peaks as features; therefore, they are the only methods potentially affected by blacklist-mapped artifacts. We tested whether we would observe any change in downstream clustering performance upon opting to perform a blacklist removal step. Through a qualitative and quantitative comparison of clustering performance across the datasets generated by the three different technologies (10X Genomics, sciATAC, and Fluidigm C 1), we determined that methods, which remove features according to blacklist annotations show no considerable advantage over those that permitted such features (Additional file 1: Figure S21).

Rare cell type-specific peak detection

As all cell identities may not be pre-defined in complex tissue types, we sought to examine PB-All and PB-Clades strategies to infer a chromatin accessibility feature set from the scATAC-seq libraries directly. To achieve this, we established a simulation setting where we mixed bulk ATAC-seq data from three sorted populations (B cells, CD4+ T cells, and monocytes from the 10X PBMCs dataset) that would be mixed in complex tissue (i.e., peripheral blood mononuclear cells) (Additional file 1: Figure S18b). After peak calling on both

the synthetic bulk and isolated reads from each cell type, we inferred the proportion of cell type-specific peaks from the minor cell population that were captured by the peak calling in the synthetic bulk mixture (see the “Methods” section). Overall, the results indicate that cell type-specific peaks may be vastly underestimated from performing peak calling on the mixture of single cells (PB-All) (Additional file 1: Figure S18b). Specifically, only ~ 18% of cell type-specific peaks from very rare (1% prevalence) or ~40% from rare (5% prevalence) cell populations were detected when peaks were called when treating the heterogenous source as a synthetic bulk experiment. Consequently, as these peaks would be vastly underrepresented in a consensus peak set, virtually all computational algorithms will fail to identify rare populations. Moreover, as many common quality-control measures for scATAC involve filtering based on the proportion of reads in peaks, these cell populations may be underrepresented in quality-controlled datasets. As observed in other studies (DA Cusanovich et al. 2018, AT Satpathy et al 2019), these results suggest calling peaks on PB-All may result in suboptimal performance. Alternatively, when isolated populations have been profiled (for example by FACS), peak sets can be defined by calling peaks using data from cells in each predefined population separately as discussed in the previous section since this enables the resolution of rare subpopulations (for example HSC in the hematopoietic system).

Frequency-based peak selection vs intensity-based peak selection

Cusanovich2018 selects peaks that are present in at least a specified percentage of cells before performing TF-IDF transformation, while scABC selects peaks with the most reads to cluster cells. To evaluate the effect of selecting peaks based on their representation in the cell population or based on their intensity (defined as the sum of reads in that peak in all samples), we focus on the two methods that implement the step of peak selection, Cusanovich2018 and Control-Naïve (equivalent to scABC+PCA).

To assess the two peak selection strategies, we ran both Cusanovich2018 and Control-Naïve on both simulated bone marrow dataset at noise level of 0.2 with a coverage of 2500 fragments and the Buenrostro2018 dataset by varying the cutoffs for peak inclusion (Additional file 1: Figures S22-S23). We calculated the intensity of peaks by counting the number of reads across all cells and calculated the frequency of peaks by counting the number of cells in which a peak is observed. For this analysis, we selected the top peaks based on intensity and frequency with the following cutoffs: top 100%, 80%, 60%, 40%, 20%, 10%, 8%, 6%, 4%, 2%, and 1%.

For both Cusanovich2018 and Control-Naïve, the two peak selection strategies have similar clustering result scores when varying the cutoff (Additional file 1: Figures S22a-b, S23a-b). We observed reasonable and stable clustering performance using more than 20% of the ranked peaks. As the number of peaks is reduced, the scores start to decline noticeably and decrease almost monotonically. Below 1%, both methods perform poorly. In

addition, we observed that the Louvain method produces more stable results than hierarchical clustering and k-means across the considered settings.

Running time of different methods

In our analysis, we also collected the running time of each method on both simulated and real datasets (see Additional file 1: Note S6). For the simulated datasets, we only reported the execution time necessary to build a feature matrix starting from a peaks-by-cells count matrix. For real datasets, we considered the execution time to build a feature matrix from bam files. The running times are shown in Additional file 1: Figure S24 (Additional file 1: Table S20). All the tests were run on a machine with an Intel Xeon E5-2600 v4 X CPU with 44 cores and 1 TB of RAM with the CentOS 7 operating system. When analyzing real datasets with methods that rely on peaks but do not provide an explicit function to construct a peaks-by-cells matrix (Cusanovich2018, Cicero, Gene Scoring, and Scasat), we ran the same script on a Linux cluster to obtain the peaks-by-cells matrix such that the execution time of this step is equivalent across these methods. It is worthwhile to mention that not all the methods of this benchmark support parallel computing. For the methods that support parallel computing, including SnapATAC, chromVAR, and cisTopic, the execution time was reported using 10 cores. For the rest of the methods, we run them using a single core. We selected this number reasoning that a typical lab may not have access to a machine with 44 cores and instead may use a mid-size computing node with 8 –12 cores. Notably, SnapATAC was the only method capable of processing the full sci-ATAC-seq mouse dataset (~80,000 single cells).

As shown in Additional file 1: Figure S24, BROCK MAN and SCRAT have the largest execution times in all the real datasets while the methods that use a custom script to obtain a peaks-by-cells matrix tend to have shorter execution times (e.g., Scasat, Cusanovich2018, Gene Scoring). We also assessed the scalability of methods with respect to the increasing coverage (250, 500, 1000, 2500, and 5000 fragments per peaks). We observe that with the increase of read coverage, for cisTopic, there is an exponential increase of the running time whereas for other methods, the running time stays stable or increases linearly (Additional file 1: Figure S24, Additional file 1: Table S21).

Finally, we compared execution time vs clustering performance (Fig. 7c). Interestingly, the most accurate methods (SnapATAC, cisTopic, and Cusanovich2018) have a reasonable running time while outperforming the other methods for clustering quality across all the datasets. Considering the computational time as an important factor that must be carefully evaluated before the implementation of any bioinformatics pipeline, we believe that Cusanovich2018 is the best in balancing clustering performance with execution time.

Discussions

scATAC technologies enable the epigenetic profiling of thousands of single cells, and many computational methods have been developed to analyze and interpret this data. However, the sparsity of scATAC-seq datasets provides unique challenges that must be addressed in order to perform essential analyses such as cluster identification, visualization, and trajectory inference (H. Chen et al. 2019, X Qui et al. 2017). Moreover, the rapid technological innovations that facilitate profiling accessible chromatin landscapes of 104 or 105 cells provide additional computational challenges to efficiently store and analyze data.

In this study, we compared ten computational methods developed to construct informative feature matrices for the downstream analysis of scATAC-seq data. We developed a uniform processing framework that ranks methods based on their ability to discriminate cell types when combined with three common unsupervised clustering approaches, followed by evaluation of three well-accepted clustering metrics. We evaluated these methods on 13 datasets, three of those obtained using different technologies (Fluidigm C1, 10X, and sci-ATAC) and five consisting of simulated data with varying noise levels. These datasets comprise cells from different tissues in both mouse and human. In addition to identifying various methodologies that perform optimally on real and simulated data, our benchmarking examination of scATAC-seq methodologies reveals general principles that will inform the development of future algorithms. First, peak-level or bin- level feature counting generally performs better in distinguishing different cell types followed in turn by k-mer-level, TF motif-level, and gene-centric-level summarization. We interpret this finding as an indication of the complexity of gene regulatory circuits where precise enhancer elements may have distinct functions that cannot be sufficiently approximated by sequence context or proximity to gene bodies alone. Second, we note that the methods that implement a dimensionality reduction step generally perform better in the separation of cell types, since this step may help to remove the redundancy between a large number of raw features and to mitigate the effect of noise. Third, for the methods that do not implement a dimensionality reduction step, simply adding a PCA step could significantly improve the clustering results. In fact, PCA generally boosts Louvain clustering results. For methods that do not account for the differing sequencing coverage of cells, the first PC could be used to capture and correct for sample depth differences. In this case, removing the first PC may improve the performance of these methods. Fourth, we observe that the Louvain method overall performs more consistently and accurately than k-means and hierarchical clustering. In contrast, k-means and hierarchical clustering are more sensitive to outliers and may result in suboptimal clustering solutions since some of clusters may correspond to single or few outlier cells. Fifth, the robustness of different methods to noise and coverage varies among different datasets. Among the top three methods, cisTopic is the most penalized by low coverage. Sixth, it was also observed that inappropriate transformations, such as log2 transformation and

normalization based on region size as implemented in SCRAT, may impact negatively clustering performance.

We observe that many methods fail to scale to larger datasets, which are now available due to improvements in split-pool technology and droplet microfluidics. As technologies improve and individual labs and international consortia lead efforts to generate ever larger single-cell datasets, scalability will be an unavoidable goal of method developments on a par with accuracy. As many of our evaluated methods were designed in the context of data generated from the Fluidigm C1 platform (which produces ~ 102 cells), such approaches were often incapable of analyzing large datasets. In particular, the sci-ATAC-seq mouse dataset served as a useful resource to test the scalability of the methods that were benchmarked (~ 80,000 cells). Notably, our evaluation demonstrates that only SnapATAC was able to scale to process and analyze this large dataset. Future methods must be capable of processing datasets of this size especially adopting efficient data structures that allow out of core computing. Our findings reinforce the need for methods that not only are accurate but highly scalable for scATAC-seq data processing. Defining regions is an important step in constructing feature matrices. Selecting informative regions generally improves downstream analyses such as clustering to capture heterogeneity within cell populations. Peak calling is a popular and straightforward way to define regions of interest. We observe that clustering performance is not generally impacted by using peaks defined from bulk ATAC-seq data vs using peaks obtained from aggregating single-cell data based on FACS-sorting labels. However, performing peak calling by simply pooling reads from single cells may obfuscate peaks specific to rare cell populations leading to failures in uncovering them. In addition, the Cusanovich2018 approach to identify pseudo-bulk clades is a promising unsupervised way to perform in silico sorting without relying on FACS-sorting labels. This strategy potentially serves as a suitable way to preserve peaks specific to rare cell types. Also choosing an appropriate number of peaks is important for improving the downstream analysis (for example based on intensity/frequency-based given that they perform similarly). We are aware of the current limitations in our benchmarking effort. We have compared single-cell ATAC-seq methods based on their ability to separate discrete cell populations; however, this might not be ideal when dealing with a continuous cell lineage landscape. We observe that chromVAR generally works better in preserving a continuous space while SnapATAC tends to break a putative landscape into discrete populations. The choice of method is ultimately case-specific and may be driven by the downstream application. For example, the feature matrix obtained by chromVAR may be more suitable for trajectory inference (H Chen et al. 2019) while the one obtained from SnapATAC may be more appropriate to better identify discrete and well-separated cell populations by clustering. We acknowledge also that not all tested methods were specifically designed to produce clustering results. For example, chromVAR, Cicero, and Gene Scoring were designed to determine important marker genes and their regulatory logic or to infer enriched TF binding sites within accessible

chromatin regions. However, because clustering is a critical part of single-cell analysis and researchers frequently use output from all methods to produce clustering results (DA Cusanovich et al. 2018), we felt that evaluating the clustering abilities using feature matrices produced by each method was a useful measure. An additional limitation of our study is that it is impossible to create a simulation framework that models an experimental outcome with perfect accuracy. Several assumptions were made to enable our simulation of the data; these assumptions are described in the “Methods” section, where we detail explicitly how the simulated data was generated.

Interestingly, we learnt that some combinations of feature matrices with the simple clustering approaches included in our benchmarking framework perform even better than the original combination proposed by the respective authors. This highlights the value of this dual characterization (user vs designer perspective) and provides a summary of both perspectives to the readers. We believe it is important to stress the distinction between biological realities and computational performance, especially in the context of unsupervised clustering. A big and critical assumption (or hope) of our field is that an unsupervised clustering procedure will provide clustering solutions that recapitulate different populations corresponding to different cell types/states. Given that for several real datasets the ground truth is not known, a current compromise during the exploratory clustering analysis is to use known marker genes, sorted populations, or known tissues to validate the clustering solutions based on classic metrics. If we embrace this assumption, keeping in mind that additional validation is required to truly delineate the subpopulation structure of a population of cells, the two views, biological and computational, can be reconciled. Our benchmark procedure is aimed to provide some guidelines based on explorative analyses that are currently adopted in several published papers.

Looking forward, due to the wealth of data being produced by new scATAC technologies, we hypothesize that more powerful machine learning frameworks may be able to uncover complex cis and trans relationships that define cell-cell relatedness. Specifically, we anticipate autoencoder-like models that integrate genomic sequence context, gene body positions, and precise accessible chromatin information will yield information enrich features and that more advanced manifold learning methods will help to remove redundancy and better preserve heterogeneity within single-cell populations. Such achievements may enable us to overcome the inherent sparsity and high dimensionality that characterizes scATAC-seq data.

Material & Methods

Our assessment of methods was based on public scATAC-seq datasets made available in public repositories by the respective authors (see the “Availability of data and materials” section). As such, we refer to the original publications

for further details on the experimental design and data pre-processing/alignment. For peak calling, we used the ENCODE pipeline (https://www.encodepro-ject.org/atac-seq/) except for the 10X PBMCs data for which peaks were already available through the Cell Ranger pipeline optimized for this technology. Whenever changes were required for running a given method, those are noted in the respective sections.

Datasets

Human hematopoiesis I (Buenrostro et al. (JD Buenrostro et al. 2018))

This dataset comprised of 10 FACS-sorted cell populations from CD34 + human bone marrow, namely, hematopoietic stem cells (HSCs), multipotent progenitors (MPPs), lymphoid-primed multipotent progenitors (LMPPs), common myeloid progenitors (CMPs), granulocyte-macrophage progenitors (GMPs), megakaryocyte-erythrocyte progenitors (MEPs), common lymphoid progenitors (CLPs), plasmacytoid dendritic cells (pDCs), monocytes, and an uncharacterized CD34+ CD38− CD45RA+ CD123− cell population. A total of 2034 cells from six human donors were used for analysis. A peak file (including 491,437 peaks) obtained from bulk ATAC-seq dataset was provided.

sci-ATAC-seq mouse tissues (Cusanovich et al. (DA Cusanovich et al. 2018))

This dataset comprises cells from 13 tissues of adult mouse, namely, the bone marrow, cerebellum, heart, kidney, large intestine, liver, lung, prefrontal cortex, small intestine, spleen, testes, thymus, and whole brain, with over 2000 cells per tissue. A total of 81,173 cells from 5 mice were used for analysis. A subset was obtained by randomly downsampling 15% cells from each tissue and was comprised of 12,178 cells.

Human hematopoiesis II (10X PBMCs)

This dataset is composed of peripheral blood mononuclear cells (PBMCs) from one healthy donor. A total of 5335 cells were used for analysis.

Simulated scATAC-seq datasets

In order to evaluate and benchmark various approaches, we generated synthetic (labeled) data from downsampling 18 FACS-sorted bulk populations that were previously described (J.C. Ulirsch et al. 2019). For ease of interpretation, we considered only 6 isolated populations (HSC, CMP, NK, CD4, CD8, erythroblast). For the erythropoiesis simulation, eight additional populations (P1–P8) originally described in (Ludwig S. Leif et al. 2019) were also considered.

Our simulation framework starts with a peak x cell type counts matrix (from bulk ATAC-seq) and generates a single-cell counts matrix C for an arbitrary

number of synthetic single cells. Explicitly, for a simulated single cell j and corresponding peak i from bulk cell type t ,we seek to generate $c_{i,j}$, where $c_{i,j} \in \{0,1,2\}$, noting that these values correspond to possible observations in a diploid genome. Next, we define the rate (r_i^t) at which the peak i is prevalent in the bulk ATAC-seq data for cell type t. This rate is determined by the ratio of reads observed in peak i over the sum of all reads. Assuming a total of k peaks for the matrix C and for user-defined parameters q (noise parameter; $q \in [0,1]$) and n (number of simulated fragments), we define $c_{i,j}$ as follows:

$$c_{i,j} \quad rbinom\left(2, p_i^t\right)$$

where

$$p_i^t = (r_i^t)\left(\frac{1}{2}n\right)(1-q) + (1/k)\left(\frac{1}{2}n\right)(q)$$

Intuitively, the parameter p_i^t defines the probability that a count will be observed in peak i for a single cell. Additionally, p_i^t can be decomposed into the sum two terms. As $q \to 0$, the first term dominates, and the probability of observing a count in peak i is simply the scaled probability of the ratio of reads for that peak from the bulk ATAC-seq data (r_i^t). Thus, when $q=0$, the simulated data has no noise. Conversely, as $q \to 1$, the second term dominates, and p_i^t reduces to a flat probability that is no longer parameterized by the peak i or cell type t and thus represents a random distribution of n fragments into k peaks.

The noise level we simulated attempts to mimic the non-specific cutting from Tn5. To give a sense of the range of this parameter on real data, we considered simply the number of reads falling outside peaks over the total number of reads (excluding blacklisted regions). This calculation assumes that reads in regions defined as peaks by a bulk or pseudo-bulk measurement will be dominated by specific cutting and that regions outside peaks will be dominated by non-specific cutting. Using this approach, we estimated the following noise levels: 0.38 for the Buenrostro2018 dataset, 0.22 for the 10X PBMC dataset, and 0.62 for the sci-ATAC-seq mouse dataset. We would like to point out that these rates may be slightly underestimated; a more careful estimation would require one to consider the fact that, at any given region of the genome, reads could be observed from specific and non-specific cutting.

For bone marrow-based simulations, we simulated 200 cells per labeled cell type while for erythropoiesis-based simulation we simulated 100 cells per

labeled cell type. Eventually, we have 1200 cells for each simulated dataset. In the base simulations, we parametrized n =2500 fragments in peaks in expectation for all cells. For additional simulations that compared different data coverages, we set n to various values (5000, 2500, 1000, 500, and 250 respectively) to benchmark this effect. To evaluate the effect of noise in our simulation, we set q to three values (0, 0.2, 0.4) to benchmark the robustness to noise. At values of $q>0.4$, no method could reliably separate all the subpopulations. Finally, since our simulation started at the reads in the peak level, for some methods, the core algorithm associated with the method was extracted in order to benchmark it in this setting. Additionally, full code to reproduce these simulated dataset matrices has been made available with our online code resources.

Peak calling

For real datasets, peaks were called using the ENCODE ATAC-seq processing pipeline (https://www.encodeproject.org/atac-seq). Briefly, single cells were aggregated into cell populations according to cell type, obtained either by FACS sorting or by tissue of origin. Peaks were called for each cell population and merged into a single file with bedtools (AR Quinlan et al. 2010).

Building the feature matrix

BROCKMAN

This method starts by defining regions of interest, which will be scanned for k-mer content, as 50 bp windows around each transposon integration site and merging overlapping regions. Then, a frequency matrix of k-mers-by-cells is built by counting all possible gapped k-mers (for k from 1 to 8) within the previously defined windows. This frequency matrix is scaled so that each k-mer has mean 0 and standard deviation 1. Principal component analysis (PCA) is applied to the scaled k-mers-by-cells frequency matrix, and significant principal components (PCs) as estimated with the jackstraw method are selected to build a final feature matrix for downstream analyses.

chromVAR

This method starts by counting reads under chromatin-accessible peaks in order to build a count matrix of peaks-by-cells (X). Then, a set of chromatin features such as transcription factor (TF) motifs or k-mers are considered. Reads mapping to each peak that contains a given TF motif (or k-mer) are counted in order to build a count matrix of motifs-by-cells or k-mers-by-cells (M). Moreover, a raw accessibility deviation matrix of motifs (or k-mers)-by-cells (Y) is generated by calculating the difference between M and the expected number of fragments based on X. Then, background peak sets are created for each motif (or k-mer) to remove technical confounders.

Background motifs (or k-mers)-by-cells raw accessibility deviations are then used to calculate a bias-corrected deviation matrix and to compute a deviation z-score used for downstream analyses.

cisTopic

This method starts by building a peaks-by-cells binary matrix by checking if a peak region is accessible, i.e., at least one read falls within the peak region. Then, latent Dirichlet allocation (LDA) is performed on this binary matrix, and two probability distributions are generated, a topics-by-cells probability matrix and a regions-by-topics probability matrix. The former is the final feature matrix for downstream analyses.

Cicero

This method defines promoter peaks as the union of annotated TSS minus 500 base pairs and macs2 defined peaks around the TSS. It takes as input the peaks-by-cells binary matrix. It also requires either pseudo temporal ordering or coordinates in a low-dimensional space (t-SNE) so that cells can be readily grouped. It then computes the co-accessibility scores between sites using Graphical Lasso. To get the gene activity scores, it selects sites that are proximal to gene TSS or distal sites linked to them and weight them by their co-accessibility. Then, all the sites are summed and weighted according to their co-accessibility to produce a genes-by-cells feature matrix that is used in this benchmarking analysis.

Gene Scoring

This method first constructs a peaks-by-cells count matrix and defines regions of interest as the 50 kb up-stream and downstream of gene TSSs. Then, it finds the overlap between ATAC-seq peaks and TSS regions and the peaks are weighted by a function of the distance to the linked genes. Finally, the peaks-by-cells count matrix is converted into genes-by-cells weighted count matrix by multiplying the weighted peaks by genes matrix. The genes-by-cells weighted count matrix is the final feature matrix for downstream analyses.

Cusanovich2018

This method starts by binning the genome into fixed-size windows (by default, 5 kbp), and building a binary matrix from evaluating whether any reads map to each bin. Bins that overlap ENCODE-defined blacklist regions are filtered out, and the top 20,000 most commonly used bins are retained. Then, the bins-by-cells binary matrix is normalized and rescaled using the term frequency-inverse document frequency (TF-IDF) transformation. Next, singular value decomposition (SVD) is performed to generate a PCs-by-cells LSI score matrix, which is used to group cells by hierarchical clustering into differ-ent clades. Within each clade, peak calling is performed on the aggregated scATAC-seq profiles, and identified peaks are combined into a new peaks-by-cells binary matrix. Finally, the new peaks-by-cells matrix is

trans-formed with TF-IDF and SVD as before to get a matrix of PCs-by-cells, which is the final feature matrix for downstream analyses.

scABC

This method starts by building a peaks-by-cells count matrix of read coverage within peak regions. Then, the weights of cells are calculated by a nonlinear transformation of the read coverage within the peak background, defined as a 500-kb region around peaks. Since the weights will be used as part of weighted K-medoids clustering to define cell landmarks and further perform finer re-clustering instead of normalizing the peaks-by-cells matrix, the feature matrix in scABC is defined as the peaks-by-cells count matrix.

Scasat

This method first constructs a peaks-by-cells binary accessibility matrix by checking if at least one read overlaps with the peak region. Then Jaccard distance is computed based on the binary matrix to get a cells-by-cells dissimilarity matrix. Multidimensional scaling (MDS) is further performed to reduce the dimension and to generate the final feature matrix for downstream analysis.

SCRAT

This method starts by aggregating reads from each cell according to different features (such as TF motifs or region of interest of each gene), and then building a count matrix of features-by-cells. The features-by-cells count matrix is normalized by library and region size to get the final feature matrix for downstream analyses.

SnapATAC

This method starts by binning the genome into fixed-size windows (by default 5 kb) and estimating read coverage for each bin to build a bins-by-cells binary count matrix. Bins that overlap ENCODE-defined black-list regions are filtered out, as well as those with exceedingly high or low z-scored coverage. Then, the bins-by-cell matrix is transformed into a cells-by-cells Jaccard index similarity matrix, which is further transformed by normalization and regressing out coverage bias between cells. Finally, PCA is applied to the normalized similarity matrix, and the top PCs are used to build a PCs-by-cells matrix that is the final feature matrix for downstream analyses.

Clustering

For this study, we used three commonly used clustering methods: k-means, hierarchical clustering (with default ward linkage) as implemented in the scikit-learn library (F Pedregosa et al. 2011), and Louvain clustering (a community detection-based method) (VD Blondel et al. 2008, JH Levine et al. 2015) as implemented in Scanpy (FA Wolf et al. 2018). For both hierarchical clustering and k-means, we set the number of clusters to the number of unique FACS-sorted labels or known tissues. In the 10X PBMCs scATAC-seq dataset, which lacks the FACS-sorted labels, we instead set the number of clusters to 8 since this is the expected number of populations based on previous studies (HA Pliner et al. 2019). For the Louvain algorithm, we set the size of local neighborhood to 15 for all the datasets. Since Louvain method requires "resolution" instead of the number of clusters and different number of clusters will affect the clustering evaluation, to make the comparison fair, we use the binary search algorithm on the "resolution" (ranging from 0.0 to 3.0) to find the same number of clusters as the other two clustering methods. If the precise number of clusters did not match the desired value, the "resolution" value inducing the closest number of clusters to the desired value was used.

Metrics to evaluate the clustering

To evaluate clustering solutions for datasets with a known ground truth (i.e., for each cell, we have a label that indicated the cell type), we used three well-established metrics: adjusted Rand index (ARI), mutual information, and homogeneity. Briefly, for the ARI, first, the Rand index (RI) is defined as a similarity measure between two clusters considering all pairs of samples assigned in the same or different clusters in the predicted and true clustering. Then, the raw RI score is adjusted for chance in the ARI score as described in the following formula:

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}$$

where RI is the pre-computed Rand index and E is the expected Rand index.

Mutual information is a measure of the mutual dependence between two variables. The mutual information value is computed according to the

following formula, where |Ui| is the number of the samples in cluster Ui and |Vj| is the number of the samples in cluster Vj:

$$\mathrm{MI}(U,V)=\sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\frac{|U_I\cap V_j|}{N}\log\frac{N|U_i\cap V_j|}{|U_i||V_j|}$$

The homogeneity score is used to check if the algorithm used for the clustering can assign to each cluster only samples belonging to a single class. Its value h is bounded between 0 and 1, and a low value indicates low homogeneity and vice versa. The score is computed as follows:

$$h=1-\frac{H(Y_{true}\vee Y_{pred})}{H(Y_{true})}$$

where H(Ytrue|Ypred) is the probability to assign true samples to a set of predicted samples, while H (Ytrue) are the labels of the samples.

To evaluate clustering solutions for the 10X PBMCs dataset, we proposed a simple score called the Residual Average Gini Index (RAGI) and compared the accessibility of housekeeping genes with previously characterized marker genes (HA Pliner et al. 2019). We reasoned that a good clustering solution should contain clusters that are enriched for accessibility of different marker genes, and each marker gene should be highly accessible in only one or a few clusters. First, to quantify the accessibility of each gene in each cell, we used the Gene Scoring approach described above. Briefly, the accessibility at each TSS is the distance-weighted sum of reads within or near the region. Second, to quantify the enrichment of each gene in each cluster of cells, we computed the mean of the accessibility values in all cells for each cluster. Third, based on the vector of mean accessibility values (one per cluster), we computed the Gini index (C Gini et al. 1997) for each marker gene. The Gini index measures how imbalanced the accessibility of a gene is across clusters. This score is bound by [0,1] where 1 means total imbalance (i.e., a gene is accessible in one cluster only) and 0 means no enrichment. This score has been previously used on scRNA-seq to perform clustering (J Linang et al. 2016, D Tsoucas et al. 2018). As a control, we also calculated the Gini index for a set of annotated housekeeping genes reported in https://m.tau.ac.il/~elieis/HKG/HK_genes.txt. Housekeeping genes should show minimal specificity for any given cluster since, by definition, they are highly expressed in all cells. Based on the set of Gini index values for marker and housekeeping genes, we calculated several metrics: (i) the mean Gini index for the two groups, (ii) the difference in means to assess the average residual specificity that a clustering solution has with respect to marker genes (this is our proposed RAGI metric), and (iii) the Kolmogorov-Smirnov statistic and its p value comparing the two groups of Gini indices for marker and housekeeping genes. We sorted the methods based on the descending order of the differences in means (Additional file1: Table S13); a positive value

indicates that the marker genes on average separate the clusters better than uninformative housekeeping genes.

Rare cell type-specific peak analysis

FACS-sorted bulk ATAC-seq data was downloaded and processed from a previously described resource (MR Corces et al. 2016). For each simulation, we created a randomly sampled set of 200 million unique (PCR-deduplicated) reads, which roughly represents a complexity similar to recommendations from the 10X Chromium scATAC-seq solution. Cell type-specific peaks were defined using the full dataset for each of the three cell types. Peaks were called using macs2 callpeak with custom parameters as in the ENCODE pipeline, i.e., "--nomodel --shift - 100 --extsize 200" to account for Tn5 insertions rather than read abundance when inferring peaks. Overlaps between the isolated minor population and the synthetic mixtures were computed using GenomicRanges (M Lawrence et al. 2013) findOverlaps function, which is equivalent to bedtools (AR Quinlan et al. 2010) overlap. For each minor population (B cell, CD4+ T cell, monocyte) and each prevalence (1, 5, 10, 20, 30%), each simulation was repeated 5 times for a total of 75 simulations. Reads from the other two (major) populations were sampled equivalently to make up the synthetic mixture for comparison.

References

Cusanovich DA, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. Cell. 2018;174(5):1309 – 24 e18.

Mereu E, et al. Benchmarking single-cell RNA sequencing protocols for cell atlas projects. BioRxiv:630087v1. 2019.

Ding J, et al. Systematic comparative analysis of single cell RNA-sequencing methods. BioRxiv:632216v2. 2019.

Schep AN, et al. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14(10):975–8.

de Boer CG, Regev A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. BMC Bioinformatics. 2018;19(1):253.

Ji Z, Zhou W, Ji H. Single-cell regulome data analysis by SCRAT. Bioinformatics. 2017;33(18):2930–2.

Corces MR, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48(10):1193–203.

Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20(5):273–82.

McInnes, L., J. Healy, and J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.

Pliner HA, et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell. 2018;71(5):858–71 e8.

Bravo González-Blas C, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nature Methods. 2019;16(5):397–400.

Cusanovich DA, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature. 2018;555(7697):538–42.

Cusanovich DA, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015;348(6237):910–4.

Lareau CA, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat Biotechnol. 2019.

Zamanighomi M, et al. Unsupervised clustering and epigenetic classification of single cells. Nat Commun. 2018;9(1):2410.

Baker SM, et al. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. Nucleic Acids Res. 2019;47(2):e10.

Fang R, et al. Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell types. BioRxiv:615179v2. 2019.

Mathelier A, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res.2015;44(D1):D110–5.

Ulirsch, J.C., et al., Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat Genet. 2019;51(4):683–93.

Leif S. Ludwig, et al., Transcriptional states and chromatin accessibility underlying human erythropoiesis. Cell Reports. 2019;27(11):3228–40.e7.

Buenrostro JD, et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell. 2018;173(6):1535–48 e16.

Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. Nat Methods. 2019;16(10):983–86.

Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

Bernstein BE, et al. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol. 2010;28(10):1045.

Yoshida H, et al. The cis-regulatory atlas of the mouse immune system. Cell. 2019;176(4):897–912 e20.

Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. R Package Version. 2011;100:4–3.

Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. Sci Rep. 2019;9(1):9354.

Satpathy, A.T., et al., Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol. 2019;37(8):925–36.

Chen, H., et al., Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat Commun. 2019;10(1):1903.

Qiu X, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979–82.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(Oct):2825–30.

Blondel VD, et al. Fast unfolding of communities in large networks. J Stat Mechanics. 2008;2008(10):P10008.

Levine JH, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97.

Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.

Gini C. Concentration and dependency ratios. Rivista di Politica Economica.1997;87:769–92.

Jiang L, et al. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016;17(1):144.

Tsoucas D, Yuan GC. GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. Genome Biol. 2018;19(1):58.

Lawrence M, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.

Chen, H., et al. Supporting data and source code for “Assessment of computational methods for the analysis of single-cell ATAC-seq data ”. 2019;Available from:https://github.com/pinellolab/scATAC-benchmarking/. Accessed 11 Nov 2019.

Figures

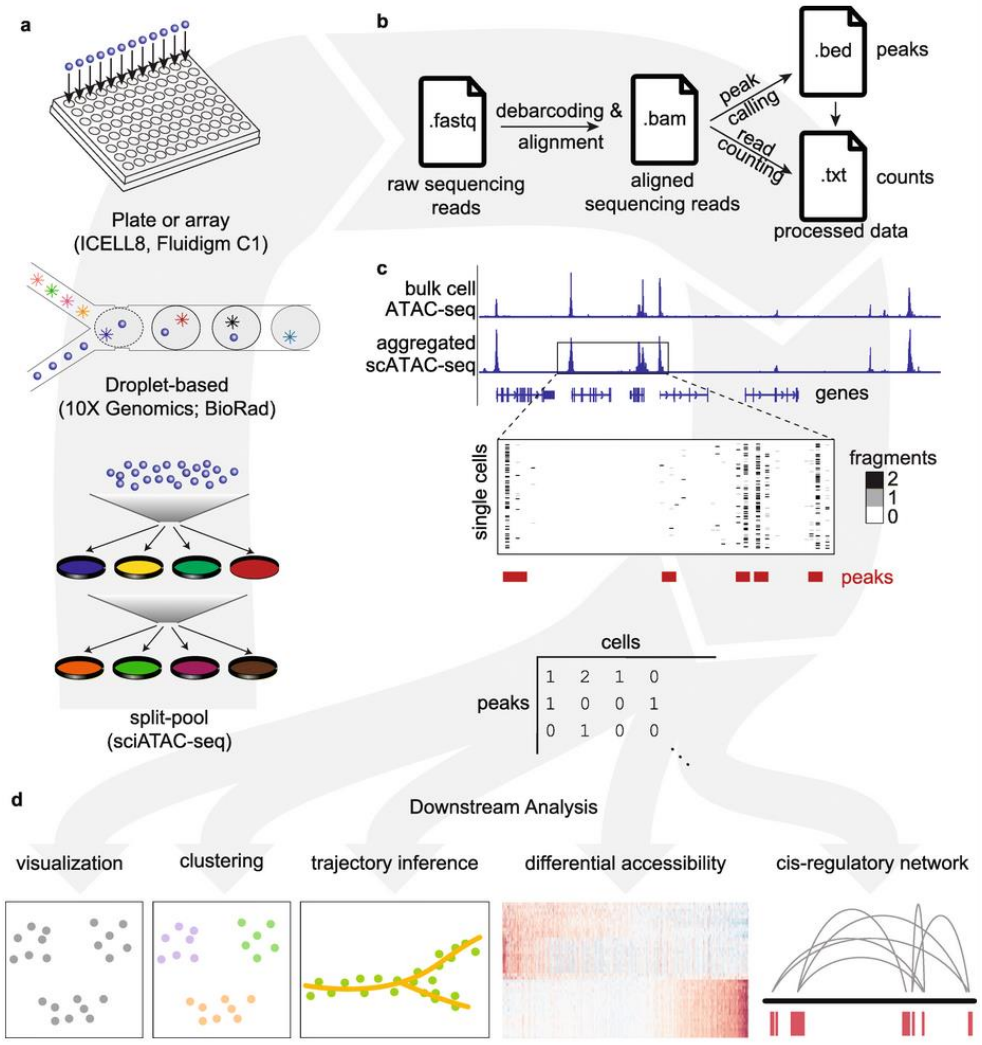


Figure 1. Schematic overview of single-cell ATAC-seq assays and analysis steps. **a** Single-cell ATAC libraries are created from single cells that have been exposed to the Tn5 transposase using one of the following three protocols: (1) Single cells are individually barcoded by a split-and-pool approach where unique barcodes added at each step can be used to identify reads originating from each cell, (2) microfluidic droplet-based technologies provided by 10X Genomics and BioRad are used to extract and label DNA from each cell, or (3) each single cell is deposited into a multi-well plate or array from ICELL8 or Fluidigm C1 for library preparation. **b** After sequencing, the raw reads obtained in .fastq format for each single cell are mapped to a reference genome, producing aligned reads in .bam format. Finally, peak calling and read counting return the genomic position and the read count files in. bed and .txt format, respectively. Data in these file formats is then used for downstream analysis. **c** ATAC-seq peaks in bulk samples can generally be recapitulated in aggregated single-cell samples, but not every single cell has

a fragment at every peak. A feature matrix can be constructed from single cells (e.g., by counting the number of reads at each peak for every cell). **d** Following the construction of the feature matrix, common downstream analyses including visualization, clustering, trajectory inference, determination of differential accessibility, and the prediction of *cis*-regulatory networks can be performed using the methods benchmarked in this manuscript

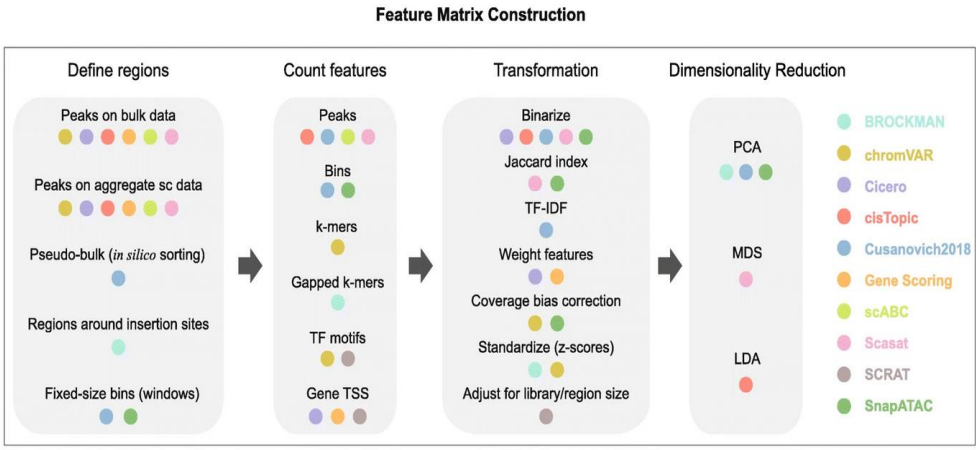


Figure 2. Benchmarking workflow. Starting from aligned read files in .bam format, feature matrices were constructed using each method. The feature matrix construction techniques used by each method were grouped into four broad categories: *define regions*, *count features*, *transformation*, and *dimensionality reduction*. A colored dot under a technique indicates that the method (signified by the respective color in the legend on the right) uses that technique. For each method, feature matrix files (defined as columns as cells and rows as features) are calculated and used to perform hierarchical, Louvain, and k-means clustering analysis. For datasets with a ground truth such as FACS-sorting labels or known tissues, clustering evaluation was performed according to the adjusted Rand index (ARI), adjusted mutual information (AMI), and homogeneity (H) scores. For datasets without ground truth, the clustering solutions were evaluated according to a Residual Average Gini Index (RAGI), a metric that compares cluster separation based on known marker genes against housekeeping genes. Lastly, a final score is assigned to each method.

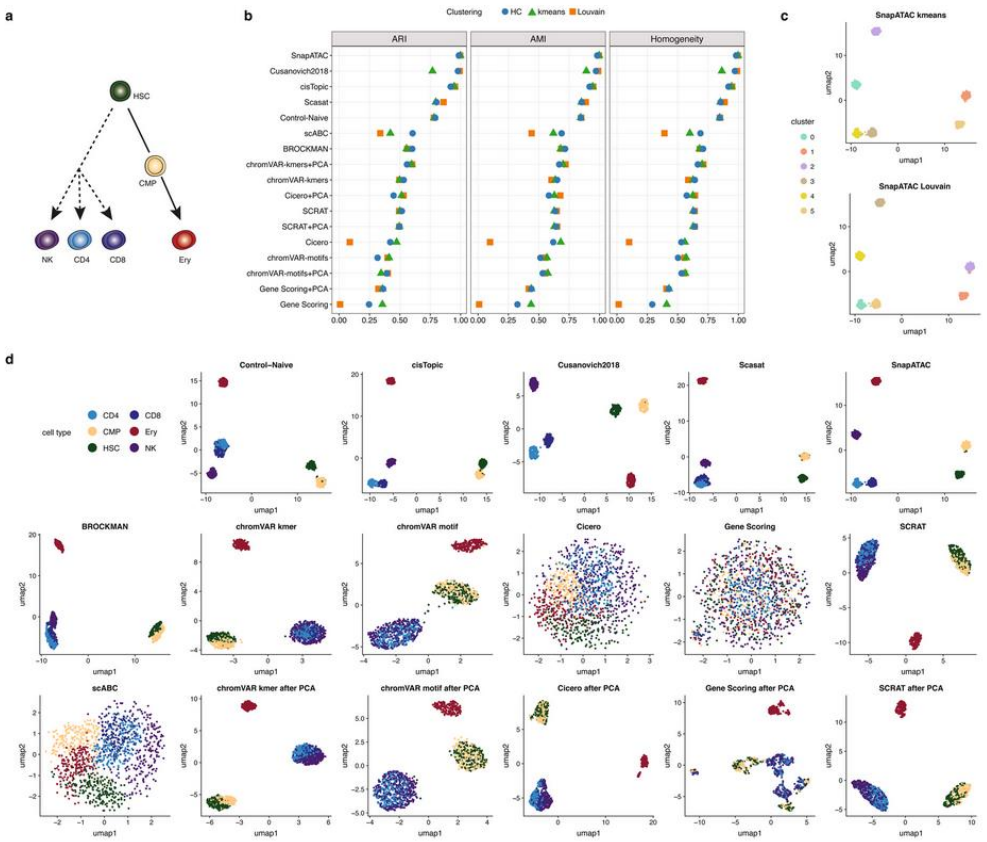


Figure 3 Benchmarking results in simulated bone marrow datasets at a noise level of 0.4 and a coverage of 2500 fragments. **a** Cell types used to create the simulated dataset. **b** Dot plot of scores for each metric to quantitatively measure the clustering performance of each method, sorted by maximum ARI score. **c** The two top-scoring pairings of scATAC-seq analysis method and clustering technique. Cell cluster assignments from each method are shown using the colors in the legend on the left. **d** UMAP visualization of the feature matrix produced by each method for the simulated dataset. Individual cells are colored indicating the cell type labels shown in **a**.

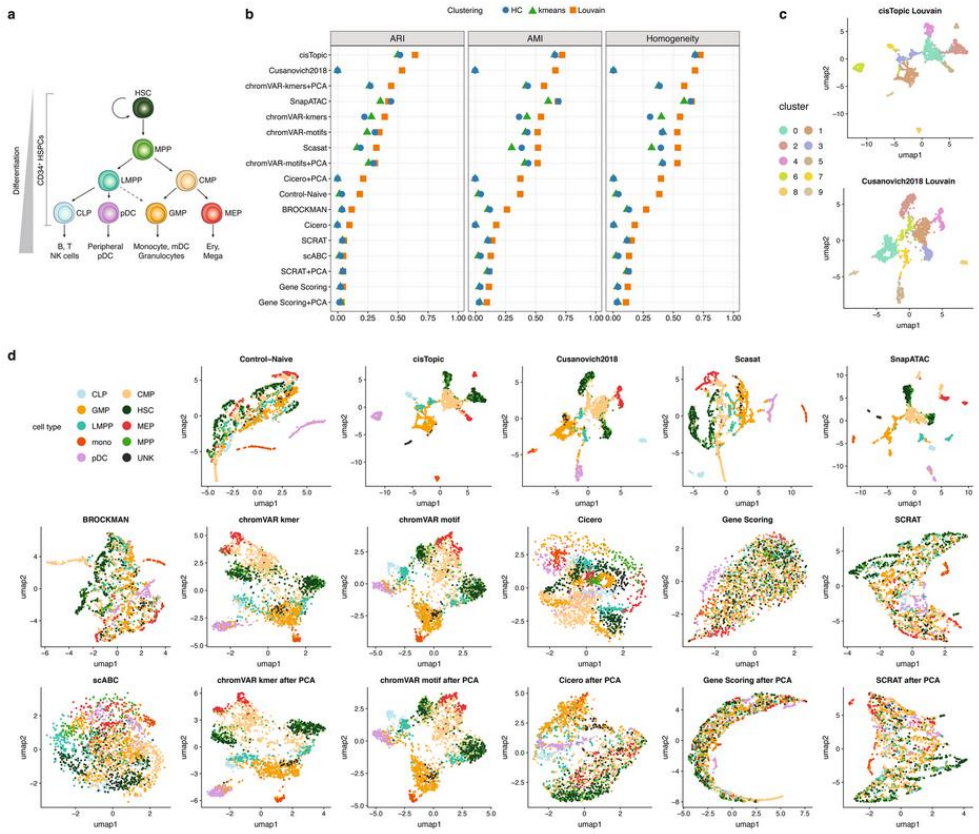


Figure 4. Benchmarking results using the *Buenrostro2018* scATAC-seq dataset. **a** Developmental roadmap of cell types analyzed. **b** Dot plot of scores for each metric to quantitatively measure the clustering performance of each method, sorted by maximum ARI score. **c** The two top-scoring pairings of scATAC-seq analysis method and clustering technique. UMAP visualization of the feature matrix produced by each method for the *Buenrostro2018* dataset. Individual cells are colored indicating the cell type labels shown in **a**.

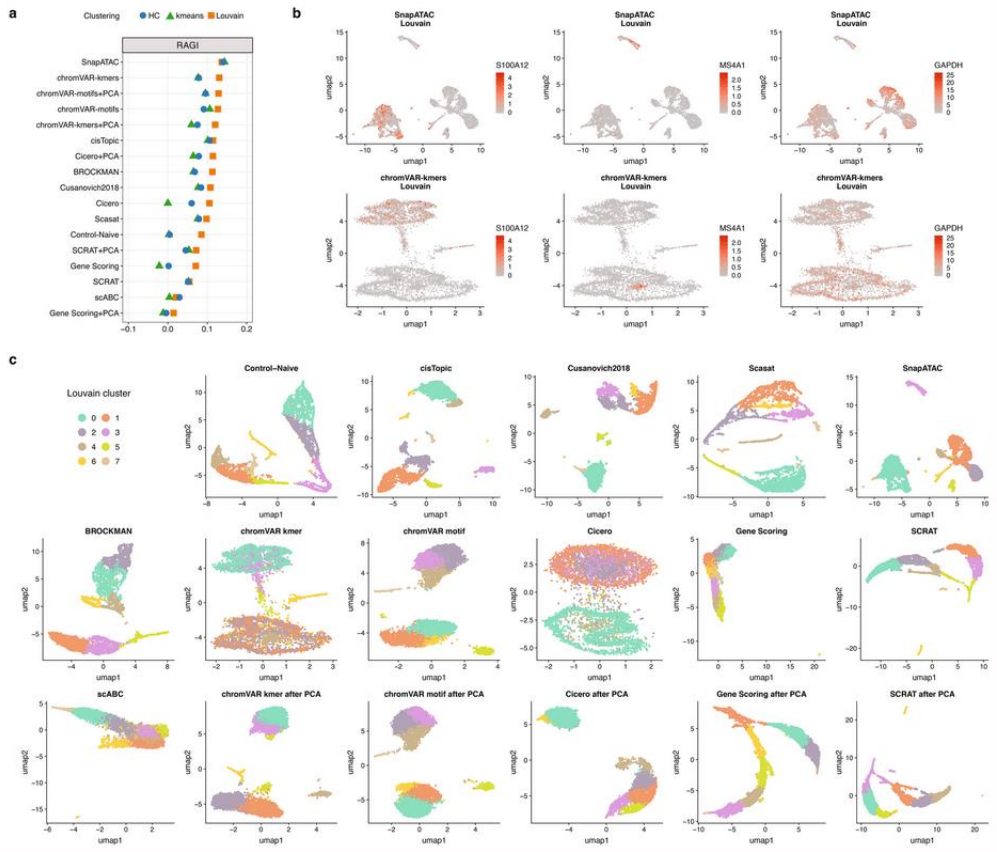


Figure 5. Benchmarking results using scATAC-seq data for 5k peripheral blood mononuclear cells (PBMCs) from 10X Genomics. **a** Dot plot of RAGI scores for each method, sorted by the maximum RAGI score. A positive RAGI value indicates that a method is able to produce a clustering of PBMCs in which chromatin accessibility of each marker gene is high in only a few clusters relative to the number of clusters with high accessibility of housekeeping genes. **b** UMAP visualization of the feature matrix produced by the top two methods (top row: SnapATAC, bottom row: chromVAR using k-mers). Chromatin accessibility of S100A12 (left, monocyte marker gene), MS4A1 (center, B cell marker gene), and GPDH (right, housekeeping gene) are projected onto the visualization. **c** UMAP visualization of the feature matrix produced by each method for the 5k PBMCs dataset from 10X Genomics. Individual cells are colored indicating cluster assignments using Louvain clustering.

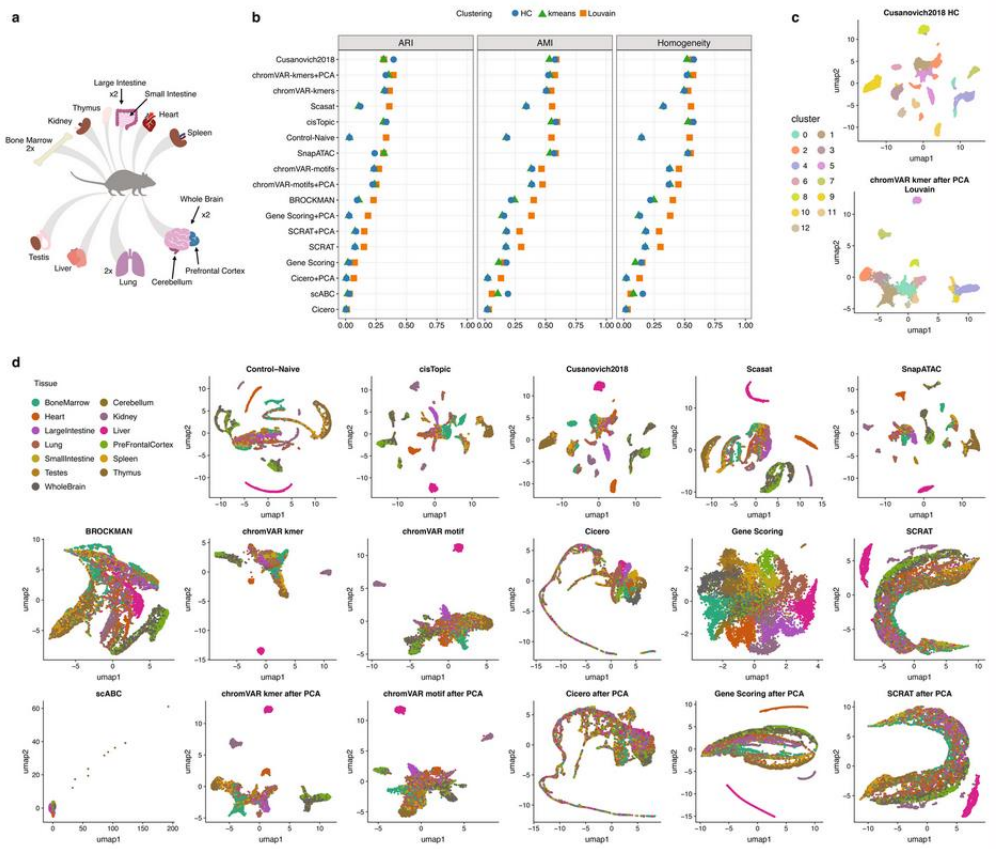


Figure 6. Benchmarking results using the downsampled sci-ATAC-seq mouse dataset from 13 adult mouse tissues. **a** schematic of 13 adult mouse tissues. Replicated tissues are indicated by “x2”. **b** Dot plot of scores for each metric to quantitatively measure the clustering performance of each method, sorted by maximum ARI score. **c** The two top-scoring pairings of scATAC-seq analysis method and clustering technique. Cell cluster assignments from each method are shown using the colors in the legend on the left. **d** UMAP visualization of the feature matrix produced by each method for the downsampled sci-ATAC-seq mouse dataset. Individual cell colors indicate the cell type.

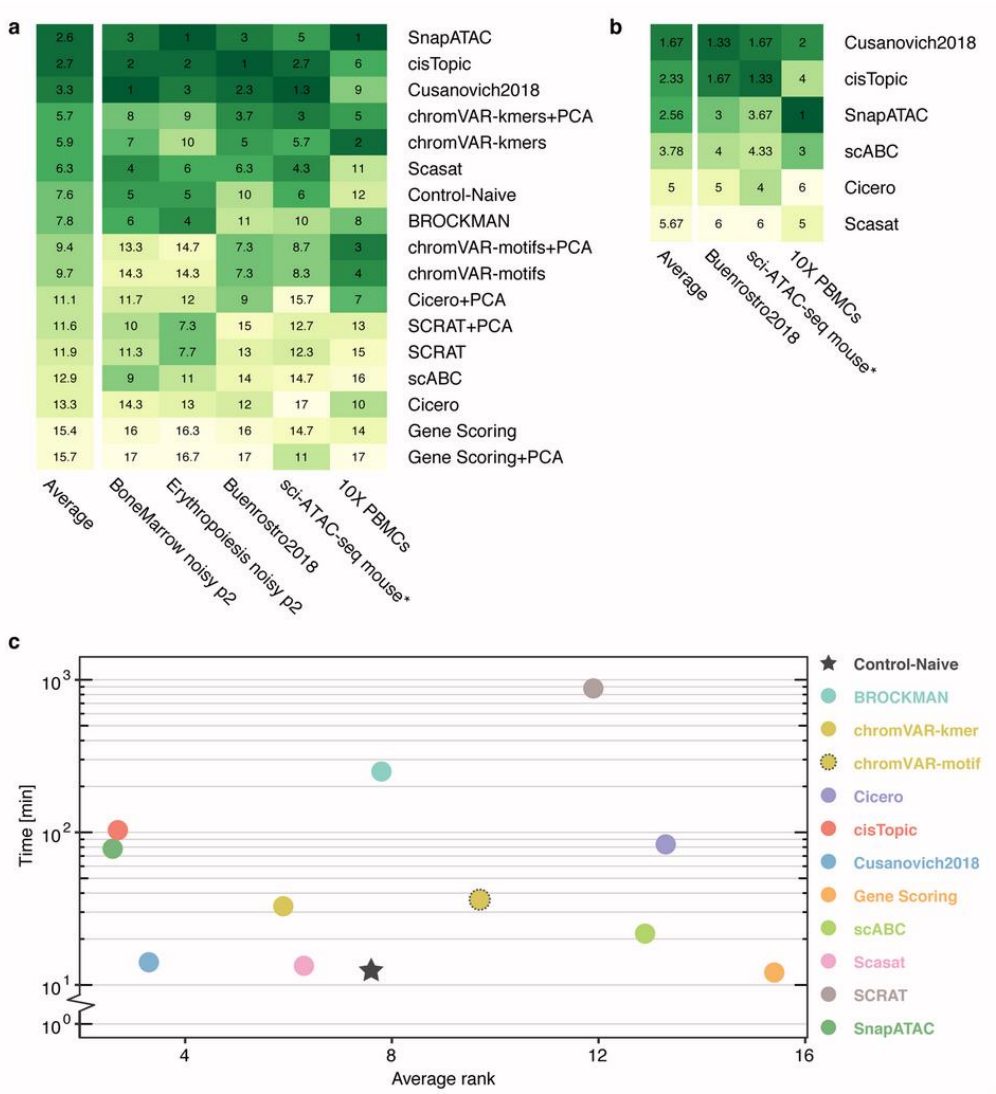


Figure 7. Aggregate benchmark results. **a** For each method, the rank based on the best-performing clustering method is measured for each metric (e.g., ARI, AMI, H, or RAGI). The average metric ranks for each dataset were used to calculate a performance score for each method. Each method was then assigned a cumulative average score based on its performance across all datasets. The asterisk indicates a downsampled dataset of the indicated original dataset. **b** For methods that specify an end-to-end clustering pipeline, average rank and cumulative average scores for each method were calculated as in **a**. **c** Plot of running time against performance for each method. Cumulative average scores, which were calculated in part **a** are shown on the *x*-axis, and the average running time across the three real datasets (*Buenrostro2018*, 10X PBMCs, and downsampled sci-ATAC-seq mouse) is shown on the *y*-axis.

Supplemental information

For simplicity and space constraints I will explain which file I have produced as supplementary that can be retrieved from the online paper:

Supplementary Note 5: Memory requirements and implementation choices

As mentioned in the main text, SnapATAC is the only method that allows to process successfully large datasets, as the sciATAC-seq mouse dataset with ~80000 cells. Here we investigate why the other methods failed to analyze this large dataset. We hypothesize that the main reason is related to the way the methods load/process the data in memory. In fact, we discovered that several methods require to load the entire dataset in the central memory (RAM).

BROCKMAN, Cicero and Gene Scoring try to load the entire dataset in memory using the *read.table* function or the *fread* function within the *data.table* package in R. Other methods such as: Cusanovic, Scrat, chromVAR, scABC and Scasat, store the entire dataset in memory within a *Matrix* object in R. CisTopic, has an optimized step to map the reads into the genome using the *Rsubread* function. This function creates a hash table of the entire genome and allows the user to select the amount of memory to use. At the end, the entire dataset is stored in the computer memory in a *CisTopicObject* data structure.

SnapATAC, preprocess the entire dataset and store it a *.snap* file. This file is based on the HDF5 technology that allows out of core computation. In SnapATAC the Python library h5py (a wrapper for HDF5 core library) is used to create the custom snap file format. More information about this custom file are available here : https://github.com/r3fang/SnapTools/blob/master/docs/snap_format.docx .

Additional file 1: Figure S9:

Heatmap for the average accessibility across clusters (columns) and the marker genes (rows) that are used to calculate the RAGI metric on the 10X PBMCs dataset. **(a)** Louvain clustering solution **(b)** k-means clustering solution **(c)** hierarchical clustering (HC) clustering solution.

Additional file 1: Figure S10:

UMAP visualization of cells colored by the accessibility of marker genes: **(a)** S100A12 and **(b)** MS4A1 and **(c)** GAPDH (housekeeping gene) and on the 10X PBMCs dataset.

Additional file1: Figure S24

Running time results. **(a)** Running time, in minutes for each method applied to the *Buenrostro2018*, 10X PBMCs, and downsampled sci-ATAC-seq mouse datasets. **(b)** Running time, in minutes for each method on the simulated bone marrow dataset at a noise level of 0.2 with read coverages of 250, 500, 1000, 2500, and 5000 fragments.

Supplementary Table 1-21

4.5 Preamble

This work is the result of my stay in the Computational Biology and Data Mining Lab and the Institute of Organismic and Molecular Evolution of the Johannes Gutenberg University Mainz under the supervision of Prof. Dr. Miguel-A Andrade-Navarro. The project was conceived together with Prof. Miguel-A Andrade-Navarro and my contribution to the work is as follows: I have designed the experiment, developed the method, made the whole computational analysis, produced the figures and wrote the manuscript. Also, I have collaborated with Steffen Albrecht for the retrieval of the data from the ENCODE portal. The analysis can be reproduced following this link: https://github.com/tAndreani/IPVARIABLE.

4.6 Chapter 3

Identification of cell-specific variable regions in ChIP-seq data (Andreani T, Albrecht S. et al, *Nucleic Acids Research,* doi: 10.1093/nar/gkaa180)

Abstract

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is used to identify genome-wide DNA regions bound by proteins. Several sources of variation can affect the reproducibility of a particular ChIP-seq assay, which can lead to a misinterpretation of where the protein under investigation binds to the genome in a particular cell type. Given one ChIP-seq experiment with replicates, binding sites not observed in all the replicates will usually be interpreted as noise and discarded. However, the recent discovery of high-occupancy target (HOT) regions suggests that there are regions where binding of multiple transcription factors can be identified. To investigate these regions, we developed a reproducibility score and a method that identifies cell-specific variable regions in ChIP-seq data by integrating replicated ChIP-seq experiments for multiple protein targets on a particular cell type. Using our method, we found variable regions in human cell lines K562, GM12878, HepG2, MCF-7, and in mouse embryonic stem cells, defined as protein binding regions with non-reproducible results across replicated experiments. These variable-occupancy target (VOT) regions are CG dinucleotide rich, and show enrichment at promoters and R-loops. They overlap significantly with HOT regions, but are not blacklisted regions producing non-specific binding ChIP-seq peaks. Interestingly, among various genomic features, DNA accessibility is a better predictor of VOT than CpG islands or epigenetic marks. Our method can be useful to point to such regions along the genome in a given cell type of interest, to improve the downstream interpretative analysis before follow up experiments.

Introduction

A series of genome-wide experiments are largely adopted to study biological systems in relation to a given protein. They contribute to our understanding of particular molecular mechanisms at the basis of biological processes such as transcription and development, just to mention a few. In particular, ChIP-seq evaluates the genomic positions bound by a protein in the genome. Standard ChIP-seq experiments typically include replicated measurements in the experimental design in order to have the proper statistical power for the identification of reliable binding sites (or ChIP-seq peaks).

Previous results have indicated for several model organisms such as yeast, Drosophila and Caenorhabditis elegans the existence of genomic regions that are bound more often with respect to others, even in genomic positions in which a binding site is not expected for the protein under investigation. These regions have been previously characterized and described as “hyper-ChIPable” in yeast (1) and confirmed later in Drosophila, C. elegans and mouse and referred to as “phantom peaks” (2). Furthermore, other regions defined here as variable regions, have protein binding that tends to variate stochastically and is difficult to interpret because their inconsistency in the reproducibility of the results. Current approaches to analyse ChIP-seq experiments do not report to the users regions that misbehave before downstream interpretative analysis; this might lead to misinterpretation of the ChIP-seq results in terms of the function associated to the protein under investigation.

Here, we present a method that uses replicated ChIP-seq data for several proteins on the same cell line to detect regions that misbehave in ChIP-seq experiments. We assigned the term variable for a given genomic region if a protein binding site (or ChIP-seq peak) was not consistently detected in several experimental replicates of the same protein and for several independent proteins in a given cell type.

These assignments can increase the value of ChIP-seq experiments by categorizing certain peaks as having cell-specific variability. Possible reasons for this variation might be the adoption of variable genomic structures (3), the high expression of a nearby gene (2), the specificity of the antibody used and the conformation of the chromatin during the immunoprecipitation. By finding variable regions, we expect to be able to characterize the origins of this variability and its potential relation to biological processes.

During the last years, the ENCODE consortium (4) addressed the problem of data collection for ChIP-seq experiments as well as other sequencing datasets creating the metadata of all the experiments. This effort is praiseworthy because at the time of reusing specific datasets it is important to know in detail how the data were produced, from which laboratory and according to which experimental criteria. This information allows controlling possible confounding factors in our study that focuses on local variability potentially caused by local genomic structural conformation or activity. Thus, we used data from the ENCODE consortium, and we controlled how the

experiments were performed, from which laboratory and bioinformatics tools used for data handling among other parameters.

In this work, we took advantage of the metadata provided by the ENCODE consortia, as indicated above, to select experiments in a consistent and comparable manner to implement a sliding window approach to classify genomic regions as variable or not.

Our results show that the method can identify variable regions for every cell line tested and that, particularly for the K562 cell line, for which many datasets are currently available, it improves the separation of the samples in a PCA to promote a better downstream interpretative analysis. Method and scripts can be found online at this link: https://github.com/tAndreani/IPVARIABLE.

Results

Variable regions can be detected in all tested cell lines

Our method to detect variable regions in ChIP-seq datasets for a given cell line relies on having several proteins tested for the particular cell line with replicates coming from the same laboratory and using the same platform (see Methods for details). Currently, the number of such sets in the ENCODE database is limited. While we were able to obtain a suitable set for K562 with four proteins and three replicates each, it is more usual to have a lower number of replicates, typically two.

For this reason, we applied our method to identify variable regions using just two replicates per protein for the human cell lines K562, GM12878, HepG2, MCF-7, and for mouse ESCs (mESCs). We found that K562 cells have a higher and significant number of variable regions for a total amount of 483 (p-val = 6.3e-103; see Methods for details) whereas the other three human cell lines GM12878, HepG2, MCF-7 have similar lower but also significant numbers: 61, 76 and 62, respectively (p-val = 1.1e-07, 5.9e-6, or 4e-5, respectively) (Fig. 2A-D). This might be due to the instability of the K562 cell's genome as reported in previous publications (12). Furthermore, we used another popular cell line used for developmental studies, mouse embryonic stem cells ES-14. Also in this cell line, we identified a number of variable regions for a total amount of 332 (p-val = 6.9e-3) (Fig 2E).

Finally, we wanted to see to what extent increasing the number of replicates in the experimental design would increase the number of variable regions. We were able to design a proper experiment only for K562 cell lines, which is the cell line with a higher number of experiments in the ENCODE project. Using three replicates per protein and four proteins (see Methods for details), the number of variable regions detected was drastically higher (a total of 3012).

Variable regions are rich in CG dinucleotides and enriched along gene body features

Next, we tested the CG dinucleotide frequency of the variable regions in K562 and mESCs. We found a higher frequency of CG dinucleotides compared to a random set of control genomic regions in K562 (p val 3.8e-4) and mESCs (p val 8.4e-8) (Fig. 3A and 3B, respectively). The protein targets in the ChIP-seq experiments used do not have DNA-binding motifs particularly affine for CG dinucleotides (motifs from the Jaspar database (13); Supplementary Fig. S2), hence the CG composition is specifically related to the variable behaviour and not to the DNA-motifs bound by the proteins selected.

Furthermore, variable regions were highly enriched for different features among K562 and mESCs cells with K562 showing a high enrichment for promoters and R-loops (Fig. 3C) and mESCs showing a high enrichment for 3UTR and Promoters (Fig. 3D). In a recently published work (3), the authors reported that previously characterized regions as "high occupancy target"

(HOT) (18,19) are likely to be a ChIP-seq artefact. Among the properties of these regions, they reported GC/CpG-rich kmers and RNA–DNA hybrids (R-loops). Since also in our work we found these characteristics for the variable regions, we downloaded the regions reported in (3) and checked for a possible enrichment. We found significant enrichment for all the cell lines tested except for GM12878. Especially for the K562 and HepG2 cell lines the enrichment was of 52 and 17 fold change, respectively (observed vs expected, two-sided Fisher test p val. 1.46e-87, 3.9e-3, respectively). Furthermore, also mESCs showed a significant enrichment of 8 fold change (observed vs expected, two-sided Fisher test p val. 3.2e-3) (Fig. 3E).

Variable regions are not blacklisted regions from ENCODE project

The ENCODE consortium provides a detailed description about the meaning of “blacklisted sites”. These genomic positions often produce artefact signal in certain loci mainly because of excessive unstructured anomalous sequences. Reads mapping to them are uniquely mappable so simple mappability filters do not remove them. These regions are often found at specific types of repeats such as centromeres, telomeres and satellite repeats. Given the high variability of the ChIP-seq peaks of the regions described in this manuscript we thought to check whether our method was detecting the already described and characterized “blacklisted regions” or not. To answer this question, we analysed the overlap of the variable regions obtained in all the cell lines we have used (K562, HepG2, MCF-7, GM12878 and mESC) with the public available ENCODE blacklisted regions (20). We found no overlap except for K562 (significant depletion, 35 observed vs 261 in random model, two-sided fisher test p val. 2.92e-46) and mESC (significant depletion, 7 observed vs 29 random, p val. 2.1 e-4). These results confirm that our variable regions are not associated with the ENCODE blacklisted regions, hence need to be considered for new detection methods.

The removal of variable regions improves the interpretation of the PCA in K562 cell lines

Variable regions may reflect cell-specific effects that are not target-specific. While this information might be indicating biological function, we hypothesized that the removal of such target non-specific data could result in an improvement of the separation of the replicates points in a PCA. In order to test such potential benefit in removing the variable regions for downstream interpretative analysis, we performed a PCA of the ChIP-seq samples obtained for the K562 cell line (Fig. 4A). The PCA was performed using (i) all the segments bound by each protein in the respective replicates in the original datasets and (ii) without the segments within the variable regions. We found that the separation along the components improves after removing the segments within the variable regions. Furthermore, the replicates of the proteins tend to cluster better without the segments within the variable regions

and this is reflected with a lower Euclidean distance in pairwise comparisons between replicates of the same proteins (Supplementary Fig. S3). We note that this does not mean that data from these regions should be discarded, but that they should be considered differently. Further research is needed to characterize these regions and find out if they have a cell-specific biological function.

DNA accessible regions are predictive of the variable behaviour in K562 and mESCs

Finally, we searched for genomic features that can be predictive of variable behaviour. We evaluated the possible association of different genomic features with variable regions using a random forest classifier. The classifier (random forest) was trained with a positive set consisting of the variable regions detected in mESCs (332 variable regions), and with a negative set consisting of genomic sequences with size and nucleotide composition similar to those of the positive set (see Methods for details about the training and about the set of genomic features).

The algorithm was able to classify the variable regions (area under ROC curve = 0.82) and returned as best predictors DNA accessible regions, together with regions lowly methylated and oxidative products of TET enzymes 5hmC and 5caC (Fig. 5A and 5B). These modifications are highly frequent at distal regulatory elements (8) and promoters (14) and we speculate that the turnover of these modifications might affect the binding of the proteins to the DNA leading to stochastic variation of the binding sites.

To observe the reproducibility of these results, we studied next data from K562 cells (483 variable regions). Again, as best predictor of variable regions, we found DNA accessible regions, together with K3K4me1 and H3K27ac chromatin marks (area under the ROC curve = 0.97) (Fig. 5C and 5D).

Material and Methods

Collection of ChIP-seq data

The ENCODE data portal provides comprehensive information about the meta-data of each experiment generated by the ENCODE consortium. We selected experiments according to specific parameters in order to avoid unwanted variability and to maintain consistency on the parameters of the downloaded data. The experiments were selected according to the following

criteria: (i) laboratory producing the data as Snyder, (ii) identical untreated isogenic human cell lines (K562, MCF-7, GM12878, and HepG2) and ES-E14 mouse embryonic stem cells (mESCs), (iii) data processed with the standard ENCODE pipeline that uses the optimal IDR threshold as statistical method to obtain the significant peaks (5) , (iv) status as released corresponding to a possible usage of the data, (v) the experiments of each biological replicate correspond to a peak file compared with appropriate input control experiment and (vi) peaks significance selected with a false discovery rate (FDR) lower than or equal to 5%. The metadata presented in JSON format was extracted and stored in a relational SQL database (see Supplementary File 1, Fig. S1). For every cell we selected the following targets: for HepG2 we used MAFK, MNT, TBX3 and ZNF24 with two biological replicates; for MCF-7 we used CREB1, CLOCK, NFIB and ZNF512B with two biological replicates; for GM12878 we used BHLHE40, EP300, IKFZ2 and ZNF143 with two biological replicates; for K562 we used ARNT, NCOR1, MNT and ZNF24 with three biological replicates; for ES-14 mESCs we used HCFC1, MAFK, ZC3H11A and ZNF384 with two biological replicates (see Supplementary Table S1 for details).

Reproducibility score implementation

After the identification of suitable experiments, the genome is binned in consecutive segments of 200 base pairs (bp) and the experimental ChIP-seq peaks are mapped to each segment. We formalized the reproducibility and not reproducibility of the segments for a given protein as illustrated in Fig. 1A and as follows:

Let S be the genomic segments for a given genome;

Let N be the number of replicate ChIP-seq experiments for a given protein;

For each segment in S;

Let P be the number of peaks detected in the segment;

Reproducibility score = NA if P = 0;

Else Reproducibility score = 1 if the segment itself or one of its neighbours* has P=N;

Otherwise Reproducibility score = 0

* Neighbours are all consecutive segments with P > 0

In the following paragraph, we explain the procedure described by the pseudocode above in words. For our study, segments of the genome are defined considering a window size of 200 base pairs, N represents the number

of replicates for each protein under investigation in a given cell type, and P is the number of replicated ChIP-seq peaks detected in a genomic segment (the signal). Consecutive segments without any signal (P=0; no peaks) are assigned with a NA. Consecutive segments in between two NA segments with a signal P reaching as a maximum value N are considered as reproducible regions and assigned a value of 1. On the contrary, consecutive segments in between two NA segments reaching a maximum value lower than N are considered as variable regions and assigned a value of 0 (Fig. 1A). The results of each protein under investigation are aggregated in a Reproducibility Score Matrix (RSM) (Fig. 1B) where rows show segments and columns show their reproducibility score for each protein and a final score (FS) defined as the average value of the row (or NA if more than 1 reproducibility score equals NA).

Statistical test of scored regions

To assess whether the number of reproducible or variable regions associated with a particular score is significant, a suitable control had to be identified. The appropriate null distribution was built by randomizing the RSM. We performed this task using the "sample" function in R. The randomization was performed 1000 times, and regions at particular scores (0, 0.25, 0.33, 0.5, 0.66, 0.75 and 1) of the null distribution were counted. Afterwards, a z-score was computed according to this formula:

$$z - score = \frac{\delta - \delta rand}{\theta \delta rand}$$

where δ is the number of regions observed with a particular score, and δ rand and $\sigma\delta$ *rand* are the mean value and the standard deviation of the null distribution, respectively. Assuming normality of the null distribution, it is possible to analytically calculate the corresponding p-value for a given z-score with significance level $\alpha = 0.05$. The regions for each particular score were subjected to the test.

Principal Component Analysis and Euclidean distances

Principal Component Analysis (PCA) was performed using the Python package scikit-learn version 0.19.1. The dots represented in the PCA are biological replicates for a given protein. Each colour represents a specific protein and the features set used to perform the PCA are all the segments detected in all the proteins. In order to test the effect of the removal of the variable regions in the PCA, segments within the variable regions were

removed from the features set. The similarity distances between replicates of the same protein in the PCA were computed with the Python package SciPy version 0.19.1 using as a metric the Euclidean distance. Boxplot and Dotplot were performed using the Python library Matplotlib version 2.2.2.

Enrichment analysis at regulatory elements

We collected genomic coordinates of the following gene-related features from the UCSC table browser database in hg19 and mm10 annotations: promoter, 5UTR, coding exons, intron and 3UTR. Furthermore, we also used regions with R-loops (6) since they were previously reported as a potential feature associated with misbehaving ChIP peaks (3). For a set of regions (e.g. variable), the enrichment for each feature is obtained by dividing the number of regions overlapping a regulatory feature by the number of randomized regions overlapping the same feature.

$$Enrichment = \frac{Variable\ Regions \cap Feature}{Simulated\ Regions \cap Feature}$$

Randomized regions were obtained using the command shuffleBed in bedtools version v2.25.0 (7).

CG and AT dinucleotide frequency calculation

The percentage of CG and AT dinucleotides in the mouse and human genomes was calculated with the nuc function in the bedtools version v2.25.0 toolkit. To compute the CG and AT enrichment in the variable regions of mESCs and K562 cells we used a set of control regions using the shuffleBed function in bedtools version v2.25.0 (7). The differences in dinucleotide composition between the variable regions and the set of control regions were tested for significance using a t-test.

Prediction of variable regions in K562 and mESCs using genomic features

In order to find out whether different genomic features (active chromatin marks, repressive chromatin marks, DNA accessible regions, CpG islands,

etc.) could be used by a random forest classifier to predict variable regions in mESCs or in K562 cells we used a large panel of datasets. For mouse ESCs, we also considered regions undergoing TET oxidation and bivalent domains. We used the following published data: 5hmC, 5fC and 5caC (8); CpG islands extracted from UCSC table browser for mm10 annotation; H3K4me1, H4K4me3, H3K79me2, H3K27ac, H3K27me3, H3K36me3 (9), LMR (10), DNAse-seq from ID:ENCSR000CMW experiment in the ENCODE portal. mm9 genome features were converted to mm10 using the Batch Coordinate Conversion (liftOver) tool from the UCSC Genome Browser Utilities (https://genome.ucsc.edu/cgi-bin/hgLiftOver). For K562 cells, we used the following data downloaded from the ENCODE data portal: H3K27ac (ID:ENCFF044JNJ), H3K27me3 (ID: ENCFF145UOC), H3K4me1 (ID: ENCFF183UQD), H3K4me3 (ID:ENCFF261REY), H3K79me3 (ID: ENCFF350GQM), H3K36me3 (ID: ENCFF537EUG), DNAse-seq (ID: ENCFF856MFN). CpG islands were extracted from UCSC table browser for hg19 annotation.

To train and test the random forest model we used the function randomForest from the R package randomForest version 4.6-14 (DOI:10.1023/A:1010933404324). As a positive set we have used the variable regions estimated by our method and as a negative set, we used a set of regions obtained with the package gkmSVM version 2.0 (11). This package has a function named genNullSeqs capable of using the positive set of sequences and learning their nucleotide composition. Subsequently, the function generates a set of genomic locations with sequences of nucleotide composition and length similar to those in the positive test set. This approach was successfully used previously for the prediction of double-strand breaks at CTCF and accessible chromatin sites (17). The train and test set were obtained using the function sample.split in the R package caTools version 1.17. The function is used to split the data used during classification into train and test subsets. We decided to split train and test in a 30:70 ratio. We also split the dataset using other ratios such as 10:90, 20:80 and 40:60 and this did not change the performance of the prediction. The evaluation of the prediction was performed by computing the ROC curve using the package pROC version 1.13.0 (https://cran.r-project.org/web/packages/pROC/pROC.pdf).

DISCUSSION

During the latest years, several laboratories tried to study the regulation of gene expression in different model organisms. For this scope, ChIP-seq was adopted as a standard technique but the extent of its usage raised some questions in terms of reliability (1,2,15,16). In particular, in a previous work (3), Wreczycka and colleagues presented a method that considers the nature of phantom peaks and hyper-ChIPable regions to define high-occupancy target (HOT) regions where un-specific binding to multiple targets would be found even in the absence of expected binding motifs. They concluded that the un-specificity of binding sites in HOT regions is associated with CG dinucleotide rich regions

and as a consequence at R-loops (that are CG rich) and DNA tertiary structures. Though, this is a common concern for ChIP-seq assays and since the beginning, the technique was known to be biased toward GC-rich contents during fragment selection in the steps of the library preparation and amplification during the sequencing (15). Here we have found evidence that supports that such regions could also be responsible for variable behaviour in ChIP-Seq in a cell-specific fashion. Our method evaluates replicated ChIP-seq experiments for multiple targets in a cell type, to find regions where target binding is not reproduced in all replicates for multiple targets. These variable-occupancy target (VOT) regions are cell-specific and share structural features with HOT regions. However, differently to HOTs, VOTs do not produce consistent un-specific target binding. Accordingly, VOTs do not overlap blacklisted ENCODE regions. Together, the cell-specificity of VOTs and our finding that VOTs can be predicted using DNA accessibility suggest their dependency on gene expression and epigenetic state.

While the most consistent enrichment of variable regions observed was for promoters and 5UTR regions in both K562 cells and mESCs, the differences observed for variable regions at R-loops suggests that it is not possible to drive a certain conclusion relating where exactly this variation occurs. On the other hand, the fact that we are able to predict variable regions using genomic features alone with relatively high accuracy indicates that there is certainly a relationship between genomic features and variability that could be eventually detected. Taking these results together, we assume that with the future availability of further ChIP-Seq datasets testing multiple proteins in the same cell lines it will be possible to assess the sources of variability in ChIP-Seq with more certainty.

With our method, we propose a systematic approach using ChIP-seq experiments and replicated measurements in a given cell-type to identify misbehaving DNA regions that have to be treated differently in the post-processing downstream analysis. We have shown that discarding data from these regions can improve studies focusing on target-specific effects. However, further research is needed to study potential cell-specific functions of VOTs as hubs or sponges for transcriptional regulatory complexes, which could be verified with other experimental assays like ChIP-qPCR. We suggest applying our approach as a post-processing quality check of the data before

starting to follow up experiments and driving biological conclusions. We must point out that our method requires enough replicates for the same protein in a given tissue. For example, predictions for organisms like the fly *Drosophila melanogaster* using data from modENCODE and modERN are not yet possible, At this point, the only datasets we find suitable for our analyses are in ENCODE.

Finally, we note that similar approaches to the one used for our method to point to variable regions in ChIP-seq datasets could be eventually developed and applied to any type of next-generation sequencing datasets that uses replicated measurements under various conditions (for example, ATAC-seq from multiple cell types). This could open avenues for the discovery of other types of variability leading to a more informed use of sequencing-based data. The study of the similarities and differences between variable regions obtained with different techniques might be crucial to increase our understanding of the inter-relation between genomic structural flexibility and regulatory function.

REFERENCES

1. Teytelman, L., Thurtle, D. M., Rine, J., & van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proceedings of the National Academy of Sciences , 110 (46), 18602–18607.

2. Jain, D., Baldi, S., Zabel, A., Straub, T., & Becker, P. B. (2015). Active promoters give rise to false positive “Phantom Peaks” in ChIP-seq experiments. Nucleic Acids Research 43 (14), 6959–6968.

3. Wreczycka K., F. Vedran, U. Bora, R. Wurmus, S. Bulut, B. Tursun, A. Akalin (2019). HOT or not: examining the basis of high-occupancy target regions, Nucleic Acids Research, 47(11), 5735-5745.

4. Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., & Chen, Y. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research 22 (9), 1813-1831.

5. Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. Annals of Applied Statistics, 5 (3), 1752–1779.

6. Sanz, L. A., Hartono, S. R., Lim, Y. W., Steyaert, S., Rajpurkar, A., Ginno, P. A. & Chédin, F. (2016). Prevalent, dynamic, and conserved R-loop structures associated with specific epigenomic signatures in mammals. Molecular cell, 63(1), 167-178.

7. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26 (6), 841–842.

8. Shen, L., Wu, H., Diep, D., Yamaguchi, S., D’Alessio, A. C., Fung, H. L., ... & Zhang, Y. (2013). Genome-wide analysis reveals TET-and TDG-dependent 5-methylcytosine oxidation dynamics. Cell, 153 (3), 692-706.

9. Whyte, W. A., Bilodeau, S., Orlando, D. A., Hoke, H. A., Frampton, G. M., Foster, C. T., ... & Young, R. A. (2012). Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature, 482 (7384), 221.

10. Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., ...& Tiwari, V. K. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480 (7378), 490.

11. Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., & Beer, M. A. (2016). gkmSVM: an R package for gapped-kmer SVM. Bioinformatics, 32 (14), 2205-2207.

12. Ramachandran, P., Palidwor, G. A., & Perkins, T. J. (2015). BIDCHIPS: bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates. Epigenetics & chromatin, 8 (1), 33.

13. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., .. & Baranasic, D. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic acids research, 46 (D1), D260-D266.

14. Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., ...& Oliviero, S. (2015). Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA methylation dynamics. Cell reports, 10 (5), 674-683.

15. Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nature reviews genetics, 10 (10), 669.

16. Park, D., Lee, Y., Bhupindersingh, G., & Iyer, V. R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. PloS one 8(12), e83506.

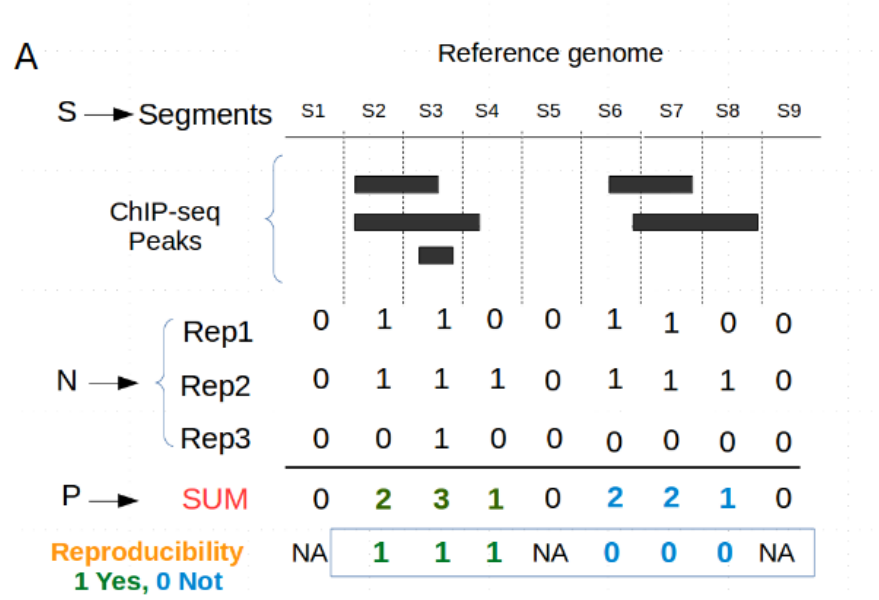
17. Mourad, Raphaël, et al. "Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution."Genome biology 19.1 (2018): 34.

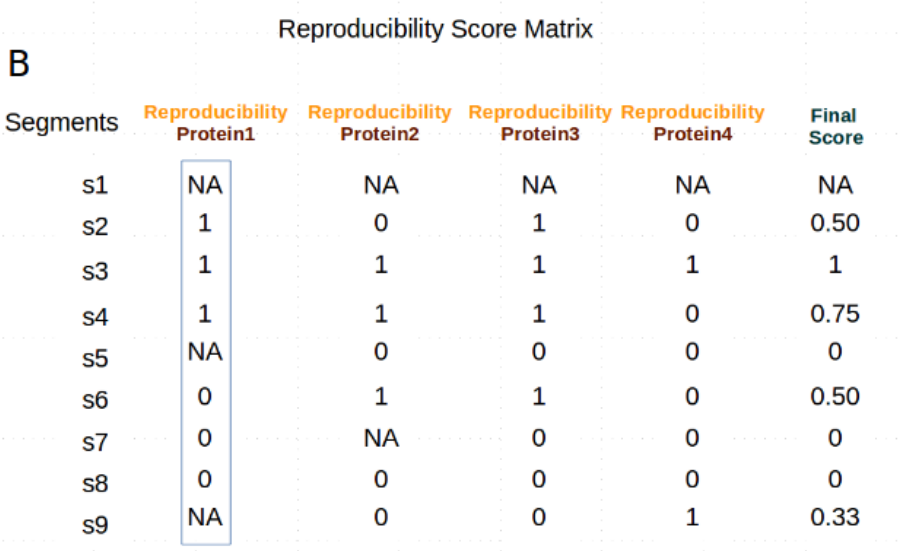
18. Xie D., Dan X., Boyle A.P., Linfeng W., Jie Z., Trupti K. et al. . Dynamic trans-acting factor colocalization in human cells. Cell. 2013; 155:713–724.

19. Boyle A.P., Araya C.L., Cathleen B., Philip C., Chao C., Yong C., Gardner K., Hillier L.W., Janette J., Jiang L. et al. . Comparative analysis of regulatory information and circuits across distant species. Nature. 2014; 512:453–456.

20. Haley M. Amemiya, Anshul Kundaje and Alan P. Boyle. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific Reports 9.1 (2019): 9354.

Figures





Segments	Reproducibility Protein1	Reproducibility Protein2	Reproducibility Protein3	Reproducibility Protein4	Final Score
s1	NA	NA	NA	NA	NA
s2	1	0	1	0	0.50
s3	1	1	1	1	1
s4	1	1	1	0	0.75
s5	NA	0	0	0	0
s6	0	1	1	0	0.50
s7	0	NA	0	0	0
s8	0	0	0	0	0
s9	NA	0	0	1	0.33

Figure 1. Method to annotate genomic regions with a reproducibility score. (A) From the genomic segments S, segments with ChIP-seq peaks for a given protein 1 in a given cell, in N=3 replicates, are converted to a binary format (Rep1 to Rep3). The sum at each segment of the values for the replicates (SUM or P) allows to define blocks of consecutive segments between zero-scored segments devoid of peaks (here two blocks; green and blue). All segments in a block are identified as indicating a reproducible region (Reproducibility=1; green) if the block holds at least one segment with value 3. Otherwise they are given a value indicating a non-reproducible region (Reproducibility=0; blue). (B) Average reproducibility values for ChIP-seq experiments from four different proteins in cell type A are combined in a final

score that ranges from 0 (not reproduced in the four proteins) to 1 (reproduced in the four proteins). Only segments with values for at least 3 proteins were considered.

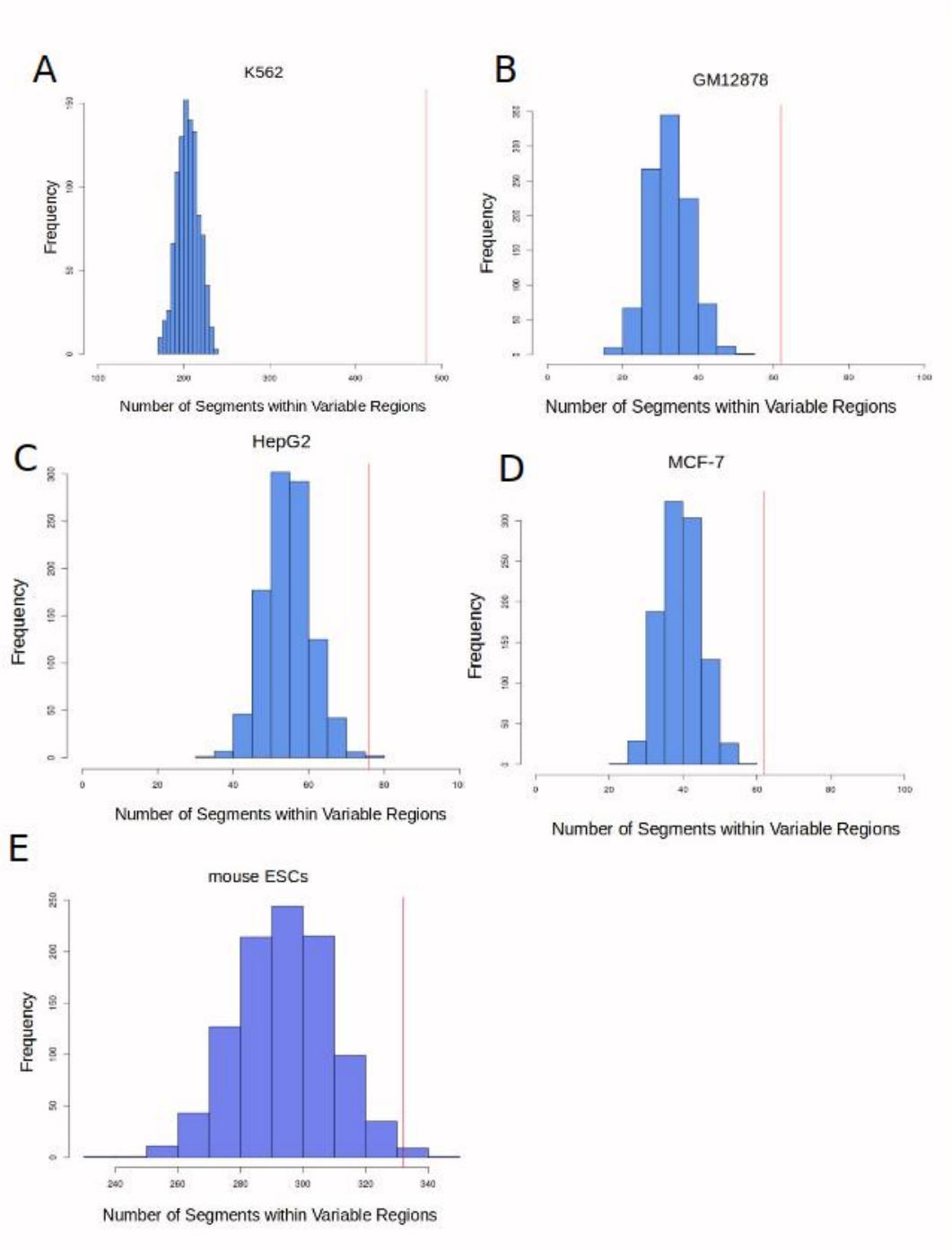


Figure 2. Significance of the observed number of variable regions for different cell types. The number of variable regions observed in each cell line (red line) is significantly higher than the corresponding computed null distribution (blue). (A) K562, (B) GM12878, (C) HepG2, (D) MCF-7 and (E) mouse ESCs cells;

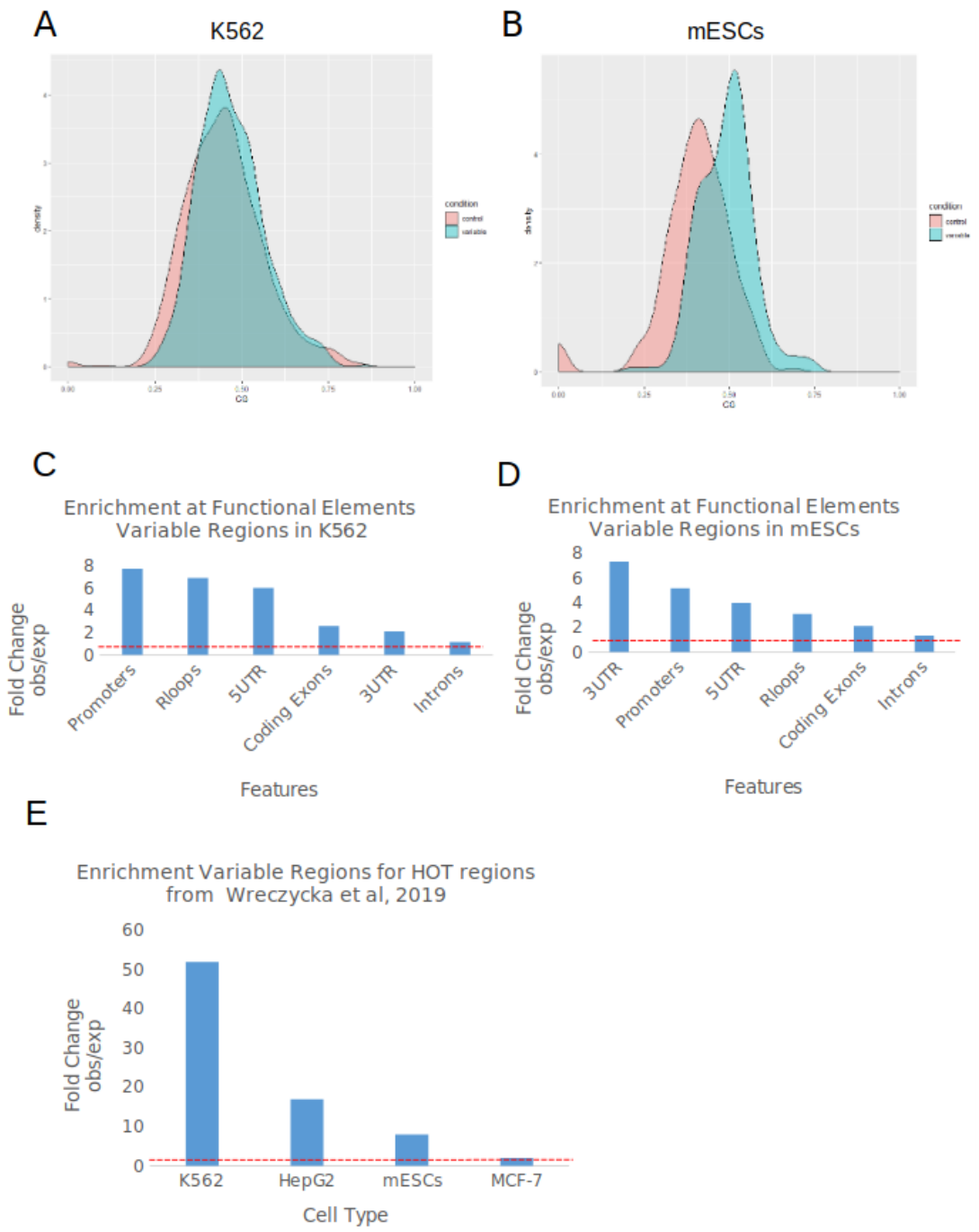


Figure 3. Properties of variable regions in K562 and mESCs. CG dinucleotide composition in K562 human cell lines (A) and in mouse ESCs (B). Control regions are a set of randomly sampled genomic regions of similar size. Enrichment of variable regions in K562 human cell lines (C) and mouse ESCs (D) at several gene body features and R-loops. (E) Enrichment of variable regions at HOT regions (3).

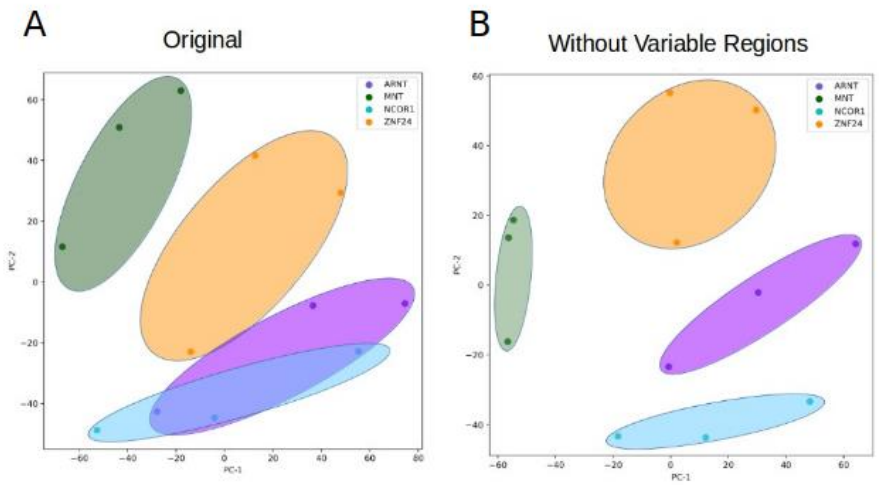


Figure 4. Effect of removing variable regions. PCA using presence of peaks in the set of 200 bp segments on the genome as features with the original data and without 200 bp segments within the variable regions in K562 cell lines (A and B, respectively). Each dot represents a biological replicate and each colour the protein target.

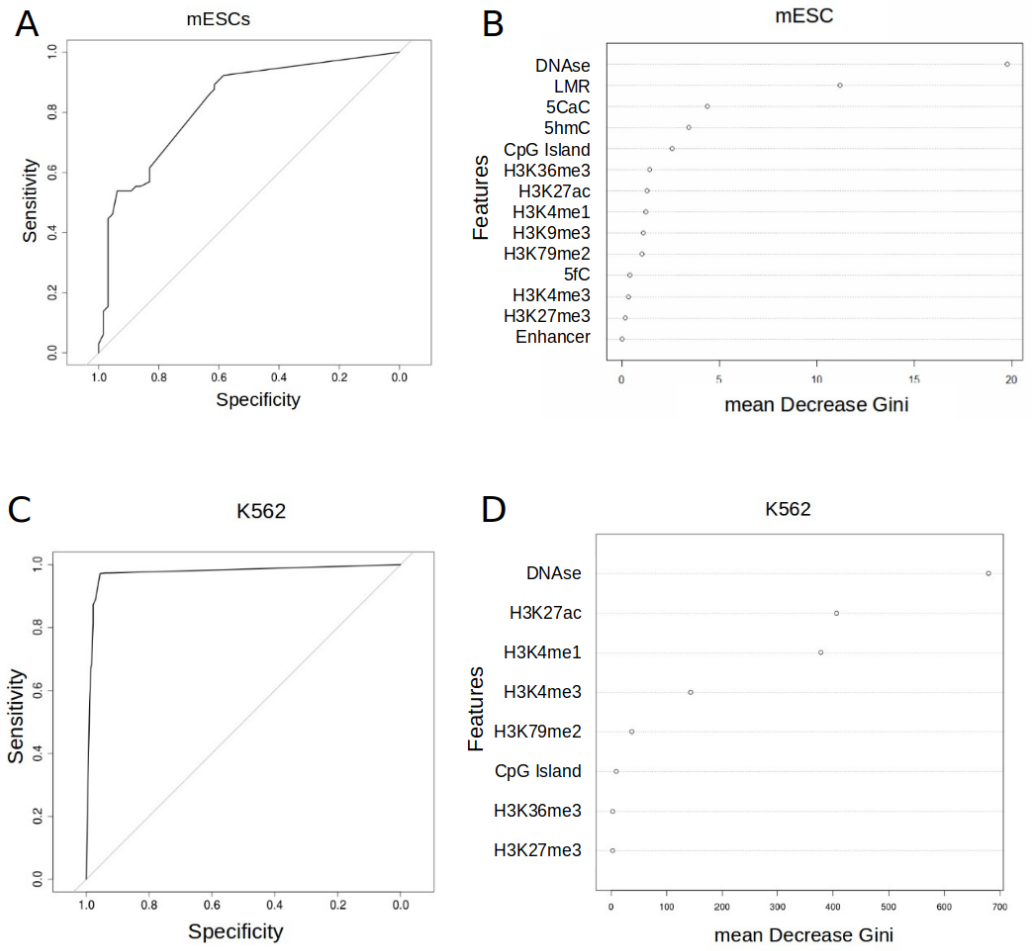


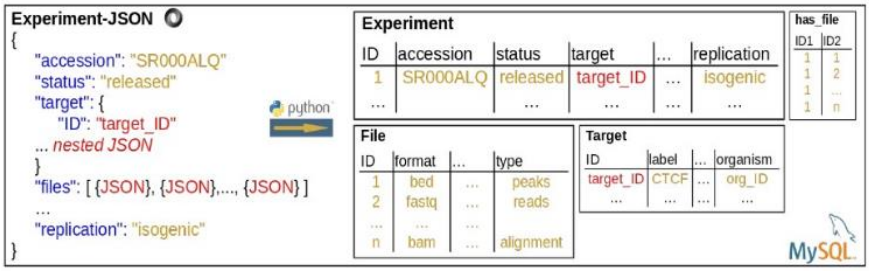
Figure 5. Prediction of variable regions in K562 and mESCs. (A) Receiver operating characteristic (ROC) as a quality measure of the predictability of the variable regions in mESCs and (B) importance of the features for predicting the variable regions in mESC measured as mean decrease Gini in the random forest. See text for details about every feature. (C) Receiver operating characteristic (ROC) as a quality measure of the predictability of the variable regions in K562 cells and (D) importance of the features for predicting the variable regions in K562 cells measured as mean decrease Gini in the random forest.

Supplemental information

Conversion of JSON-based files to a relational SQL database

The metadata in ENCODE is represented in JSON-format. There is an API (Application Program Interface) that enables downloading JSON files for many experiments automatically. In addition, it is possible to restrict the experiments to be downloaded by specifying conditions the experimental metadata has to comply with. However, a JSON file for one experiment can include up to 11,000 lines and this can cause two problems. The first is that the extraction of information can be complicated and requires the further implementation of small scripts / programs. The second is that parsing such files can be computationally expensive especially if there are more than 1000 files to be processed. To avoid these problems and enabling the analysis of the metadata in a flexible and fast way we decided to convert the information from JSON into a MySQL database. For objects like experiment, target, organism, replicate, etc. we downloaded all JSON files. A Python script was implemented that extracts information encoded in each JSON file and automatically fills the SQL database with this information. The procedure in the Python script takes also cares about creating the tables for each object and furthermore creating also relational tables for the connection between two tables (objects). Relational information is also automatically extracted and stored into SQL. Figure 1 represents a simplified case with one experiment. The brackets ”{}” define the beginning and end of a JSON. There are simple features like accession and status for which the value is directly given. Those simple values can also appear in lists represented by the brackets ”[]”. However, one JSON can include other JSONs. This is shown by the target in the experiment that is completely described by a so called nested JSON. It is also possible that one JSON includes a list of JSONs. The example in Figure S1 shows that for each file that is related to the experiment there is one JSON within the experiment. These nested JSONs are the reason why those files are getting so large. Another advantage coming with the conversion from JSON into SQL is that objects are not stored redundantly. Taking as example the target that describes the protein that is targeted in a ChIP-seq experiment. All JSONs for experiments with the same target store the same information for the same target multiple times. In SQL there is only one entry for a specific target in the target table and experiments with this target are just linked to this entry by the target_ID. An example for such a link is represented by the target in Supplementary Figure S1 below and has_file is a relational table.

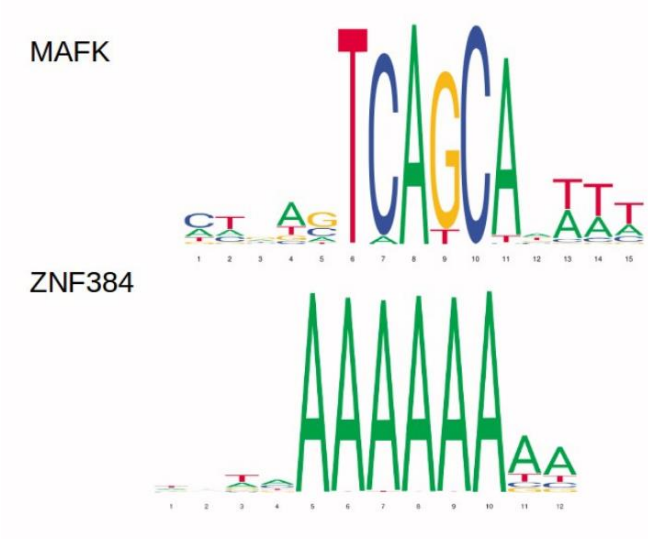
Figure S1



1) Same Cell Isogenic Line (K562)
2) Same condition/treatment
3) Same Laboratory (Snyder)
4) Same Bioinformatics ENCODE Pipeline
5) Same Statistical Test (IDR threshold)

Figure S2

(A) DNA binding motifs of the protein targets of the ChIP-seq samples used to detect variable regions in mouse ESCs from the Jaspar database. No data for ZC3H11A and HCFC1.



(B) DNA binding motifs of the protein targets of the ChIP-seq samples used to detect variable regions in K562 cell lines

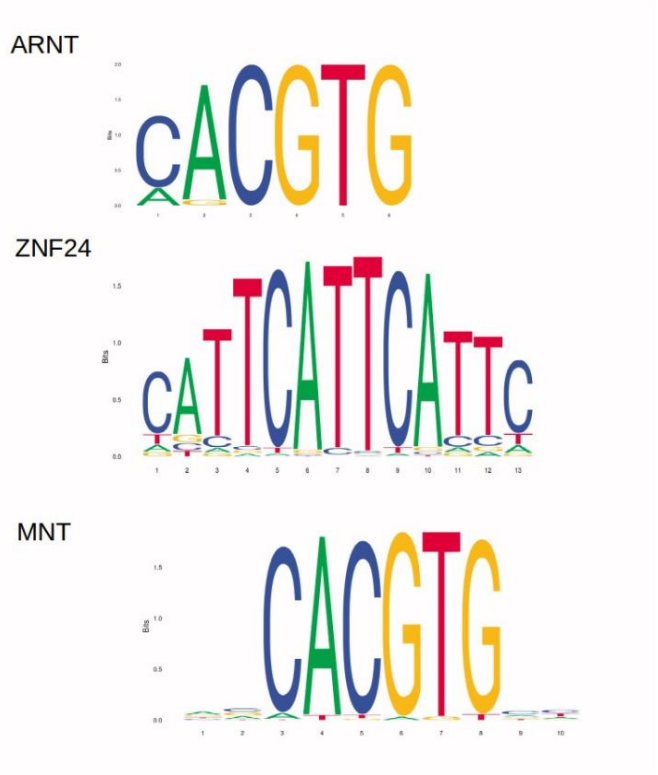
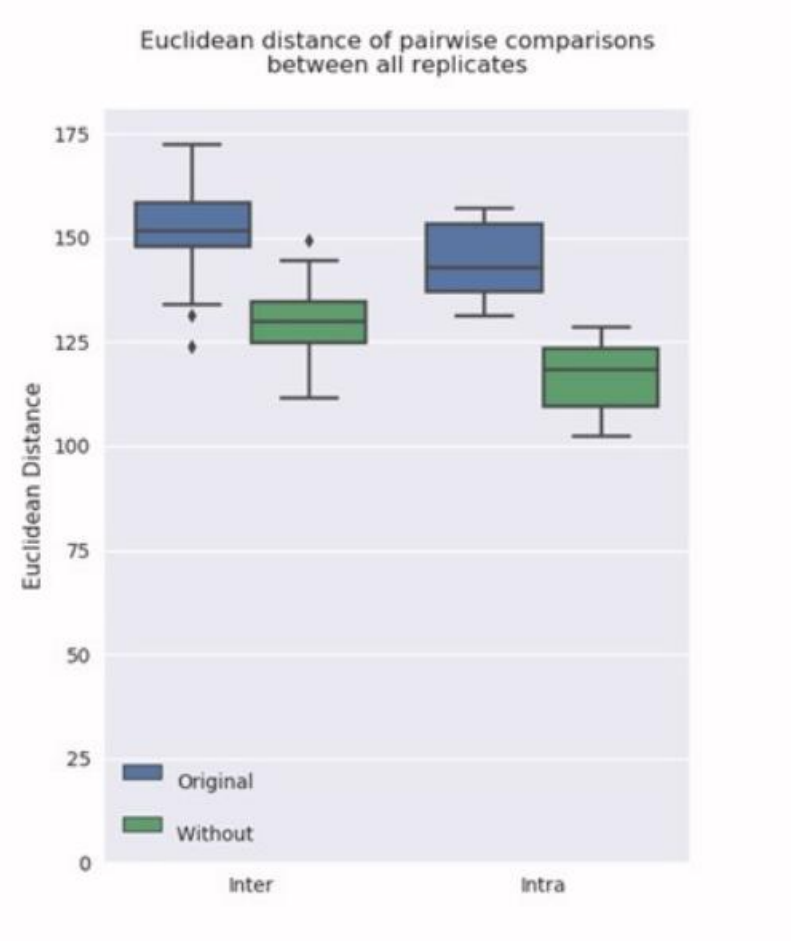

ARNT
ZNF24
MNT

Figure S3

Pairwise comparisons of Euclidean distances between replicates of different proteins (inter) and within replicates of the same proteins (intra) in the original dataset and after removing the segments within the variable regions for the K562 ChIP-seq dataset (four proteins with three replicates).



5. General Discussions

One of the most heavily researched fields in biology today is exploring how genes are regulated and which mechanisms are responsible for their physiological expression. The understanding of different genetic and epigenetic components involved in the activation or repression of specific genes is important for understanding genetically-linked diseases. Thus, uncovering the interrelationship between regulatory elements and DNA modifications that regulate gene expression as well as detecting unwanted sources of variation in sequencing data will be crucial to promote disease intervention.

The dynamics of DNA methylation and demethylation at promoters and other regulatory elements such as enhancers (Stadler et al. 2011, Shen et al. 2013), has attracted attention in the epigenetic field to elucidate the early steps of embryonic development. The advancement in cell cultures has revealed mESCs as a good model to study the early steps of embryonic development. In fact, the possibility of mESCs to become any cell type of an organism and also transiently differentiate into “2C-like” cells, which resemble the 2-cell stage of embryonic development, has allowed me to investigate the potential role of GADD45 proteins in promoting active DNA demethylation. The role of enhancers in mESC differentiation (White et al. 2012, Greenberg et al. 2019, Battle et al. 2019) and, in particular, the role of DNA methylation at enhancers in mESCs (Lu et al. 2014) and during early development (Jadhav et al. 2019) has motivated me to investigate the role of GADD45 proteins in promoting active DNA demethylation at these elements using mESCs. In addition, the availability of other sources of data, such as the annotation of regulatory elements and marker genes that are specifically expressed at the 2C-like stage, allowed me to integrate them into my analysis and ask whether demethylated sites by GADD45 proteins are located at enhancers in the proximity of 2C-like state-specific genes.

The methylome analysis of wild type mESCs contrasted with Gadd45 triple-knockouts (TKOs) revealed that the bimodal distribution pattern of the Gadd45-TKO samples is skewed toward hypermethylation (Fig. 2A in chapter 4.2), confirming the role of GADD45 proteins in promoting DNA demethylation. Furthermore, the enrichment of the hyper-differentially methylated regions (hyper-DMRs) of Gadd45-TKOs for the hyper-DMRs in Tet-TKOs in mESCs (Fig. 2B in chapter 4.2) and the oxidative intermediates such as 5fC and 5caC observed in the Tdg knockout mESCs (Fig. 2B in chapter 4.2, Fig. 2C, Fig. 2D in chapter 4.2) revealed a role of GADD45 proteins in promoting active DNA demethylation whereby the removal of these modifications is TET and TDG dependent. Parallel to my analysis, it was discovered that down-regulated genes in Gadd45-TKO mESCs were significantly associated with 2C-like specific genes and hence uncovered a potential role of Gadd45 in promoting the 2C-like state by modulating gene expression (Fig. 4A and 4B in chapter 4.2). This has motivated me to ask

whether the hyper-DMRs in Gadd45-TKOs were also associated with 2C-like specific genes. Surprisingly, my analysis showed a significant association between genes in the vicinity of the hyper-DMRs in Gadd45-TKO and 2C-like specific genes (Fig. 4C in chapter 4.2). Finally, motif analysis of the hyper-DMRs in Gadd45-TKO mESCs showed enrichment for a 2-cell stage marker named Zscan4c (Fig. 3A in chapter 4.2), further emphasizing a potential role of GADD45 proteins in regulating the 2C-like state of mESCs. These computational analyses could be integrated with additional experimental assays, in which the deletion of the hyper-DMRs located in Gadd45-TKO near the 2C-like genes should theoretically compromise the promotion of the 2C-like state in mESCs.

The investigation of potential regulatory roles of enhancers in mESCs together with the underlying mechanism of DNA methylation and demethylation turnover has motivated me to explore novel computational methods and algorithms designed to process ATAC-seq data (Buenrostro 2018). This sequencing technology has been applied to map the enhancers and open chromatin regions in whole organisms such as *Caenorhabditis elegans* (Daugherty et al. 2017) and *Drosophila (*Bozek et al. 2019). Furthermore, this sequencing technique has been improved at the single-cell resolution in order to map enhancers of specific tissues of various mouse organs and *Drosophila* developmental stages (Cusanovich 2015, Cusanovich 2018 and Cusanovich 2018). Cell-specific enhancers belonging to the three main branches of the blood lineages (red blood cells, lymphocytes, and myeloid) (Buenrostro et al., 2018) as well as to a population of cells called Peripheral Blood Mononuclear Cells (PBMCs) were also identified by applying scATAC-seq. Thus, the benchmarking of the different available methods to process and analyze scATAC-seq datasets is of primary importance to guide people that want to use this technology into which computational method is more suitable for the analysis of a specific dataset.

In this thesis, I have examined the performance of all available methods of analyzing different scATAC-seq datasets selected based on the size (number of single cells sequenced) and whether the cells had labels or not. Furthermore, since scATAC-seq is able to detect DNA sequences and not a fixed number of genes like in single-cell RNA-seq, the featurization strategy implemented on the different datasets was evaluated based on the capability (or lack thereof) of separating the different populations of cells with various coverages and noise levels.

In these experiments, methods such as *Cusanovich2018,* cisTopic, and SanpATAC were able to clearly separate the differentiated bone marrow cell types at a noise level of 0.4 and a coverage of 2500 fragments (Fig. 3-D, in chapter 4.4), with the highest evaluation score for all the metrics used such as AMI, ARI, and homogeneity score (Fig. 3B, in chapter 4.4). For datasets with labels, such as *Buenrostro2018*, methods like *Cusanovich2018,* cisTopic, and chromVAR-kmers were among the top-performing in separating the cells into their respective groups (Fig. 4D, in chapter 4.4), which was also

confirmed by the highest evaluation score for all the metrics used such as AMI, ARI, homogeneity score (Fig. 4B, in chapter 4.4). For datasets without labels (such as the PBMCs), SnapATAC, chromVAR, and cisTopic were the computational methods capable of clearly separating the cells into groups (Fig. 5C, in chapter 4.4) and this was confirmed also by my proposed metric (Residual Average Gini Index [RAGI]), in which they show the highest RAGI score (Fig. 5A, in chapter 4.4). Finally, another labelled dataset was investigated on sub-sampled scATAC-seq data of 13 adult mouse tissues (Fig. 6, in chapter 4.4). For this dataset, *Cusanovich2018*, chromVAR and cisTopic clearly separated the cells into different clusters, each one corresponding to a specific tissue (Fig. 6D, in chapter 4.4), and this observation was confirmed with the highest score for all the metrics used, such as AMI, ARI, homogeneity score (Fig. 6B, in chapter 4.4). A subset of data was used because most of the methods failed in processing the entire dataset composed of approximately 80000 single cells. In fact, the only method capable of processing this dataset was SnapATAC (Supplemental Note 5, in chapter 4.4), which uses HDF5 data structure technology to allow core computation.

Lastly, with the aim of determining which method is capable of clearly separating cells into corresponding groups within a reasonable amount of time, we decided to compare the execution time of each method with its performance in separating the single cells into groups. These groups were based on the average rank of the evaluation scores (AMI, ARI, Homogeneity, and RAGI) obtained for all the methods applied to *Buenrostro2018*, PBMCs, and down-sampled scATAC-seq mouse cell datasets (Fi. 7C, in chapter 4.4). It can be concluded from the plot that the SnapATAC, cisTopic, and *Cusanovich2018* methods are the best compromise in terms of execution time and capability of separating cells into the different cell types.

Open chromatin regions are hot spots for transcription factors (TFs) that bind to the DNA in order to regulate the expression of nearby genes. Given the wealth of data provided by the ENCODE consortia for ChIP-seq experiments, the use of such datasets can be very useful for the scientific community when the information within them is combined with other analyses to address a particular biological question. However, the reliability of the ChIP-seq peaks for a given cell type of interest can be compromised by the presence of artefacts or by highly variable regions that show inconsistent results.

Several lines of evidence have reported that for some DNA regions, such as the ”blacklisted regions”, it is more likely to have false-positive results in the form of statistically significant ChIP-seq peaks with high fold change (Amemiya et al. 2019). Furthermore, other regions, named high occupancy target (HOT) regions, are located at R-loops and promoters of highly transcribed genes and show statistically significant peaks even when the protein under investigation has been removed from the genome (Wreczycka et al. 2019, Jain et al. 2015).

In order to better understand these regions, I have developed a computational method that is able to detect variable regions in ChIP-seq results in a cell type-specific fashion. The algorithm takes in a transcription factor, each one with at least two replicates, and counts how many times in the genome the same region consistently shows a peak for all the replicates. In case peaks are observed in all of the replicates, the method assigns a reproducibility value of 1 to the region; contrarily, a reproducibility value of 0 is assigned if the peak is not observed in all the replicates (Fig. 1A, in chapter 4.6). After aggregating these values for several transcription factors, regions with a final score of 0 will be considered “variable regions” (Fig. 1B in chapter 4.6) and tested for significance by examining whether they appear randomly along the genome or not (Fig. 2 A to E in chapter 4.6). As shown in Fig. 2A-E, all the cell lines tested have a significant result, so I decided to take these regions and perform downstream analyses.

First, in order to check the validity of my method, I overlapped the variable regions obtained with the HOT regions published by Wreczycka et al. (2019) and found a strong enrichment for most of the cell lines tested, especially K562 and HepG2, in which the enrichment was of 52 and 17 fold change, respectively (observed vs expected, two-sided Fisher test p val. 1.46e-87, 3.9e-3, respectively) (Fig. 3 E in chapter 4.6). Second, given the validity of the method, I checked whether the removal of the variable regions can improve the separation of the samples in a PCA. As reported (Fig. 4 A in chapter 4.6), the removal of such regions improves the separation of the samples and this result can be interpreted as an improvement in the specificity of the peaks for downstream analysis of the TFs investigated. Finally, I searched for particular genomic features that can be predictive of the variable regions in K562 and mESCs. These two cell types showed the highest amount of variable regions and have several publicly available datasets that can be used to predict the variable regions. The random forest algorithm shows a good performance in predicting the variable regions (Fig. 5 A and C in chapter 4.6). Furthermore, the most predictive features for both cell types are the DNA accessible regions (Fig. 5 B and D in chapter 4.6), confirming that open chromatin regions, such as enhancers and promoters, show high variability in the binding sites, which has also been reported in previous studies (Jain et al. 2015, Theytelman et al. 2013, D. Park et al. 2013, Wreczycka et al. 2019).

In line with the previous finding, I believe my method can be implemented in the case someone has obtained ChIP-seq results with replicates from a particular cell type and would like to know whether the significant regions need to be considered with caution before downstream interpretative analysis.

6. References

1) Singh, Shashank, et al. "Predicting enhancer-promoter interaction from genomic sequence with deep neural networks." *Quantitative Biology* (2019): 1-16.

2) Talukder, Amlan, et al. "EPIP: a novel approach for condition-specific enhancer–promoter interaction prediction." *Bioinformatics* 35.20 (2019): 3877-3883.

3) Zeng, Wanwen, Mengmeng Wu, and Rui Jiang. "Prediction of enhancer-promoter interactions via natural language processing." *BMC genomics* 19.2 (2018): 84.

4) Belokopytova, Polina, et al. "Quantitative prediction of enhancer-promoter interactions." *bioRxiv* (2019): 541011.

5) Gao, Tianshun, and Jiang Qian. "EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species." *Nucleic acids research* (2019).

6) Poetsch, Anna R., and Christoph Plass. "Transcriptional regulation by DNA methylation." *Cancer treatment reviews* 37 (2011): S8-S12.

7) Morton Bradbury, E. "Reversible histone modification and the chromosome cell cycle." *Bioessays* 14.1 (1992): 9-16.

8) Ramchandani, Shyam, et al. "DNA methylation is a reversible biological signal." *Proceedings of the National Academy of Sciences* 96.11 (1999): 6107-6112.

9) Butler, Jill S., and Sharon YR Dent. "The role of chromatin modifiers in normal and malignant hematopoiesis." *Blood, The Journal of the American Society of Hematology* 121.16 (2013): 3076-3084.

10) Greenberg, Maxim VC, and Deborah Bourc'his. "The diverse roles of DNA methylation in mammalian development and disease." *Nature Reviews Molecular Cell Biology* (2019): 1-18.

11) Li, Xuewen, et al. "Control of germline stem cell differentiation by Polycomb and Trithorax group genes in the niche microenvironment." *Development* 143.19 (2016): 3449-3458.

12) Aloia, Luigi, Bruno Di Stefano, and Luciano Di Croce. "Polycomb complexes in stem cells and embryonic development." *Development* 140.12 (2013): 2525-2534.

13) Arab, Khelifa, et al. "Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A." *Molecular cell* 55.4 (2014): 604-614.

14) Arab, Khelifa, et al. "GADD45A binds R-loops and recruits TET1 to CpG island promoters." *Nature genetics* 51.2 (2019): 217.

15) Zhang, Rui, et al. "LncRNAs and cancer." *Oncology letters* 12.2 (2016): 1233-1239.

16) Bernstein, Bradley E., et al. "The NIH roadmap epigenomics mapping consortium." *Nature biotechnology* 28.10 (2010): 1045.

17) Tolhuis, Bas, et al. "Looping and interaction between hypersensitive sites in the active β-globin locus." *Molecular cell* 10.6 (2002): 1453-1465.

18) Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung H-L, Zhang K, Zhang Y. 2013. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics, Cell 153: 692–706.

19) Lu F, Liu Y, Jiang L, Yamaguchi S, Zhang Y. 2014. Role of Tet proteins in enhancer activity and telomere elongation, Genes Dev 28: 2103–2119.

20) Schäfer, Andrea, et al. "Impaired DNA demethylation of C/EBP sites causes premature aging." *Genes & development* 32.11-12 (2018): 742-762.

21) Zhang, Wu, and Jie Xu. "DNA methyltransferases and their roles in tumorigenesis." *Biomarker research* 5.1 (2017): 1.

22) Okano, Masaki, et al. "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development." *Cell* 99.3 (1999): 247-257.

23) Feng, Jian, et al. "Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons." *Nature neuroscience* 13.4 (2010): 423.

24) Barau, Joan, et al. "The DNA methyltransferase DNMT3C protects male germ cells from transposon activity." *Science* 354.6314 (2016): 909-912.

25) Kohli, Rahul M., and Yi Zhang. "TET enzymes, TDG and the dynamics of DNA demethylation." *Nature* 502.7472 (2013): 472-479.

26) Schomacher, Lars, et al. "Neil DNA glycosylases promote substrate turnover by Tdg during DNA demethylation." *Nature structural & molecular biology* 23.2 (2016): 116.

27) Cortellino, Salvatore, et al. "Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair." *Cell* 146.1 (2011): 67-79.

28) Li, Zheng, et al. "Gadd45a promotes DNA demethylation through TDG." *Nucleic acids research* 43.8 (2015): 3986-3997.

29) Kienhöfer, Sabine, et al. "GADD45a physically and functionally interacts with TET1." *Differentiation* 90.1-3 (2015): 59-68.

30) Sabag, Ofra, et al. "Establishment of methylation patterns in ES cells." *Nature structural & molecular biology* 21.1 (2014): 110.

31) Auclair, Ghislain, et al. "Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse." *Genome biology* 15.12 (2014): 545.

32) Seisenberger, Stefanie, et al. "The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells." *Molecular cell* 48.6 (2012): 849-862.

33) Smith, Zachary D., et al. "A unique regulatory phase of DNA methylation in the early mammalian embryo." *Nature* 484.7394 (2012): 339.

34) Messerschmidt, Daniel M., Barbara B. Knowles, and Davor Solter. "DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos." *Genes & development* 28.8 (2014): 812-828.

35) Pfaffeneder, Toni, et al. "Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA." *Nature chemical biology* 10.7 (2014): 574.

36) Xiao, Chuan-Le, et al. "N6-methyladenine DNA modification in the human genome." *Molecular cell* 71.2 (2018): 306-318.

37) Fornace, Albert J., Isaac Alamo, and M. Christine Hollander. "DNA damage-inducible transcripts in mammalian cells." *Proceedings of the National Academy of Sciences* 85.23 (1988): 8800-8804.

38) Fornace, A. J., et al. "Mammalian genes coordinately regulated by growth arrest signals and DNA-damaging agents." *Molecular and cellular biology* 9.10 (1989): 4196-4203.

39) Liebermann, Dan A., and Barbara Hoffman. "Gadd45 in stress signaling." *Journal of molecular signaling* 3.1 (2008): 15.

40) E Tamura, R., et al. "GADD45 proteins: central players in tumorigenesis." *Current molecular medicine* 12.5 (2012): 634-651.

41) Chi, Hongbo, et al. "GADD45β/GADD45γ and MEKK4 comprise a genetic pathway mediating STAT4□independent IFNγ production in T cells." *The EMBO journal* 23.7 (2004): 1576-1586.

42) Kelman, Zvi, and Jerard Hurwitz. "Protein–PCNA interactions: a DNA-scanning mechanism?." *Trends in biochemical sciences* 23.7 (1998): 236-238.

43) Azam, Naiyer, et al. "Interaction of CR6 (GADD45γ) with proliferating cell nuclear antigen impedes negative growth control." *Journal of Biological Chemistry* 276.4 (2001): 2766-2774.

44) Kearsey, Jonathan M., et al. "Gadd45 is a nuclear cell cycle regulated protein which interacts with p21Cip1." *Oncogene* 11.9 (1995): 1675-1683.

44) Zhan, Qimin, et al. "Association with Cdc2 and inhibition of Cdc2/Cyclin B1 kinase activity by the p53-regulated protein Gadd45." *Oncogene* 18.18 (1999): 2892.

45) Jin, Shunqian, et al. "GADD45-induced cell cycle G2-M arrest associates with altered subcellular distribution of cyclin B1 and is independent of p38 kinase activity." *Oncogene* 21.57 (2002): 8696.

46) Vairapandi, Mariappan, et al. "GADD45b and GADD45g are cdc2/cyclinB1 kinase inhibitors with a role in S and G2/M cell cycle checkpoints induced by genotoxic stress." *Journal of cellular physiology* 192.3 (2002): 327-338.

47) Hollander, M. Christine, and Albert J. Fornace Jr. "Genomic instability, centrosome amplification, cell cycle checkpoints and Gadd45a." *Oncogene* 21.40 (2002): 6228.

48) Kaufmann, Lilian T., Mathias S. Gierl, and Christof Niehrs. "Gadd45a, Gadd45b and Gadd45g expression during mouse embryonic development." *Gene Expression Patterns* 11.8 (2011): 465-470.

49) Gierl, Mathias S., et al. "GADD45G functions in male sex determination by promoting p38 signaling and Sry expression." *Developmental cell* 23.5 (2012): 1032-1042.

50) Ma, Dengke K., et al. "Neuronal activity–induced Gadd45b promotes epigenetic DNA demethylation and adult neurogenesis." *Science* 323.5917 (2009): 1074-1077.

51) Barreto, Guillermo, et al. "Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation." *nature* 445.7128 (2007): 671.

52) Schäfer, Andrea, et al. "Ing1 functions in DNA demethylation by directing Gadd45a to H3K4me3." *Genes & development* 27.3 (2013): 261-273.

53) Niehrs, Christof, and Andrea Schäfer. "Active DNA demethylation by Gadd45 and DNA repair." *Trends in cell biology* 22.4 (2012): 220-227.

54) Marikawa, Yusuke, and Vernadeth B. Alarcón. "Establishment of trophectoderm and inner cell mass lineages in the mouse embryo." *Molecular Reproduction and Development: Incorporating Gamete Research* 76.11 (2009): 1019-1032.

55) Bradley, Allan, et al. "Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines." *Nature* 309.5965 (1984): 255-256.

56) Nichols, Jennifer, and Austin Smith. "Naive and primed pluripotent states." *Cell stem cell* 4.6 (2009): 487-492.

57) Brons, I. Gabrielle M., et al. "Derivation of pluripotent epiblast stem cells from mammalian embryos." *Nature* 448.7150 (2007): 191.

58) Tesar, Paul J., et al. "New cell lines from mouse epiblast share defining features with human embryonic stem cells." *Nature* 448.7150 (2007): 196.

59) Fujikura, Junji, et al. "Differentiation of embryonic stem cells is induced by GATA factors." *Genes & development* 16.7 (2002): 784-789.

60) Okamoto, Koji, et al. "A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells." *Cell* 60.3 (1990): 461-472.

61) Rosner, Mitchell H., et al. "Oct-3 is a maternal factor required for the first mouse embryonic division." *Cell* 64.6 (1991): 1103-1110.

62) Schöler, H. R., et al. "Oct□4: a germline□specific transcription factor mapping to the mouse t□complex." *The EMBO journal* 9.7 (1990): 2185-2195.

63) Niwa, Hitoshi, Jun-ichi Miyazaki, and Austin G. Smith. "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells." *Nature genetics* 24.4 (2000): 372.

64) Habibi, Ehsan, et al. "Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells." *Cell stem cell* 13.3 (2013): 360-369.

65) Torres-Padilla, Maria-Elena, and Ian Chambers. "Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage." *Development* 141.11 (2014): 2173-2181.

66) Tsai, Ping-Hsing, et al. "Ash2l interacts with Oct4-stemness circuitry to promote super-enhancer-driven pluripotency network."

67) Rulands, Steffen, et al. "Genome-scale oscillations in DNA methylation during exit from pluripotency." *Cell systems* 7.1 (2018): 63-76.

68) Macfarlan, Todd S., et al. "Embryonic stem cell potency fluctuates with endogenous retrovirus activity." *Nature* 487.7405 (2012): 57.

69) De Iaco, Alberto, et al. "DUX-family transcription factors regulate zygotic genome activation in placental mammals." *Nature genetics* 49.6 (2017): 941.

70) Hendrickson, Peter G., et al. "Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons." *Nature genetics* 49.6 (2017): 925.

71) De Iaco, Alberto, et al. "DPPA2 and DPPA4 are necessary to establish a 2C□ like state in mouse embryonic stem cells." *EMBO reports* 20.5 (2019).

72) Eckersley-Maslin, Mélanie, et al. "Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program." *Genes & development* 33.3-4 (2019): 194-208.

73) Reuter, Jason A., Damek V. Spacek, and Michael P. Snyder. "High-throughput sequencing technologies." *Molecular cell* 58.4 (2015): 586-597.

74) Johnson, David S., et al. "Genome-wide mapping of in vivo protein-DNA interactions." *Science* 316.5830 (2007): 1497-1502.

75) Mikkelsen, Tarjei S., et al. "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." *Nature* 448.7153 (2007): 553.

76) Barski, Artem, et al. "High-resolution profiling of histone methylations in the human genome." *Cell* 129.4 (2007): 823-837.

77) Lickwar, Colin R., et al. "Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function." *Nature* 484.7393 (2012): 251.

78) Furey, Terrence S. "ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions." *Nature Reviews Genetics* 13.12 (2012): 840.

79) Buenrostro, Jason D., et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature methods* 10.12 (2013): 1213.

80) Buenrostro, Jason D., et al. "Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation." *Cell* 173.6 (2018): 1535-1548.

81) Lareau, Caleb A., et al. "Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility." *Nature Biotechnology* (2019): 1.

82) Lareau, Caleb, et al. "Inference and effects of barcode multiplets in droplet-based single-cell assays." *bioRxiv* (2019): 824003.

83) Cusanovich, Darren A., et al. "The cis-regulatory dynamics of embryonic development at single-cell resolution." *Nature* 555.7697 (2018): 538.

84) Cusanovich, Darren A., et al. "A single-cell atlas of in vivo mammalian chromatin accessibility." *Cell* 174.5 (2018): 1309-1324.

85) Cusanovich, Darren A., et al. "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing." *Science* 348.6237 (2015): 910-914.

86) Chen, Huidong, et al. "Assessment of computational methods for the analysis of single-cell ATAC-seq data." *Genome biology* 20.1 (2019): 1-25.

87) Hardcastle, Thomas J. "High-throughput sequencing of cytosine methylation in plant DNA." *Plant methods* 9.1 (2013): 16.

88) Meissner, Alexander, et al. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis." *Nucleic acids research* 33.18 (2005): 5868-5877.

89) Staševskij, Zdislav, et al. "Tethered oligonucleotide-primed sequencing, TOP-Seq: a high-resolution economical approach for DNA epigenome profiling." *Molecular cell* 65.3 (2017): 554-564.

90) ENCODE Project Consortium. "The ENCODE (ENCyclopedia of DNA elements) project." *Science* 306.5696 (2004): 636-640.

91) Amemiya, Haley M., Anshul Kundaje, and Alan P. Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome." *Scientific Reports* 9.1 (2019): 9354.

92) Teytelman, Leonid, et al. "Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins." *Proceedings of the National Academy of Sciences* 110.46 (2013): 18602-18607.

93) Park, Daechan, et al. "Widespread misinterpretable ChIP-seq bias in yeast." *PloS one* 8.12 (2013): e83506.

94) Jain, Dhawal, et al. "Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments." *Nucleic acids research* 43.14 (2015): 6959-6968.

95) Wreczycka, Katarzyna, et al. "HOT or not: examining the basis of high-occupancy target regions." *Nucleic acids research* 47.11 (2019): 5735-5745.

96) GTEx Consortium. "The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." *Science* 348.6235 (2015): 648-660.

97) Wu, Angela R., et al. "Quantitative assessment of single-cell RNA-sequencing methods." *Nature methods* 11.1 (2014): 41.

98) Grosselin, Kevin, et al. "High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer." *Nature genetics* 51.6 (2019): 1060.

99) Rotem, Assaf, et al. "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state." *Nature biotechnology* 33.11 (2015): 1165.

100) Luo, Chongyuan, et al. "Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex." *Science* 357.6351 (2017): 600-604.

101) Mulqueen, Ryan M., et al. "Highly scalable generation of DNA methylation profiles in single cells." *Nature biotechnology* 36.5 (2018): 428.

102) Saliba, Antoine-Emmanuel, et al. "Single-cell RNA-seq: advances and future challenges." *Nucleic acids research* 42.14 (2014): 8845-8860.

103) Wolock, Samuel L., Romain Lopez, and Allon M. Klein. "Scrublet: computational identification of cell doublets in single-cell transcriptomic data." *Cell systems* 8.4 (2019): 281-291.

104) Kiselev, Vladimir Yu, Tallulah S. Andrews, and Martin Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data." *Nature Reviews Genetics* 20.5 (2019): 273-282.

105) Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines." *Experimental & molecular medicine* 50.8 (2018): 1-14.

106) Kiselev, Vladimir Yu, et al. "SC3: consensus clustering of single-cell RNA-seq data." *Nature methods* 14.5 (2017): 483.

107) Yau, Christopher. "pcaReduce: hierarchical clustering of single cell transcriptional profiles." *BMC bioinformatics* 17.1 (2016): 140.

108) de Boer, Carl G., and Aviv Regev. "BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization." *BMC bioinformatics* 19.1 (2018): 253.

109) Schep, Alicia N., et al. "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data." *Nature methods* 14.10 (2017): 975.

110) Pliner, Hannah A., et al. "Cicero predicts cis-regulatory DNA Interactions from single-cell chromatin accessibility data." *Molecular cell* 71.5 (2018): 858-871.

111) González-Blas, Carmen Bravo, et al. "cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data." *Nature methods* 16.5 (2019): 397.

112) Zamanighomi, Mahdi, et al. "Unsupervised clustering and epigenetic classification of single cells." *Nature communications* 9.1 (2018): 2410.

113) Baker, Syed Murtuza, et al. "Classifying cells with Scasat, a single-cell ATAC-seq analysis tool." *Nucleic acids research* 47.2 (2018): e10-e10.

114) Ji, Zhicheng, Weiqiang Zhou, and Hongkai Ji. "Single-cell regulome data analysis by SCRAT." *Bioinformatics* 33.18 (2017): 2930-2932.

115) Fang, Rongxin, et al. "Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types." *bioRxiv* (2019): 615179.

116) Danese, Anna, et al. "epiScanpy: integrated single-cell epigenomic analysis." *bioRxiv* (2019): 648097.

117) de Souza, Camila PE, et al. "Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data." *bioRxiv* (2018): 414482.

118) Hui, Tony, et al. "High-resolution single-cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations." *Stem cell reports* 11.2 (2018): 578-592.

119) Angermueller, Christof, et al. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning." *Genome biology* 18.1 (2017): 67.

120) Kapourani, Chantriolnt-Andreas, and Guido Sanguinetti. "Melissa: Bayesian clustering and imputation of single-cell methylomes." *Genome biology* 20.1 (2019): 61.

121) Abdelaal, Tamim, et al. "A comparison of automatic cell identification methods for single-cell RNA-sequencing data." *bioRxiv* (2019): 644435.

122) Wagner, Florian, and Itai Yanai. "Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data." *BioRxiv* (2018): 456129.

123) Domanskyi, Sergii, et al. "Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters." *bioRxiv* (2019): 539833.

124) Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Mahrt E, et al. SCINA: semi-supervised analysis of single cells in silico. BioRxiv. 2019; 559872.

125) Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053 – 1058.

126) Cao Z-J, Wei L, Lu S, Yang D-C, Gao G. Cell BLAST: searching large-scale scRNA-seq databases via unbiased cell embedding. BioRxiv. 2019; 587360.

127) Ma F, Pellegrini M. Automated identification of cell types in single cell RNA sequencing. BioRxiv. 2019; 532093.

128) Johnson TS, Wang T, Huang Z, Yu CY, Wu Y, Han Y, et al. LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves dsubtype detection. Bioinformatics. 2019.

129) Alquicira-Hernandez J, Nguyen Q, Powell JE. scPred: scPred: cell type prediction at single-cell resolution. BioRxiv. 2018; 369538.

130) Kanter JK de, Lijnzaad P, Candelli T, Margaritis T, Holstege F. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. BioRxiv. 2019; 558908.

131) Lieberman Y, Rokach L, Shay T. CaSTLe – classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. PLoS One. 2018;13: e0205499.

132) Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163 –172.

133) Boufea K, Seth S, Batada NN. scID: identification of equivalent transcriptional cell populations across single cell RNA-seq data using discriminant analysis. bioRxiv. 2019; 470203 .

134) Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. bioRxiv. 2018; 508085.

135) Stadler, Michael B., et al. "DNA-binding factors shape the mouse methylome at distal regulatory regions." *Nature* 480.7378 (2011): 490.

136) Whyte, Warren A., et al. "Enhancer decommissioning by LSD1 during embryonic stem cell differentiation." *Nature* 482.7384 (2012): 221.

137) Greenberg, Maxim VC, and Deborah Bourc'his. "The diverse roles of DNA methylation in mammalian development and disease." *Nature Reviews Molecular Cell Biology* (2019): 1-18.

138) Battle, Stephanie L., et al. "Enhancer Chromatin and 3D Genome Architecture Changes from Naive to Primed Human Embryonic Stem Cell States." *Stem cell reports* 12.5 (2019): 1129-1144.

139) Jadhav, Unmesh, et al. "Extensive recovery of embryonic enhancer and gene memory stored in hypomethylated enhancer DNA." *Molecular cell* 74.3 (2019): 542-554.

140) Daugherty, Aaron C., et al. "Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans." *Genome research* 27.12 (2017): 2096-2107.

141) Bozek, Marta, et al. "ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm." *Genome research* 29.5 (2019): 771-783.

142) Teytelman, Leonid, et al. "Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins." *Proceedings of the National Academy of Sciences* 110.46 (2013): 18602-18607.

143) Steinacher, Roland, et al. "SUMOylation coordinates BERosome assembly in active DNA demethylation during cell differentiation." *The EMBO journal* 38.1 (2019).

144) Amouroux, Rachel, et al. "De novo DNA methylation drives 5hmC accumulation in mouse zygotes." *Nature cell biology* 18.2 (2016): 225.

145) Zhang, Rui-peng, Jian-zhong Shao, and Li-xin Xiang. "GADD45A protein plays an essential role in active DNA demethylation during terminal osteogenic differentiation of adipose-derived mesenchymal stem cells." *Journal of Biological Chemistry* 286.47 (2011): 41083-41094.

146) Boyer, Laurie A., et al. "Core transcriptional regulatory circuitry in human embryonic stem cells." *cell* 122.6 (2005): 947-956.

147) Thomson, Matt, et al. "Pluripotency factors in embryonic stem cells regulate differentiation into germ layers." *Cell* 145.6 (2011): 875-889.

148) Chen, Zhiyuan, and Yi Zhang. "Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development." *Nature genetics* (2019): 1.

149) Kawasaki, Fumiko, et al. "Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania." *Genome biology* 18.1 (2017): 23.

150) Greene JR, Morrissey LM, Foster LM, Geiduschek EP. DNA binding by the bacteriophage SPO1-encoded type II DNA-binding protein, transcription factor 1. Formation of nested complexes at a selective binding site. J Biol Chem. 1986;261:12820–7.

151) Seifermann, Marco, et al. "Role of the DNA repair glycosylase OGG1 in the activation of murine splenocytes." *DNA repair* 58 (2017): 13-20.

152) Gu, Tian-Peng, et al. "The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes." *Nature* 477.7366 (2011): 606.

153) Hildesheim, Jeffrey, et al. "Gadd45a protects against UV irradiation-induced skin tumors, and promotes apoptosis and stress signaling via MAPK and p53." *Cancer research* 62.24 (2002): 7305-7315.

154) Dawlaty, Meelad M., et al. "Loss of Tet enzymes compromises proper differentiation of embryonic stem cells." *Developmental cell* 29.1 (2014): 102-111.

155) Jackson, Melany, et al. "Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells." *Molecular and cellular biology* 24.20 (2004): 8862-8871.

156) Okano, Masaki, et al. "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development." *Cell* 99.3 (1999): 247-257.

157) Lei H, Oh SP, Okano M, Jüttermann R, Goss KA, Jaenisch R, Li E. 1996. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells, Development 122: 3195–3205

7. List of abbreviations

5mC 5-methyl-cytosine

5hmC 5-hydroxy-methyl-cytosine

5fC 5-formyl-cytosine

5caC 5-carboxy-methyl-cytosine

mESCs mouse embryonic stem cells

WGBS whole-genome bisulfite sequencing

RRBS reduced representation bisulfite sequencing

DMRs differentially methylated regions

GADD45 Growth Arrest and DNA Damage

TET Ten-eleven translocation methylcytosine dioxygenase

2C 2 cell

TDG Thymine DNA glycosylase

scATAC single-cell assay for transposase-accessible chromatin using sequencing

RAGI residual average Gini index

PBMCs peripheral blood mononuclear cells

HOT high occupancy target regions

VOT variable occupancy target regions

ChIP-seq chromatin immunoprecipitation followed by sequencing

ENCODE encyclopedia of DNA elements

8. Acknowledgements

9. Lebenslauf