UNIVERSITÀ DEGLI STUDI DI TORINO

Doctoral School in Life and Health Sciences

PhD Programme in Complex Systems for Life Sciences

XXXIII Cycle

Integrative analysis of multi-omics data for the study of gene regulation in stem cell differentiation and cancer

TUTOR:
Prof. Salvatore Oliviero

PhD CANDIDATE:
Andrea Lauria

Abstract

In recent years, the advent of high-throughput technologies enabled the investigation of complex biological systems at multiple molecular levels, such as the genome, transcriptome and epigenome. During my PhD experience, I worked at two main projects, which involved the integration and interpretation of multiple omics data sources in order to elucidate the gene regulatory mechanisms driving both pathological – cancer – and physiological – stem cell differentiation – phenotypes.

I. *Identification of altered biological processes in heterogeneous cohorts by discretization of expression profiles*

Heterogeneity is a fundamental feature of complex phenotypes. So far, genomic screenings have profiled thousands of samples providing insights into the transcriptome of the cell. However, disentangling the heterogeneity of these transcriptomic Big Data to identify defective biological processes remains challenging. Genomic studies have shown that distinct genes sharing common biological functions or involved in the same biological process (*i.e.* gene sets) can be altered in different patients affected by same condition. In this context, the gene set analysis (GSA) aims at identifying groups of genes whose cumulative expression is altered in the phenotype of interest. Although widely used, conventional GSA algorithms have been developed to treat microarray data of small-sized case-control studies. Therefore, their application is not suitable to RNA-sequencing profiles of large heterogeneous cohorts. During the first part of my PhD, my work has been focused on the development and testing of a new GSA algorithm, namely GSECA (Gene Set Enrichment Class Analysis), that employs a model-based data discretization approach to exploit the specific bimodal behaviour of RNA-sequencing profiles and increase signal-to-noise ratio in heterogeneous high-volume datasets. Using simulated and experimental RNA-sequencing data sets, we showed that GSECA provides higher performances than other available algorithms in detecting truly altered biological processes in large cohorts. Applied to 5941 samples from 14 different cancer types, GSECA correctly identified the alteration of the PI3K/AKT signaling pathway driven by the somatic loss of PTEN and verified the emerging role of PTEN in modulating immune-related processes. In particular, we showed that, in prostate cancer, PTEN loss appears to establish an immunosuppressive tumor microenvironment through the activation of STAT3, and low PTEN expression levels have a detrimental impact on patient disease-free survival.

II. *Dissecting the functional role of de novo DNA methylation during embryonic lineage specification*

The correct establishment of DNA methylation patterns is essential for cell fate specification in the developing embryo. However, the molecular targets as well as the mechanisms that determine the specificity of the *de novo* methylation machinery during differentiation are not completely elucidated. In the second part of my PhD, we investigated the role of the *de novo* DNA methyltransferases in controlling lineage-fate decision during mouse early development. Using a combination of *in vitro* stem cell differentiation models, loss of function experiments and high-throughput multi-omics approaches – WGBS, ChIP-, bulk-, and single-cell-RNA sequencing -, we demonstrated that *Dnmt3b*-dependent methylation is essential for the correct specification of the meso-endodermal lineages. Our results showed that, in the transition from the pre- (Embryonic Stem Cells) to the early post-implantation embryo (Epiblast Stem Cells), *Dnmt3b* activity is directed towards regulatory regions associated with key developmental transcription factors, acting as an epigenetic priming that ensures flawless commitment at later stages. We found that the differentiation into meso-endodermal progenitors is impaired in *Dnm3b* knockout (3BKO) cells, which are redirected towards neuro-ectodermal lineages. Finally, we demonstrated that the impaired meso-endodermal induction of 3BKO cells can be rescued by silencing *Sox2*, a master regulator of neuronal differentiation.

Table of Contents

List of Figures

Part I - Chapter 1

Part I - Chapter 2

Part I - Chapter 3

Part I – Chapter 4

Part I – Chapter 5

Part II – Chapter 1

Part II – Chapter 2

Part II – Chapter 3

Part II – Chapter 4

Part II – Chapter 5

Acknowledgements

I would like to express my gratitude to:

My PhD supervisor, Professor *Salvatore Oliviero*, for giving me the opportunity to join his research group, and for the important scientific training and mentorship work carried out over the three years of my PhD research activity.

All the past and present members of *Oliviero*'s Functional Genomics/Epigenomics laboratory, for all the help, constant support and sharing of scientific experience*: Francesca Anselmi, Carlo Castiglione, Hassan Dastsooz, Danny Incarnato, Mara Maldotti, Gouhua Meng, Fatemeh Mirzadeh, Ivan Molineris, Edoardo Morandi, Francesco Neri, Caterina Parlato, Isabelle Laurence Polignano, Valentina Proserpio, Stefania Rapelli, Lisa Marie Simon, Mirko Scrivano, Annalaura Tamburrini*.

Professor *Michele Caselle,* for introducing me to the science of complex systems and the world of scientific research, for the important scientific collaboration and for all the useful discussions and advices.

Professor *Giacomo Donati* and the members of his research group, for the scientific collaboration, helpful discussions and advices.

Special thanks go to Doctor *Matteo Cereda*, head of the Cancer Genomics and Bioinformatics group at Italian Institute for Genomic Medicine, for introducing me to the field of bioinformatics and the significant scientific collaboration carried out during my PhD experience, and all the members of his research group, especially *Serena Peirone* and *Marco Del Giudice* for their precious efforts.

Turin, March 2021

Andrea Lauria

Chapter 0

Multi-omics data analysis

Over the past decades, the advent of Next Generation Sequencing (NGS) technologies opened the gates of a new era in molecular biology, promoting a huge improvement towards the understanding of complex biological systems. Technological advances in NGS-based functional genomics assays - like RNA sequencing (RNA-seq), chromatin immunoprecipitation followed by sequencing (ChIP-seq), sequencing of bisulfite-converted DNA (BS-seq) – enhanced the possibilities of biological investigations by querying multiple molecular sources of information, such as the DNA, RNA, proteins and their related biochemical modifications – known as the epigenome. The simultaneous investigation of these molecular profiles, as well as the integration of other data sources like clinical data, provided new insights into the mechanistic understanding of complex diseases, such as cancer. The increasing availability and cost-efficiency of sequencing platforms promoted an explosion of large-scale biomedical screening projects, fostering collaboration of research groups in large genomics consortia and leading to the development of public repositories hosting multi-omics 'Big Data' (*e.g.* the ENCODE consortium or The Cancer Genome Atlas). More recently, the advent of single-cell sequencing technologies enabled genome-wide profiling of cell molecular components at individual cell resolution, opening the possibility to answer new question about cell states and cell-to-cell heterogeneity with unprecedent detail. These techniques enlightened the study of processes like cell differentiation during development, enabling the possibility to track the trajectory crossed by each individual cell type and understand the molecular players as well as the regulatory mechanisms that drive cell decisions.

The interpretability of multi-omics data is challenged by the great complexity and heterogeneity proper of genomics Big Data. As a consequence, their understanding requires tailored computational analysis techniques, coupling traditional bioinformatics tools with new data science methodologies, in order to properly exploit the multiple interconnected layers of biological information and acquire a system-level knowledge of the regulatory processes responsible for the phenotype of interest. In this dissertation, two main works are reported, covering most of the research activity carried out over the three years of my PhD experience. Both projects were characterized by the integrative analysis and interpretation of multiple omics data sources, in order to elucidate the gene regulatory mechanisms driving both pathological – cancer – and physiological – stem cell differentiation – phenotypes. In the first work, we focused on the

development of a new data mining algorithm to perform gene set enrichment analysis in RNA-seq experiments, with applications to the analysis of cancer transcriptomics data available from The Cancer Genome Atlas (TCGA). In the second work, in joint efforts between the experimental and computational side of the lab, we analyzed the role played by *de novo* DNA methylation in the regulation of cell differentiation during the early stages of mouse embryonic development, exploiting a combination of *in vitro* stem cell differentiation models, loss-of-function experiments and high-throughput approaches – *i.e.* WGBS, ChIP-, bulk-, and single-cell-RNA sequencing.

Part I

Identification of altered biological processes in heterogeneous RNA-sequencing data by discretization of expression profiles

The following chapters are adapted from the journal article:

Andrea Lauria*, Serena Peirone*, Marco Del Giudice*, Francesca Priante, Prabhakar Rajan, Michele Caselle, Salvatore Oliviero, Matteo Cereda, **Identification of altered biological processes in heterogeneous RNA-sequencing data by discretization of expression profiles**, *Nucleic Acids Research*, Volume 48, Issue 4, 28 February 2020, Pages 1730–1747, https://doi.org/10.1093/nar/gkz1208

* Joint First Authors

Chapter 1

Introduction I

1.1 Gene expression

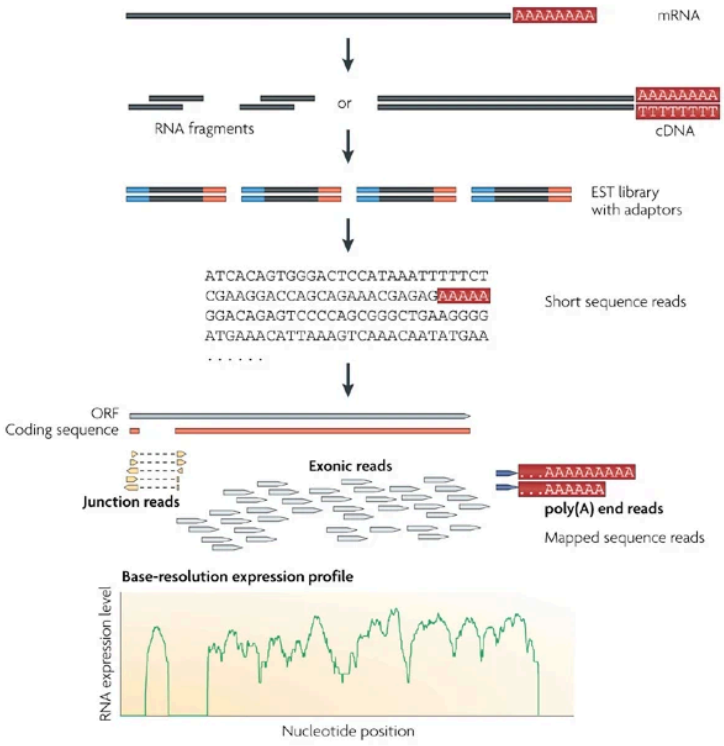
The central dogma of molecular biology, established by Francis Crick in 1958, explains the transfer of sequence information between information-carrying biopolymers in living organisms. DNA is the biomolecule that stores the genetic information necessary to the functioning, growth, development and reproduction of all known living organisms. These instructions are organized into smaller pieces of genetic material, called genes. In order to be used, they have to be transcribed into a RNA molecule (transcript). This can then be translated into a protein (in the case of protein-coding genes), or, for the non-coding genes, it can carry out other functional roles in the cell (e.g. post-transcriptional regulation, ribosomal RNA etc.). This entire process is called gene expression, and transcription represents its first step. The central dogma describes the flow of genetic information associated with gene expression as a general information transfer mechanism (i.e. believed to occur normally in most cells) being unidirectional from DNA into RNA and into protein as the final functional product, but never from protein to nucleic acids (1). A wide range of mechanisms can be used by the cells to turn genes on and off if necessary, modulating any step of the gene expression chain, including transcription – a process known as transcriptional regulation.

1.2 Transcriptomics technologies

The transcriptome is the complete set of transcripts in cells, for a specific tissue, developmental stage or condition. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease. The key aims of transcriptomics are: to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions, in order to assess the differential expression of genes at the bases of distinct phenotypes.

Different experimental techniques have been developed in order to measure and quantify the transcriptome. Popular choices before the introduction of Next Generation Sequencing (NGS) based tools were given by hybridization methods, based on DNA microarrays (2). With the emergence of NGS technologies, sequence-based approaches have become the gold-standard for transcriptome profiling experiments. RNA sequencing (RNA-seq) is most often used for analysing differential gene expression (3). The standard workflow begins with RNA extraction and purification, followed by mRNA enrichment or ribosomal RNA depletion. The remaining RNA is then fragmented and reverse transcribed into cDNA molecules. sequence adaptors are ligated and, after the selection of fragment sizes, the ends of cDNAs are sequenced on a high-throughput platform – typically, Illumina -, producing many short reads. Then follows the computational analysis: the short reads are pre-processed by removing low-quality read, artefacts or sequence errors. The pre-processed reads are aligned and/or assembled to a transcriptome, and the gene/transcript expression levels are estimated by quantifying reads that overlap transcripts. The last steps for differential expression analysis involve filtering and normalizing between samples, and statistical modelling of significant changes in the expression levels of individual genes and/or transcripts between sample groups (2,3) (Figure 1.1).

RNA-seq offers several advantages in comparison to hybridization methods. First, it is not limited to detecting transcripts that correspond to known genomic sequence. Furthermore, it has a very low background noise since the cDNA sequences can be unambiguously mapped to unique regions of the genome.



I - Figure 1. 1: RNA-seq experimental workflow. Schematic representation of the workflow of a typical RNA-seq experiment. – Adapted from (2)

Moreover, RNA-seq does not have an upper limit for quantification, so it has a high dynamic range and it is also able to capture transcripts that are expressed at very low levels. By contrast, DNA microarrays lack sensitivity for genes that are expressed at both high or low levels. Additionally, the results of RNA-seq show high levels of reproducibility(2). Nowadays, RNA-seq from bulk tissues and cells remains the conventional approach to measure gene expression; however, it provides an average reading of the transcriptome across cell populations, thus not preserving single-cell specific information. Single-cell laboratory and analysis methods enabled to move beyond bulk analysis and are rapidly being adopted by the research community (3).

1.3 Gene Set Analysis

In recent years, genomic screenings, powered by the improvements of NGS technologies, have studied RNA-seq expression profiles of large cohorts to gain insights into complex phenotypes, including cancer. Despite the fact that transcriptomic analyses are a well-established practice in biomedical research, the biological interpretation of the gene expression profiles coming from high-throughput screenings is still a major challenge. In addition to the intrinsic complexity of the biological mechanisms, a major hurdle is the presence of inter-sample heterogeneity (IH), or the variable expression of genes across samples due to genetic, environmental, demographic, and technical factors (4). A typical approach to the interpretation of the transcriptome profiles is the differential expression analysis. It consists on determining the list of differentially expressed genes between different collections of samples associated with a treatment condition or belonging to distinct phenotypes (5,6). One can then focus on a small number of "most" differentially expressed genes to discern important mechanisms related with the studied phenotype. This approach is based on single-gene level analysis and, although it gives us relevant biological information, it has some limitations. In fact, cellular processes often do not rely on the action of a single gene, but are complex mechanisms involving the interaction of multiple genes acting together. Therefore, single-gene level analysis may miss important effect on pathways. Moreover, considering the emerging heterogeneity of samples coming from high-throughput screenings, it becomes difficult to associate the effect of a single or few differentially expressed genes with a given phenotype. For these reasons, instead of focusing on single-gene level analysis, useful biological insights can be obtained by investigating the expression profiles at the level of gene sets.

Gene sets are groups of genes that share a particular property, such as the involvement in a biological process or the association with the same disease. They are defined based on prior biological knowledge, which comes from previous studies concerning molecular interactions in pathways or co-expression of genes. Hence, the gene-set analysis (GSA) aims at identifying gene sets whose cumulative expression is altered in the phenotype of interest (7,8). The gene-set approach has two major advantages over single-gene differential expression analysis. On one hand, GSA reduces the potentially large list of differentially expressed genes into a smaller list of altered

gene sets (AGS) - which include biological processes, molecular functions or biological pathways - that can give a "snapshot" of what happens to the phenotype under study at the system-level; on the other hand, considering the combined changes in the expression levels of multiple genes, GSA may identify gene set that are associated with a phenotype even when differential expression cannot be claimed with usual criteria, exploiting the fact that many genes in the gene set undergo subtle changes(9). In addition to that, the gene-set perspective makes it possible to deal with the heterogeneity of samples. Indeed, for a particular phenotype of interest, different functionally-related genes can be deregulated among different samples. By treating them as a gene set, it is still possible to associate their deregulation with the condition under study, considering the combined effect of different genes in different samples. Therefore, GSA is a powerful technique for interpreting transcriptome profiling experiments.

1.4 Features of GSA methodologies

Since the first appearance of the gene-set approach for the interpretation of gene expression profiles, a certain number of different gene set analysis methods has been proposed. The basic assumption among all proposed methods is that when a gene set is relevant to a phenotype, a considerable fraction of its gene will show a certain degree of differential expression between the two cohorts in either one or both direction (upregulated or down-regulated). Nonetheless, the different methods often rely on different formulations of the null hypothesis and test statistics.

All existent methods can be roughly classified into three main categories(9):

- *Over-representation analysis* (ORA): this is the earliest approach to gene set analysis. This class of methods rely on a contingency table analysis testing for the association between the Differential Expression (DE) status of a gene (DE or not) and its membership to a given gene set. That is, the contingency table tests if the genes of a particular gene set are over-represented in the list of differentially expressed genes, with a hypergeometric or chi-square (χ^2) distribution. However, ORA methods cannot be applied if no DE are found, and their results are heavily dependent on the threshold used to select significantly DE genes.
- *Functional class scoring* (FCS): this second class of methods overcome the need of selecting differentially expressed genes in the first place. They derive an enrichment score for all the genes of a given gene set regardless of whether or not they are differentially expressed. Typically, the genes are ranked according to their expression value, using some consistent statistics (e.g. fold change, p-value from DE analysis, signal-to-noise ratio (7)). Then, an enrichment score is defined for a gene set, based on the distance of its genes in the ranked gene list. Such methods include GSEA (Gene Set Enrichment Analysis)(7) , one of the most cited and used GSA methods.

- *Single-Sample* (SS): this third class of methods is similar to FCS, but they compute a gene set score in each individual sample from the observed gene expression levels.

Among all these methods, an important distinction is given by the formulation of the null hypothesis tested (10). In particular, GSA algorithms can be divided into 'self-contained' and 'competitive' algorithms depending on whether they identify altered gene sets (AGSs) while ignoring or not genes that are outside the gene set of interest:

- *Competitive* tests: a competitive test compares differential expression of the gene set to a standard defined by the complement of that gene set. This is the case of ORA methods. The competitive null hypothesis to be tested is:
    - *H0 = "The genes in G are at most as often differentially expressed as the genes in G^c"*

  where G is the gene set of interest and G^c its complement.
- Self-contained tests: a self-contained test compares the gene set to a fixed standard that does not depend on the measurement of genes outside the gene sets. FCS and Single-Sample methods fall in this category. The self-contained null hypothesis is:
    - *H0 = "No genes in G are differentially expressed"*

The self-contained null hypothesis is more restrictive than the competitive null hypothesis. From a practical viewpoint, the competitive methods can be applied even with one sample per cohort as they rely on genes as sampling unit. However, they cannot work if no genes outside the gene set are measured. On the other hand, the self-contained methods use the subjects as the sampling unit and hence require several samples per group to infer significance of the gene sets. Unlike the competitive methods, some of the self-contained methods can be applied even when only the genes in the gene set are profiled (11).

1.5 Limitations of existing GSA tools

Although GSA is a widely-employed technique in gene expression studies, most existing GSA methods suffer a few marked limitations (9,10,12). Firstly, GSA algorithms have been designed to handle microarray expression data and subsequently adopted to handle RNA-seq data (7,13). RNA-seq gene expression profiles are characterized by a bimodal behavior reflecting the presence of two major subpopulations of genes in cells (i.e. lowly and highly expressed genes) (14). This behavior is not observable using low-sensitive microarray experiments (15), and to date it has not been taken into account by existing GSA methods. Thus, their application to RNA-seq expression profiles may not be efficient (13). Secondly, GSA methods have been developed to handle experimental conditions in the absence of IH (i.e. altered genes are concordantly activated or repressed in the cohort of interest) (9). As a consequence, biological processes composed of genes

that exhibit a significant excess of coordinated variability (i.e. activated or repressed in different subpopulations) cannot be detected by conventional GSA methods (9). Finally, most existing GSA methods have been designed to assess gene expression of case–control studies with limited sample size and, thus, characterized by a negligible IH and high signal-to-noise ratio (13). These limitations become crucial in the analysis of RNA-seq datasets of large-scale screening projects that are characterized by a high IH. Therefore, GSA algorithms that are able to handle IH are needed.

1.6 Aim of the project

In this work, we addressed the aforementioned limitations, developing a new approach to the gene set analysis called 'Gene Set Enrichment Class Analysis' (GSECA), whose purpose is the identification of AGSs in heterogeneous high-volume RNA-seq datasets. GSECA implements a sample-specific finite mixture modeling (FMM) approach to assess the bimodal distribution of each RNA-seq profile followed by a model-based data discretization (DD) process to increase the signal-to-noise ratio. Discretized data are then evaluated in a statistical framework to detect AGSs between two groups of samples. We showed that GSECA has the highest sensitivity and specificity in detecting AGSs as compared to other 'state-of-the-art' GSA algorithms in the presence of IH on both simulated and real RNA-seq data.

We next used our approach to identify the biologically relevant gene sets that are altered upon the somatic loss of PTEN, and the subsequent alteration of the PI3K/AKT signaling cascade, in 14 different cancer types, verifying the emerging role of PTEN in modulating immune-related processes. In particular, we showed that, in prostate cancer, PTEN loss appears to establish an immunosuppressive tumor microenvironment through the activation of STAT3, and low PTEN expression levels have a detrimental impact on patient disease-free survival.

We implemented the method as R software (https://www.r-project.org/), and it is freely available from GitHub (https://github.com/matteocereda/GSECA). It can be run from the command line, within R, or with a GUI in a user-friendly R/Shiny application.

Chapter 2

Gene Set Enrichment Class Analysis

Results

2.1 Method overview

The key idea of the Gene Set Enrichment Class Analysis (GSECA) is the definition of classes of gene expression levels, in which all the quantified transcripts are grouped, as a means to increase statistical power of GSA in presence of IH. To achieve this, we implemented a 'model-based' data discretization (MDD) procedure fulfilling both a biological and a 'statistical' requirement. First, we required that the division of expression values into expression classes (ECs) must resemble the presence of two major subpopulations of lowly and highly expressed genes in the cells (14) (i.e. biological requirement). Second, we considered that the discretization process must provide an adequate distribution of genes among classes and, thus, ensure a similar degree of statistical power for the subsequent tests performed for all ECs (i.e. statistical requirement).

2.1.1 GSECA algorithm

To identify AGSs in a list of gene sets $G = \{G_1, \dots, G_n\}$ between two cohorts $A = \{a_1, \dots, a_n\}$ and $B = \{b_1, \dots, b_n\}$ of heterogeneous RNA-seq expression profiles, the algorithm runs through three sequential steps (Figure 2.1):

1. *The sample-specific FMM of gene expression levels distributions* (Figure 2.1, Step 1). Given the bimodal behavior of the gene expression levels distributions of each sample (i.e. the two subpopulations of lowly and highly expressed genes in the cell), we adopted a FMM approach to model the RNA-seq expression profiles of all protein coding genes in each sample. FMM are defined as a convex combination of a finite number of probability density functions (pdfs). Therefore, by combining the properties of the individual pdfs, they are able to approximate any arbitrary distribution, thus offering a powerful and flexible tool to model complex data (16). In particular, GSECA models the bimodal distribution of RNA-seq expression profile *x* of a given sample *i* as a mixture of two Gaussian pdfs *Φ*, as previously proposed (14):

$$f(\mathrm{x_i}) = \lambda_1 \Phi(\mathrm{x_i};\ \mu_1, \sigma_1) + \lambda_2 \Phi(\mathrm{x_i};\ \mu_2, \sigma_2) \qquad (1)$$

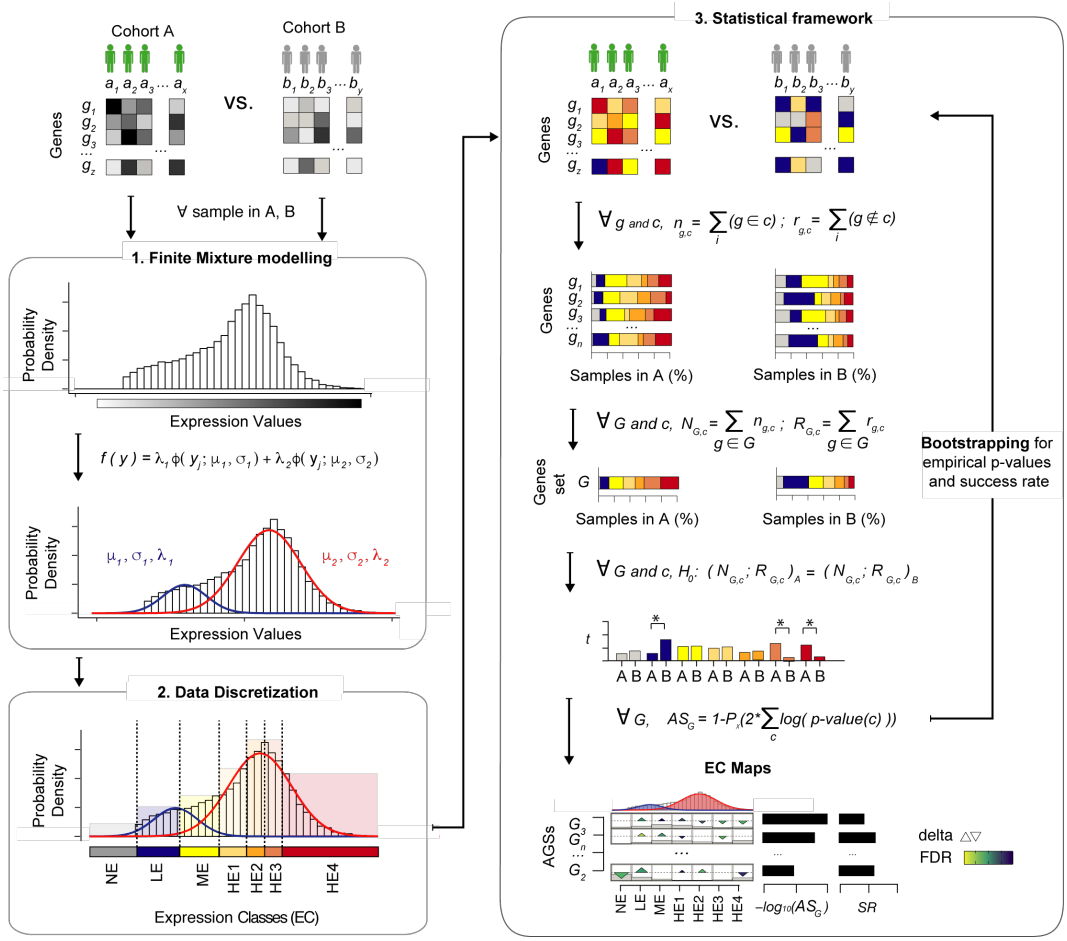
where λ is the mixing proportion, μ and σ are the mean and the standard deviation, respectively (16). To estimate the parameters μ and σ of the two components the method applies the

Expectation–Maximization (EM) algorithm (16,17). The algorithm runs iteratively until the maximum likelihood of the parameters of the two components is reached.

To ensure a consistent subdivision with the overall expression profile, we implemented an additional heuristic step. In particular, GSECA requires the mean of the first component (i.e. highly expressed genes) to be greater than the mean of the second one (i.e. lowly expressed genes). The EM step is repeated until the condition is satisfied. Besides providing an estimate of the Gaussian components, the mixture model calculates the posterior probabilities of the component membership of the mixture (16). Thus, GSECA measures the probabilities τ_1 and τ_2 of each gene to belong to the two distributions defined by the two components. Therefore, the FMM step of the algorithm accounts for the biological requirement of taking into account the bimodality of transcript populations in cells to the aim of GSA. Gene expression levels are assumed to be normalized prior to the FMM procedures with any of the suitable methods available for RNA-seq - from simple library size and transcript length accounting methods like Transcripts per Millions (TPM) or Fragments per Kilobase per Millions (FPKM), to more sophisticated ones accounting for RNA composition biases like Trimmed Mean of M-values (TMM) or Relative Log-Expression (RLE) (18). Moreover, the data are log-transformed in order to be variance-stabilized.

2. *The DD of expression values into ECs* (Figure 2.1, Step 2). To increase the signal-to-noise ratio GSECA converts the continuous measurements of expression level into discrete values. It has been recently shown that the application of DD procedures (i.e. the division of numerical features into a limited number of non-overlapped intervals) improves the accuracy of machine learning algorithms for Big Data analysis (19,20). Using the information derived from the FMM, GSECA defines seven categorical ECs and assigns each gene to the corresponding EC. Seven is the minimum number of classes that (i) ensures the minimal information loss between the discrete and the continuous expression profiles and (ii) provides an adequate distribution of genes among classes. For each sample, genes are considered as (i) not expressed (NE), or not detected, if their expression level (*e.g.* FPKM) is smaller than 0.01; (ii) lowly expressed (LE) if the probability τ_2 of belonging to the second component of the mixture is greater than 0.9; (iii) highly expressed (HE) if the probability τ_1 of belonging to the first component is greater than 0.9; or (iv) medium expressed (ME) if both the probabilities τ_1 and τ_2 are <0.9. To ensure an adequate distribution of genes among expression classes (ECs), thus a similar degree of statistical power for the subsequent tests performed for all classes, and retain as much information from the original continuous attribute as possible, HE genes are further divided accordingly to the percentiles of the expression level distribution defined by the first Gaussian component. In particular, for each sample, HE genes were assigned to (i) the first class of high expression (HE1) if their expression level is less than or

equal to the 25th percentile of the distribution of HE genes; (ii) the second class of high expression (HE2) if their expression level ranges between the 25th and the 50th percentile; (iii) the third class of high expression (HE3) if their expression level falls between the 50th and the 75th percentile; or (iv) the fourth class of high expression (HE4) if their expression level is greater than or equal to the 75th percentile. In this way, the DD step satisfies the statistical requirement of providing an equal degree of statistical power to the ECs.



I - Figure 2. 1: Schematic representation of GSECA algorithm. GSECA requires as input normalized gene expression data of two groups of samples A= {a1, . . . , an} and B = {b1, . . . , bn}, and a list of gene sets G = {G1, . . . , Gn }. The algorithm proceeds through three sequential steps: (i) the sample-specific finite mixture modeling of gene expression distribution; (ii) the sample-specific discretization of expression values into seven categorical expression classes and (iii) the statistical identification of altered gene sets (AGSs) obtained by comparing the cumulative proportion of genes of a gene set in each EC between the two cohorts using a Fisher's exact test. The expression perturbation is summarized into an association score (AS), corrected with two bootstrapping procedures for false discoveries (empirical P-value) and different sample sizes of the cohorts (success rate, SR). The AGSs are visualized as EC maps. The EC maps display the difference of the cumulative proportion of the genes of a gene set in the seven ECs between the two cohorts as triangles, whose sizes are proportional to such difference. The upper and the lower vertex of the triangles represent enrichment and depletion in cohort A as compared to B, respectively. ECmaps depict the proportion N of genes in the gene set in each EC as grey bars. GSECA orders AGSs accordingly to their AS, thus obtaining the list of the most altered processes associated with the phenotype of interest.

3. *The statistical framework for the identification of AGSs* (Figure 2.1, Step 3). The statistical analysis is thus shifted from a gene-perspective to a class-perspective: the algorithm evaluates whether the expression pattern of the genes in a gene set shows a significant displacement across the ECs in the samples of interest as compared to controls, thus suggesting a causal relationship between the condition and the phenotype.

   Given two cohorts *A* and *B*, for each gene *g* and each EC *c*, the number of samples in which *g* is and is not assigned to the class *c*, *n* and *r*, respectively, are calculated for the two cohorts:

$$\forall g \text{ and } c, \;\; n_{g,c} = \sum_i g \in c \;\; ; \quad r_{g,c} = \sum_i g \notin c \;\; ; \tag{2}$$

   where *i* are the samples in cohorts *A* and *B*. For each gene set $G = \{g_1, \dots, g_n\}$, the cumulative number of samples with genes of *G* that are and are not in each expression class across samples of *A* and *B*, *N* and *R*, respectively, are computed as follows:

$$\forall G \text{ and } c, \;\; N_{G,c} = \sum_{g \in G} n_{g,c} \;\; ; \quad R_{G,c} = \sum_{g \in G} r_{g,c} \;\; ; \tag{3}$$

   To determine whether cohort *A* is enriched or depleted of genes of a gene set *G* in an EC c as compared to cohort *B*, GSECA implements a two-tailed Fisher's Exact test. In particular, GSECA tests the null hypothesis that the cumulative proportions of genes of a gene set in each EC across samples are not different between *A* and *B*:

$$\forall G \text{ and } c, \;\; H_0 : \; (N_{G,c}; R_{G,c})_A = \;\; (N_{G,c}; R_{G,c})_B \tag{4}$$

   As a result, all seven ECs are characterized by a *P*-value representing the alterations (i.e. enrichment or depletion) of expression in the gene set. Given the contingency table defined by *N* and *R* for the two cohorts, the algorithm simulates the table under two independent binomial distributions and performs a two-tailed Fisher's Exact test. $R_{G,\,c}$ is evaluated considering all genes in the gene set that are not in the EC, regardless their class membership. Therefore, all statistical tests perform independent evaluations of the null hypothesis. In the case of multiple gene sets, the P-value of each comparison is corrected for false discoveries using either the Bonferroni or the Benjamini & Hochberg method, respectively, as defined by the user.

   Since GSECA tests the overrepresentation of genes in each EC independently from the other ECs , to quantify the degree of expression perturbation in each gene set *G* between the two cohorts *A* and *B*, the *P*-values of the seven expression classes are combined using the Fisher's method into one goodness-of-fit (χ^2) statistic, to obtain the Association Score (*AS*) (21):

$$\Psi = \; -2 \sum_c \log\big(p(c)\big) \tag{5}$$

$$AS\,(G) = P_{comb} = 1.0 - P_{\chi^2_{2k}}(\Psi) \tag{6}$$

where ψ is the combined test statistic and χ^2_{2k} is a Chi-squared distribution with *2k* degrees of freedom (*k* = number of ECs), *P* is the *P*-value and *c* is the expression class.

To calculate the significance level of the *AS* a bootstrapping procedure (random sampling with replacement) is implemented as previously described (22)(29). For 1,000 times, sample labels are shuffled and the AS is calculated for all gene sets. At the end of all iterations, for each gene set, the empirical *P*-value (p_{emp}) is measured as the number of times the AS is smaller than the observed one:

$$p_{emp}(AS_G) = \frac{1 + \left(\sum_i AS_{G,i} < AS_G\right)}{1 + \#iteration} \quad (7)$$

Finally, in case the sample sizes differ substantially between cohort *A* and *B*, a bootstrapping procedure (random sampling with replacement) is implemented to measure the success rate (SR). The algorithm down-samples the larger cohort to reach the sample size of the smaller cohort randomly 1,000 times and repeats the analysis at each iteration. At the end of all iterations, for each AS, the SR, or the proportion of significant enrichments (*P*-value < 0.01, two-tailed Fisher's Exact Test) over the total number of comparisons is calculated, as previously described in (23).

At the end of the procedure, GSECA summarizes the results for the identified AGSs in both tabular form and graphically as a heatmap, namely expression class maps (i.e. EC maps), providing an overview of the variation of expression of each gene set across the seven classes between the two cohorts (Figure 2.1).

2.2 Performance evaluation

Once the method was designed, we proceeded to assess its performance in detecting AGSs in comparison with other available GSA algorithms, employing both simulated and real data sets (which will be described in Chapter 3). In real-life systems, genes are differentially expressed with a certain degree of fold change (FC) (i.e. amplitude in up-/down-regulation) and dispersion (i.e. a measure of within-group variability, or IH) between two groups of samples (24). GSECA has been developed to dissect the contribution of IH, and thus of dispersion, in large cohorts of samples and detect the truly AGSs. To understand how well our method achieves this goal, we performed an extensive evaluation of GSECA in comparison to the others GSA methods, estimating both type I error rate and statistical power by means of *in silico* simulations.

2.2.1 GSA algorithms for comparative tests

For each analysis, we compared GSECA results with those of seven different 'state-of-art' methods: GSEA (7), GSVA (12), ssGSEA (25), Z-Score (26), PLAGE (27), ROAST (28)and Globaltest (29). We run the GSEA algorithm the using GSEA.1.0.R script available from the Broad

Institute website (http://www.broadinstitute.org/gsea). Implementations of GSVA, Z-Score, PLAGE, and ssGSEA methods were available in the R package GSVA. We run all four methods as previously described in (12), using a Poisson kernel to fit RNA-seq expression data. data. Implementations of Globaltest and ROAST were available from the R/Bioconductor package EnrichmentBrowser (30), and we run them using the *sbea* function with default parameters. We considered gene sets with corrected *P*-value ≤0.1 as significantly associated with the phenotype of interest.

To correct for false discoveries due to an unbalanced sample size of the cohorts, we implemented a measure of the SR for each comparison in a similar fashion to GSECA (see paragraph 2.2.1):

- For each method, the larger cohort is down-sampled to reach the sample size of the smaller cohort randomly 1,000 times and the analysis is repeated at each iteration. At the end of all iterations, for each comparison, the proportion of significant enrichments (*P*-value < 0.05) over the total comparisons is calculated.

2.2.2 Simulated RNA-seq data

We generated *in silico* random read counts using a Negative Binomial (NB) distribution. This model has been shown to reasonably capture biological and technical variability of RNA-seq experiments (6,31). Specifically, we modelled the number of raw reads of a gene *i* in a sample *j* as random variable Y_{ij} with NB distribution:

$$Y_{ij} \sim NB\left(mean = \mu_{ij}, size = r_{ij}\right) \tag{8}$$

with

$$E\left(Y_{ij}\right) = \mu_{ij}; Var\left(Y_{ij}\right) = \mu_{ij}\left(1 + \frac{\mu_{ij}}{r_{ij}}\right) \tag{9}$$

where μ_{ij} and $\phi_{ij} = \frac{1}{r_{ij}}$ are respectively the mean count and dispersion parameter of gene *i* in sample *j*. Dispersion describes how much the variance deviates from the mean of the read count of a gene across samples.

We derived parameters of the NB distribution from real data (specifically, from the 498 samples of the TCGA Prostate Cancer dataset, which will be extensively described in the following chapters) using the edgeR R package (31). To avoid confounding effects due to possible outliers while estimating the mean and dispersion parameters, we excluded samples with library size smaller than 40 million reads and genes not detected in all samples. We then reduced library size of all samples to that of the sample with minimum library size by binomial thinning and estimated dispersion parameter of each gene by an empirical Bayes method based on weighted conditional maximum likelihood (*estimateDisp* function in edgeR) (31).

2.2.3 Type I error rate estimation

In statistical hypothesis testing, type I error is the incorrect rejection of a true null hypothesis (often referred to as "false positive" finding). To measure the type I error rate, we generated read counts for N samples and 1000 gene sets of equal size P in the condition of no differential expression, as previously proposed in (32). Then, for each gene set, we tested the null hypothesis of no difference between the two cohorts, and we computed the type I error rate as the fraction of false positive predictions (P-value < 0.05) over the total number of gene sets tested. To examine the effects of sample and gene set sizes, we defined two tunable parameters:

1) N = sample size;
2) P = gene set size.

We ran the analysis under different parameter settings of N (60, 150, 300, 500) and P (25, 50, 100, 300), repeating each simulation ten times to obtain more stable results. GSECA resulted in being the most conservative approach, with the lowest type I error rate (average median = 0.002) as compared to the other approaches (average median = 0.05, Figure 2.2).

The conservativeness of GSECA is due to the conservativeness of the Fisher's Exact test (FET) (33) that are combined into the AS. Each FET depends on to the cumulative number of genes in the gene set in the EC across samples of the cohorts (see paragraph 2.2.1). As the sample size grows, the ability of the test to detect a small variation with high specificity and sensitivity increases. As a consequence, combining conservative FET P-values using a logarithmic scale (i.e. Fisher's Method) results in small ASs. For this reason, GSECA accounts for false positives better than the other GSA algorithms. Furthermore, GSECA specificity was not influenced by the sample and gene set sizes, remaining constant even in case of large cohorts and large gene sets (Figure 2.2).

2.2.4 Statistical power evaluation

The statistical power of a hypothesis test is the probability that the test rejects a false null hypothesis when the alternative is true: it is a measure of the probability to avoid type II errors (often referred to as "false negative" findings). To assess the statistical power of each GSA algorithm (i.e. the probability to detect an AGS when the gene set is truly altered), we tuned the parameters of the NB distribution to model several controlled settings of differential gene set expression between two cohorts A and B. To do so, besides the varying sample (N) and gene set (P) sizes, we introduced four additional tunable parameters, as previously proposed in (32):

1) β = the proportion of gene sets that contains truly differentially expressed (DE) genes;
2) γ = the percentage of genes that are truly DE in each gene set;
3) FC = the fold change in gene counts between the two cohorts;
4) D = scaling factor controlling the estimated dispersion parameter of the NB distribution.

By means of adjusting this set of parameters, we implemented two kinds of simulations:

1.1 *The FC study* (Figure 2.3, 2.5), which models the contribution of fold changes in differential gene set expression between two cohorts A and B by tuning of the FC parameter, with no changes in the estimated dispersion (D =1):

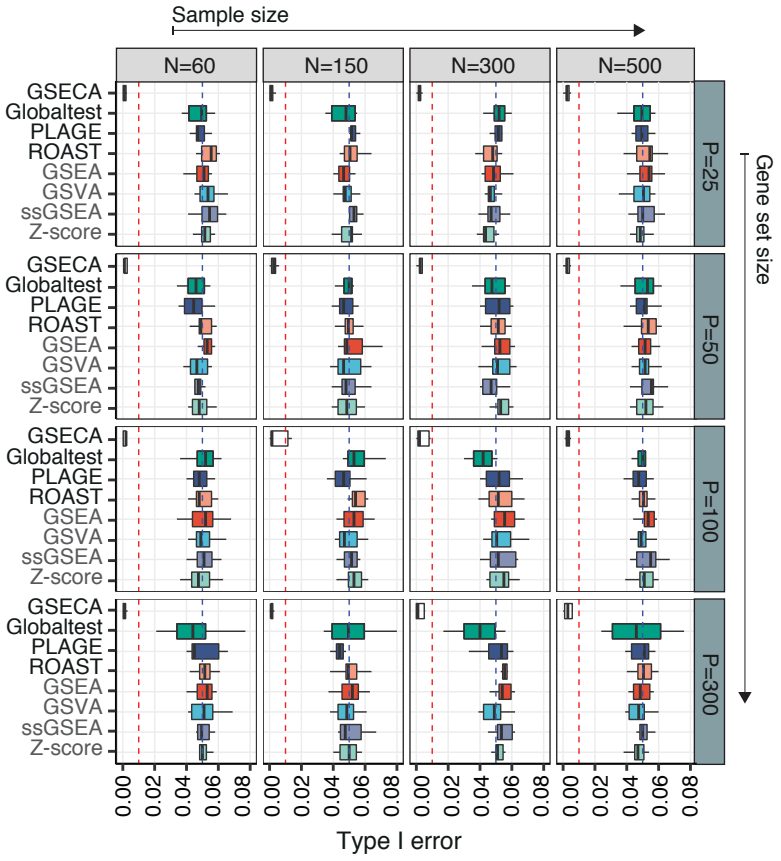
- $cohort\ A$: $NB(FC \cdot \mu_i; \varphi_i)$ $cohort\ B$: $NB(\mu_i; \varphi_i)$

We used different parameter settings for relatively small and large gene sets for cohorts of an increasing size. In particular, we varied the set of parameters as follows:

1) $\beta = \{0.05, 0.25\}$
2) $\gamma = \{0.25, 0.5\}$
3) $FC = \{1.5, 2, 2.5, 3\}$

for varying sample (N) and gene set sizes (P):

1) $N = \{60, 150, 300, 500\}$
2) $P = \{25, 100\}$



I - Figure 2. 2: Type I error rate estimation. Boxplots depicting the type I error rates for GSA methods evaluated for different settings of sample (N) and gene set sizes (P) on ten replicates. Red and blue dashed lines show the nominal α values of 0.01 and 0.05, respectively.

For all simulations, we generated S = 1,000 gene sets (median pairwise overlap of gene list < 1%) as previously proposed in (32). Briefly, we divided the N samples in two groups corresponding to distinct phenotypes A and B. Then, for each of the $(1-\beta)\cdot S$ non-DE gene sets, we generated read counts by sampling P random realizations of $NB(\mu_i;\varphi_i)$, where $1 \le i \le P$ for both cohort A and B. Conversely, for each of the $\beta \cdot S$ gene sets having truly DE genes, we sampled:

- $\frac{P}{2}$ random realizations of $NB(\mu_i;\varphi_i)$ and $NB(FC\cdot\mu_i;\varphi_i)$ under cohort A and B respectively, for $1 \le i \le \gamma \cdot \frac{P}{2}$
- $\frac{P}{2}$ random realizations of and $NB(FC\cdot\mu_i;\varphi_i)$ and $NB(\mu_i;\varphi_i)$ under cohort A and B respectively, for $\gamma \cdot \frac{P}{2} < i \le \frac{P}{2}$

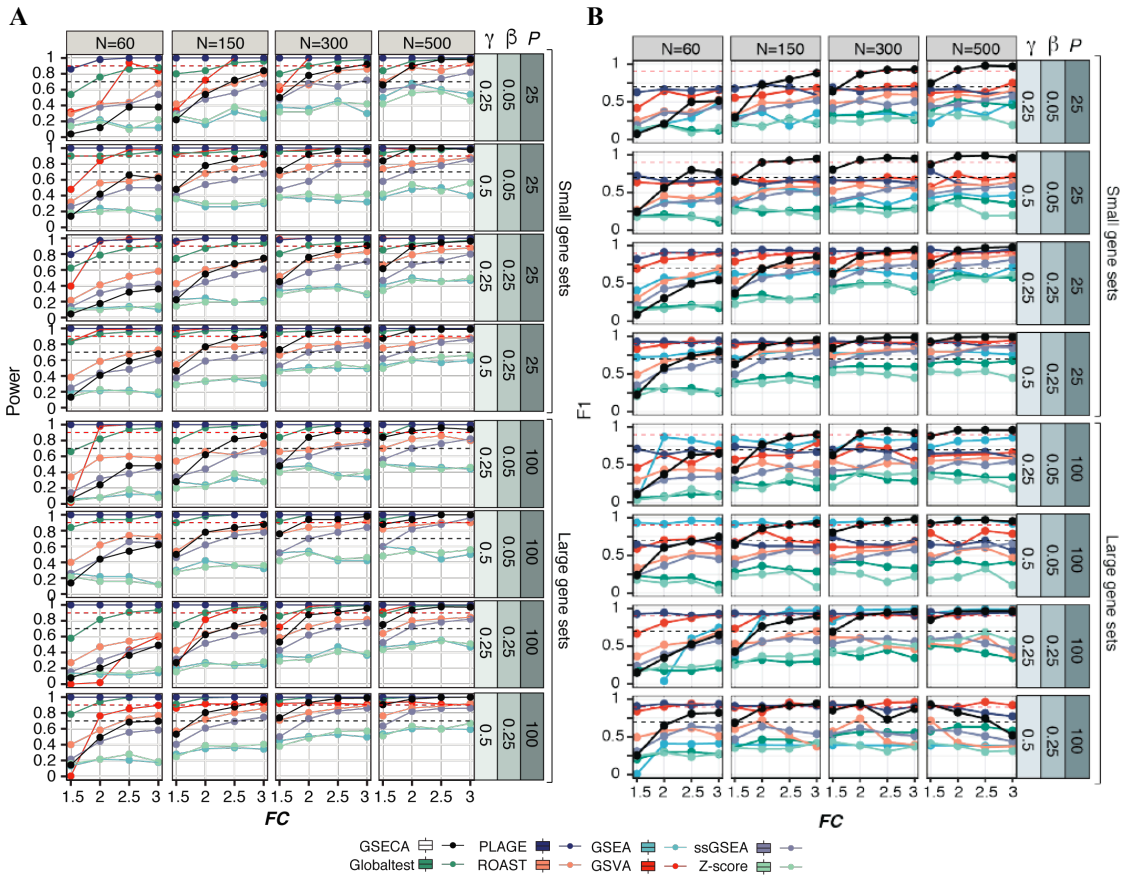
In this way, half of the $\gamma \cdot P$ DE genes in each gene set are up-regulated and half are down regulated between the two cohorts.

For each gene set, we assessed power by testing the null hypothesis of no differential expression between cohorts for all methods:

- $null\ H_0: F_A = F_B \quad alternative\ H_1: F_A \neq F_B$

Moreover, to account for the different specificity of GSA methods, we measured the F1 score, a performance evaluation metric that provides a harmonic mean of the precision and recall (34). F1 score takes both false positives and false negatives into account and it is more useful than accuracy in case of an uneven class distribution (i.e. truly AGSs and invariant gene sets, respectively) (34). For each method and simulation, we calculated precision (or positive predicted value) and recall (or sensitivity) as the fraction of true AGSs over the total amount of correctly predicted gene sets and the fraction of true AGSs over the total number of AGSs, respectively.

The FC study modeled the condition of having homogeneous expression changes across samples, driven by the degree of FC in gene set expression. In this condition, the power of GSECA increased with sample size (N) and FC values, while was not affected by the gene set size (P) and changes in the percentage of DE genes in the gene set (β and γ), as reported in Figure 2.3 and Figure 2.4. In particular, GSECA showed a power, or sensitivity, higher than 70% for medium and large sample sizes ($N \ge 150$) under different parameter settings, similar to those of the other self-contained approaches (Figure 2.3A and 2.5A). Conversely, for small sample size (N = 60), other GSA methods showed a higher power than GSECA. Overall, we noticed that GSECA predictions showed a better trade-off between precision and sensitivity (F1 score > 0.7) than all other methods even for subtle changes in gene expression for small gene set sizes (i.e. = 0.05 and P = 25, Figure 2.3A and 2.3B), reflecting the high specificity of GSECA in detecting truly AGSs.



I - Figure 2. 3: The FC simulation study. **A**. Scatter plots depicting the statistical power of each GSA algorithm at the increase of FC between cohorts for different settings of sample size *N*, gene set size *P*, the proportion of gene sets containing differentially expressed genes β, the percentage of DE genes in each gene set γ. **B.** Scatter plots depicting the F1 score of each GSA algorithm at the increase of FC between cohorts for different settings of sample size *N*, gene set size *P*, the proportion of gene sets containing differentially expressed genes β, the percentage of DE genes in each gene set γ.

1.2 *The dispersion study* (Figure 2.4-2.5), which models the contribution of IH, or within-group variability, in differential gene set expression between two cohorts A and B by fixing a small or no FC parameter and tuning the D parameter:

- $cohort\ A$: $NB(FC \cdot \mu_i; D \cdot \varphi_i)$ $cohort\ B$: $NB(\mu_i; \varphi_i)$

Again, we varied the parameter settings as follows:

1) $\beta = \{0.05, 0.25\}$
2) $\gamma = \{0.25, 0.5\}$
3) $FC = \{1, 1.1, 1.25\}$
4) $D = \{1.5, 2, 3, 5, 7, 10\}$

for varying sample (*N*) and gene set sizes (*P*):

1) $N = \{60, 150, 300, 500\}$
2) $P = \{25, 100\}$

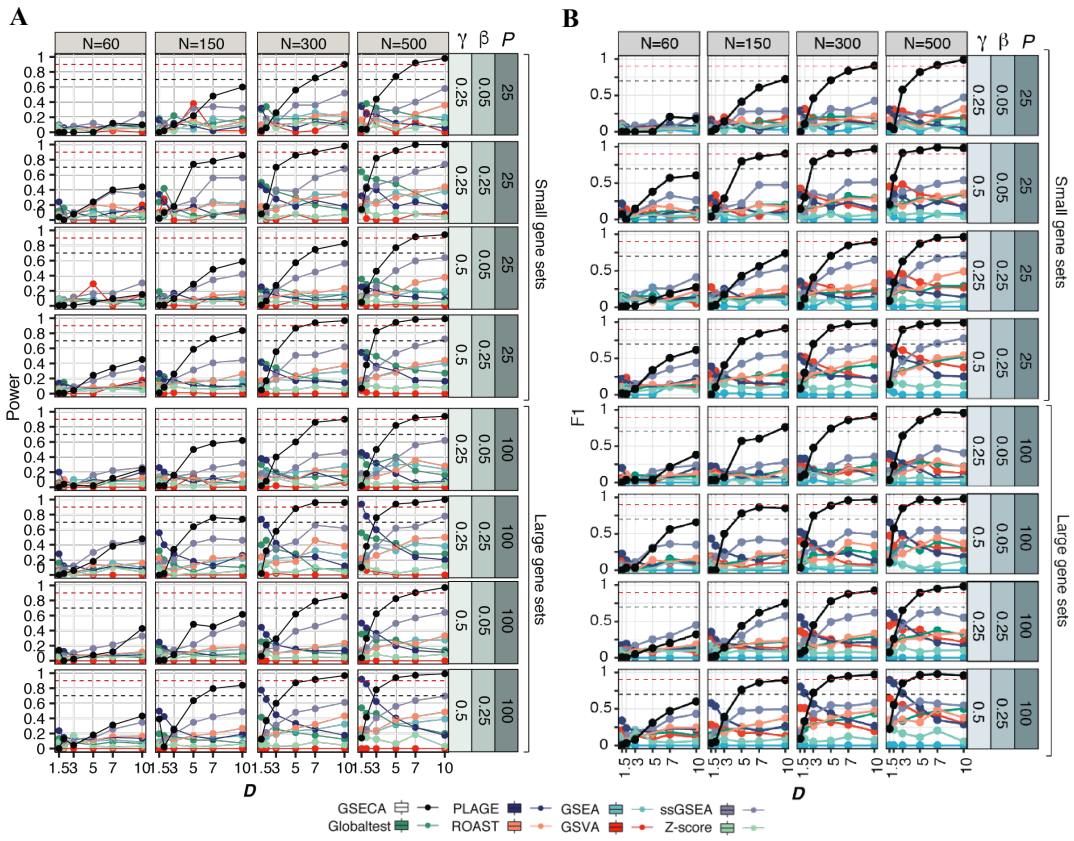
For all simulations, we generated S = 1,000 non-overlapping gene sets. We divided the N samples in two cohorts A and B. Then, for each of the $(1-\beta)\cdot S$ non-DE gene sets, we generated read counts by sampling P random realizations of $NB(\mu_i;\varphi_i)$, where $1 \le i \le P$ under both cohorts. In contrast, for each of the $\beta \cdot S$ gene sets having truly DE genes, we sampled P random realization of $NB(\mu_i;\varphi_i)$ and $NB(FC\cdot\mu_i;D\cdot\varphi_i)$ under cohort A and B respectively, for $1 \le i \le \gamma \cdot P$ (32). For each gene set, we assessed power by testing the null hypothesis of no differential expression between cohorts for all methods, and measured the F1 score.

The dispersion study modeled the effect of having heterogeneous expression changes across samples, driven by the degree of dispersion (D), or IH. In this condition, GSECA outperformed the other GSA methods, with an exponential-like grow of its statistical power at the increase of the D parameter (Figure 2.4A). Moreover, the power of GSECA increased with sample size (N), while the other methods were not affected. Even to a lesser extent, ssGSEA performed similarly to GSECA in handling heterogeneity (Figure 2.4A). Comparably to GSECA DD approach, ssGSEA brings expression profiles to a common scale collapsing the range of possible gene expression (32). In doing so, ssGSEA reduces the noise of IH (i.e. genes with similar expression levels will have the same rank), increasing its power to detect truly AGSs.

GSECA achieved the highest F1 scores, underlining its high sensitivity and specificity in case of heterogeneous gene expression (Figure 2.4B). These results did not considerably change for small variation of FC values or with gene set sizes (Figure 2.6 and Figure 2.7)

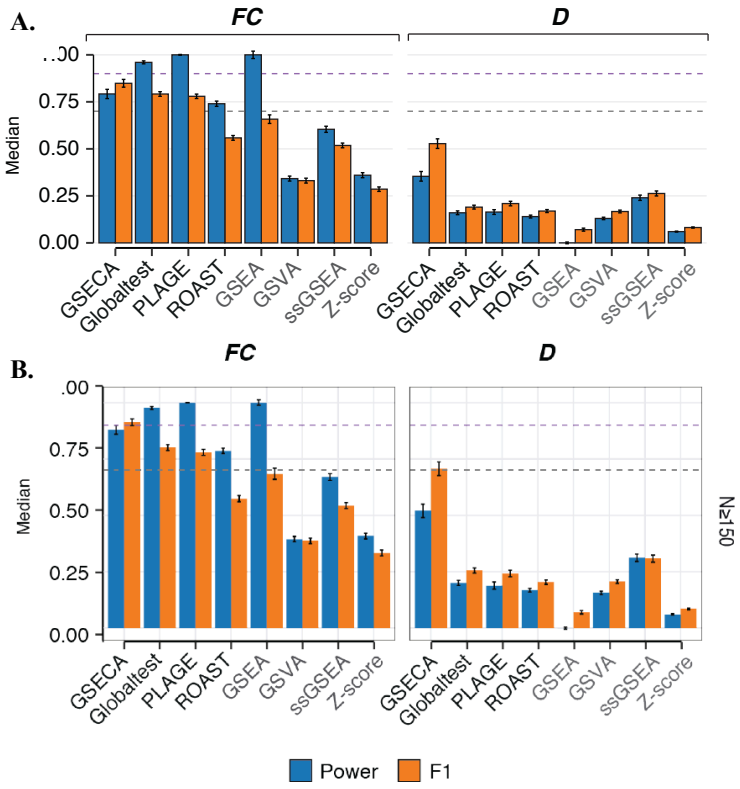
Taken together the results of the *in-silico* simulations for FC and dispersion studies, we can conclude that:

- GSECA has a high sensitivity, proper of self-contained tests (32), of identifying truly AGSs in presence of FC variations between cohorts (Figure 2.5A). Moreover, compared to the other GSA approaches, GSECA has the best balance between sensitivity and specificity, showing the highest F1 score (Figure 2.5B).
- GSECA is the most powerful GSA approach, among the tested ones, to treat dispersion, and thus IH, in gene expression between phenotypes (Figure 2.5A). The results of the simulation studies show that the performances of GSECA are enhanced in case of large cohorts (i.e. $N \ge 150$, Figure 2.5B).

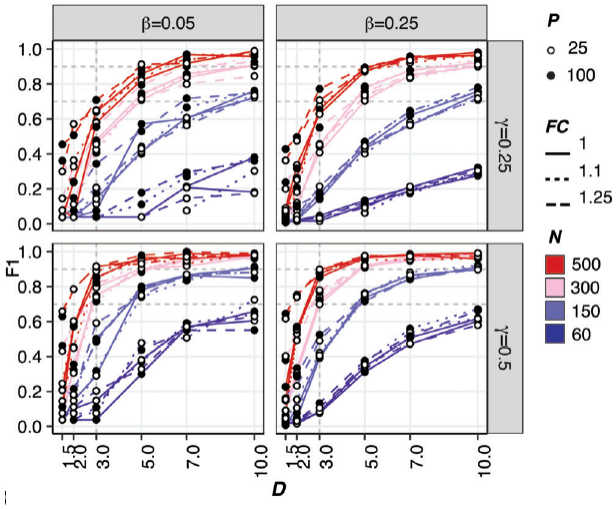


I - Figure 2. 4: The dispersion study. **A.** Scatter plots depicting the statistical power of each GSA algorithm at the increase of dispersion factor *D* at a fixed *FC* of 1.1 between cohorts for different settings of sample size *N*, gene set size *P*, the proportion of gene sets containing differentially expressed genes β, the percentage of DE genes in each gene set γ. **B.** Scatter plots depicting the F1 score of each GSA algorithm at the increase of dispersion factor *D* at a fixed *FC* of 1.1 between cohorts for different settings of sample size *N*, gene set size *P*, the proportion of gene sets containing differentially expressed genes β, the percentage of DE genes in each gene set γ.

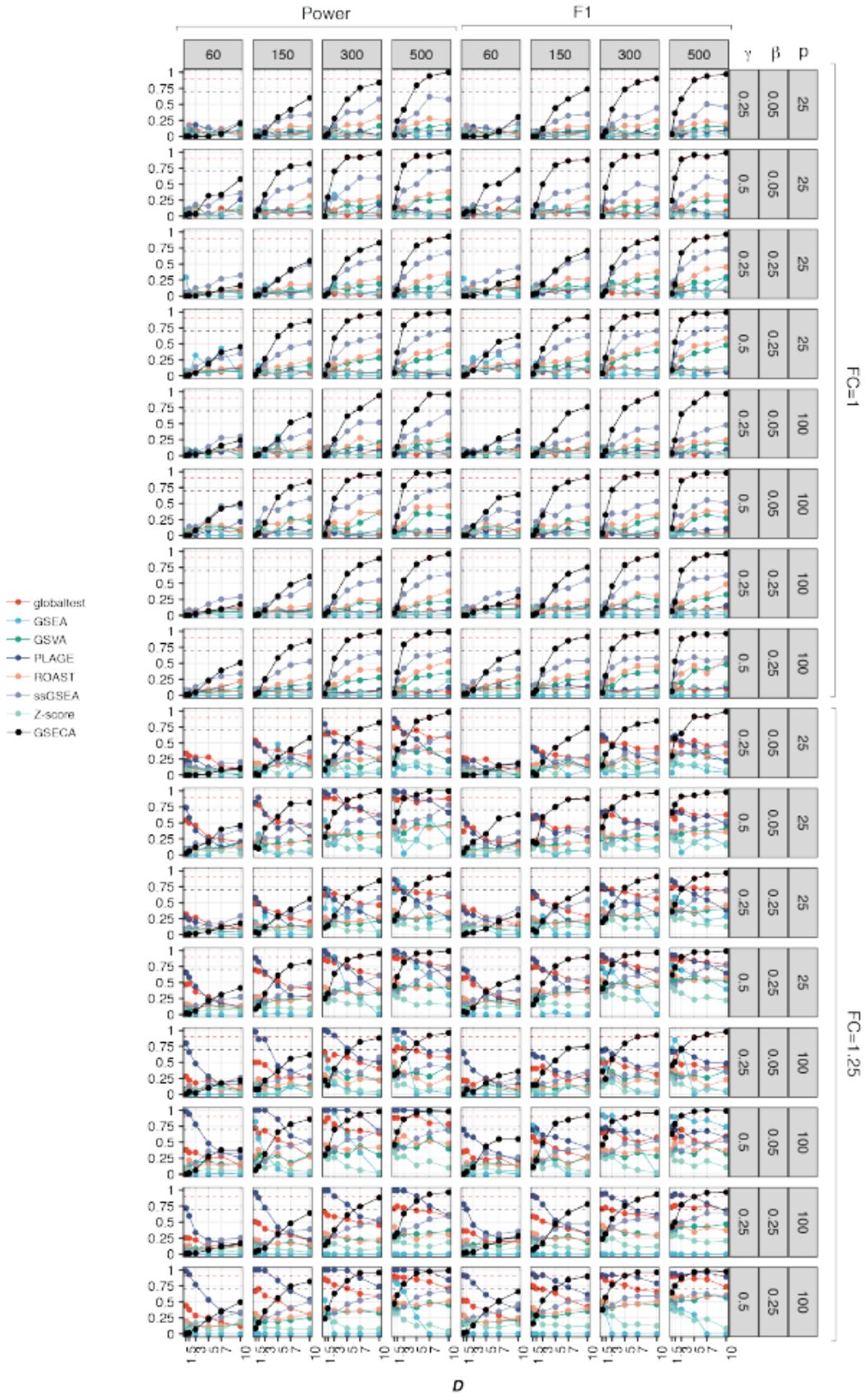
In both FC and dispersion studies, the statistical power of GSECA increased with sample size. This is a consequence of the DD process, where it is expected that small sample sizes might be not sufficient to estimate the correct distribution of data (34). When the IH noise between cohorts is negligible and the cohort size is small, GSECA requires strong FC differences to reach adequate power. In contrast, when IH noise between cohorts is relevant, GSECA performs better than any of the tested GSA methods in terms of both sensitivity and specificity, thus being the most powerful approach for detecting truly AGSs in presence of heterogeneous gene expression changes.



I - Figure 2. 5: Performance summary of GSA methods. **A.** Bar plots representing the median values of statistical power and F1 score measured for all GSA methods in all simulations for the FC and dispersion studies. Grey and purple dashed lines represent values of 0.7 and 0.9, respectively. Black error bars depict standard errors. **B.** Bar plots representing the median values of statistical power and F1 score measured for all GSA methods in all simulations for the FC and dispersion studies for sample sizes greater than 150. Grey and purple dashed lines represent values of 0.7 and 0.9, respectively. Black error bars depict standard errors.



I - Figure 2. 6: Performance of GSECA in the dispersion study simulations. Scatter plot showing GSECA statistical power at increasing values of dispersion parameter *D* at for different settings of *FC*, sample size *N*, gene set size *P*, the proportion of gene sets containing differentially expressed genes β, the percentage of DE genes in each gene set γ.



I - Figure 2. 7: Dispersion study simulations. Scatter plots depicting the statistical power and the F1 score of each GSA algorithm at the increase of dispersion factor *D* for increasing sample size *N* (left-to-right) and increasing *FC*, gene set size *P*, proportion of gene sets containing differentially expressed genes β, percentage of DE genes in each gene set γ (top-to-bottom).

Chapter 3

Identification of altered biological processes in PTEN loss prostate adenocarcinoma

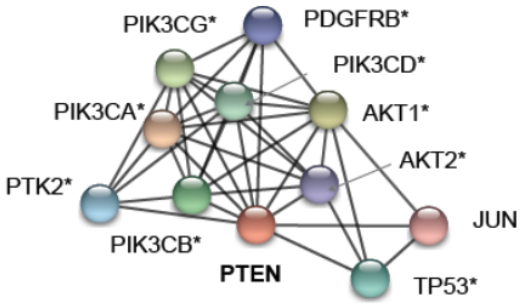
Results

3.1 The prostate cancer dataset

To assess the performance of GSECA in detecting AGSs on real datasets, we looked for a condition in which the alteration of a known biological process was expected. To this aim, we chose the Prostate Adenocarcinoma (PRAD) cancer type, available from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga), as a case study. A frequent event in prostate cancer is the somatic loss of the tumor suppressor gene PTEN (35,36), which encodes for the phosphatase and tensin homolog protein. Through its lipid phosphatase activity, the PTEN protein governs a plethora of cellular processes, including survival, proliferation, energy metabolism and cellular architecture (37). As tumour suppressor gene, PTEN acts as a negative regulator of the PI3K/AKT pathway, which is a signal transduction pathway that promotes survival and growth in response to extracellular signals. The somatic loss of PTEN results in the alteration of the PI3K/AKT signaling pathway (38) and promotes oncogenic programs (37).

To deeper characterize the relationship between PTEN and the PI3K/AKT signaling pathway, we retrieved the PTEN protein-protein interaction (PPI) network from STRING (39) (http://string-db.org) (using sources as 'textmining', 'experiment' and 'databases' as a type of evidence to measure the interactions between PTEN and other proteins). Interestingly, among the first ten top-ranked primary interactors (interaction score ≥ 0.9) of PTEN, nine genes are involved in the PI3K/AKT signaling pathway accordingly to the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Figure 3.1). Moreover, functional analysis of the PTEN interactors revealed a significant enrichment of genes in PI3K/AKT signaling pathway ($FDR \leq 0.05$).

Given these considerations, we hypothesized that stratifying human prostate adenocarcinomas (PRADs) accordingly to the somatic loss of PTEN could reveal the altered modulation of the PI3K/AKT signaling pathway, providing us the framework to test the performance of GSECA on real data, in comparison with the other selected GSA algorithms.



I - Figure 3. 1: The PTEN PPI network from STRING. First 10 top-ranked primary interactors of PTEN in the STRING PPI network.

3.1.1 Identification of PTEN loss samples

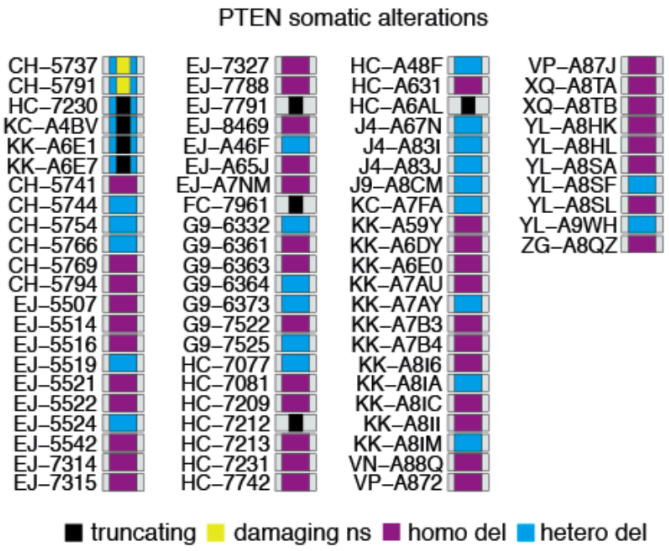
We collected genomic data for 498 PRAD samples available from TCGA. In particular, we downloaded data for somatic mutations (i.e. single nucleotide variants and small insertion/deletions (InDels)), copy number alterations (CNVs), RNA sequencing, protein expression and phosphorylation from the TCGA Data Matrix portal (Level 3, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm).

We next processed the data as previously described in (40). In particular, we considered PTEN as somatically lost in presence of at least one of the following loss of function alterations:

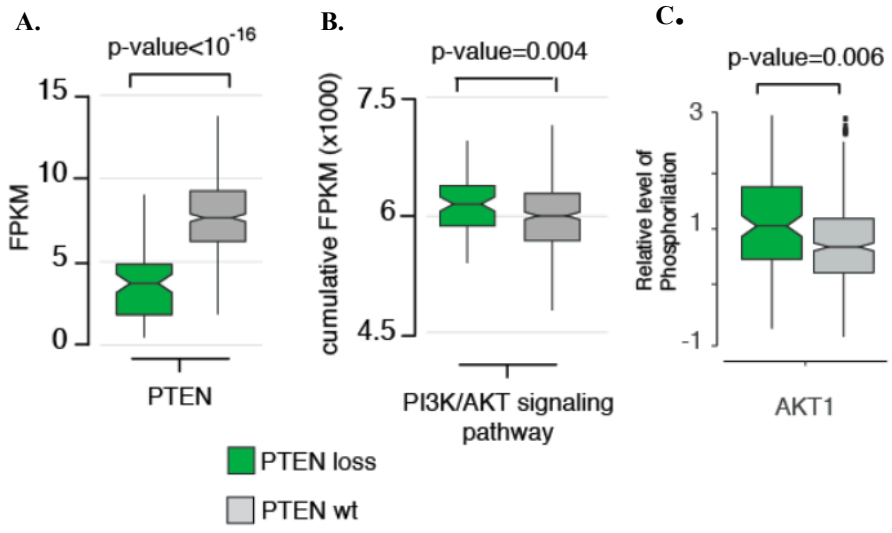
- Homozygous/heterozygous gene deletions: loss of both or single allele, assigning the copy number status as previously reported in (40).
- Truncating mutations: point mutations in DNA sequence resulting in a truncated, incomplete and non-functional protein product (i.e. stopgain, stoploss, frameshift indels).
- Damaging mutations: single nucleotide change resulting in a codon that codes for a different amino acid, with predicted damaging effects on the encoded protein. We defined damaging alterations as missense and splicing (i.e. up to two nucleotides surrounding the splice sites) mutations. Specifically, we considered missense mutations as damaging if supported by at least five out of eight function-based scores (SIFT (41), PolyPhen-2 HDIV and HVAR (42), MutationTaster (43), MutationAssessor (44), LTR (45) and FATHMM (46)) or two out of three conservation-based scores (PhyloP (47), GERP++ RS (48), SiPhy (49)). Similarly, we defined splicing mutations as damaging if supported by at least one ensemble score of dbscSNV (50).

According to these criteria, we stratified the 498 PRAD samples in two cohorts: 75 PTEN loss (PTEN-loss) and 423 PTEN wild type (PTEN-wt) tumors (*i.e.* patients without PTEN alterations). Using RNA-seq and protein data, we observed a significant lower expression of PTEN in PTEN-loss samples as compared to PTEN-wt ones (P-value < 10^{-16}, two-tailed Wilcoxon test, Figure 3.3A). Moreover, we confirmed the alteration of PI3K/AKT genes by measuring their cumulative expression levels, with a significant increase in PTEN-loss tumors (P-value = 0.004, two-tailed

Wilcoxon test, Figure 3.3B). Furthermore, we observed a significantly higher phosphorylation level of AKT1 in PTEN-loss tumours as compared to wild-type samples (*P*-value = 0.006, two-tailed Wilcoxon test, Figure 3.3C). These results show that our sample stratification, based on alterations that occur at the DNA level, produced a broad, unbalanced and heterogeneous expression dataset containing a significantly different regulation of the PI3K/AKT signaling pathway.



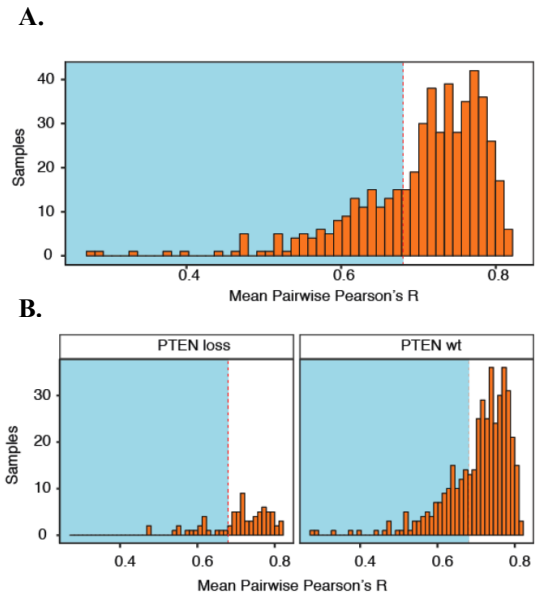
I - Figure 3. 2: Somatic alterations in PTEN loss samples. Somatic alterations (i.e. mutations and/or copy number alterations (CNVs)) affecting PTEN in the 75 PTEN loss samples.



I - Figure 3. 3: Effect of PTEN-loss sample stratification on PTEN and PI3K/AKT pathway. **A.** Boxplot distributions of PTEN normalized expression levels (i.e. FPKM) for PTEN-loss and PTEN-wt samples. **B.** Cumulative expression levels of genes involved in the PI3K/AKT signaling pathway for PTEN-loss and PTEN-wt samples. **C.** Boxplot distributions of the relative level of AKT phosphorylation, for PTEN loss and PTEN-wt samples.

3.1.2 Characterization of IH in PTEN-loss tumors

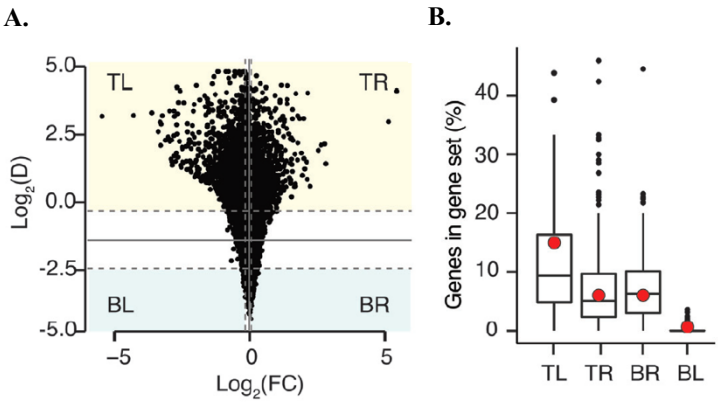
We next evaluated the level of IH of the cohorts reflected in the RNA-seq data. We first did it by correlation analyses, measuring the pairwise Pearson's correlation coefficient of the gene expression profiles and taking the average for each sample as a similarity measure. Interestingly, we found that 64% of samples had a low similarity of expression patterns with the others (Pearson's R < 0.75, Figure 3.4A and B), confirming the presence of IH in the dataset. We further characterized IH by performing differential gene expression analysis (with the DESeq2 R package (24), using default settings and filtering out genes with read counts equal to zero in all samples) between PTEN-loss and PTEN-wt cohorts and measuring the FC and dispersion of FPKM values for 19,663 protein coding genes (Figure 3.5A). We found that 56% of genes showed a reduced expression upon PTEN loss, which was reflected in the higher number of down-regulated genes (n = 631, $logFC \leq -1$ and $FDR \leq 0.1$) than up-regulated ones (n = 325, $logFC \geq 1$ and $FDR \leq 0.1$). Furthermore, out of 4830 genes with a high level of dispersion (i.e. ≥75th percentile of the dispersion distribution), 29% were activated (i.e. FC ≥ 75th percentile of the FC distribution) and 48% repressed (i.e. FC ≤ 25th percentile of the FC distribution). These findings highlight a general reduction of expression characterized by IH in the PTEN-loss cohort. This might be a possible consequence of the role of PTEN in regulating basal transcription through histones and chromatin remodelling (51).



I - Figure 3. 4: Correlation analysis of PRAD dataset. Histogram showing the distributions of mean Pearson's Correlation of pairwise comparisons, for all samples (**A.**) and for PTEN-loss and PTEN-wt samples, respectively (**B.**). Blue rectangle indicates R smaller than 0.75.

In order to perform the GSA between the two cohorts, we collected the list of 158 human gene sets of the KEGG database (52) (https://www.genome.jp/kegg/), available from MSigDb35 (version 5,

https://software.broadinstitute.org/gsea/msigdb/). This manually curated collection of human gene sets represents the current knowledge on molecular reaction and interaction networks for a wide range of biological processes and metabolic pathways. By inspecting the FC and dispersion landscape of the gene sets list between PTEN-loss and PTEN-wt tumors, we observed that the highest proportion of genes was repressed and dispersed (Figure 3.5B, TL). In particular, 15% of genes in PI3K/AKT signaling pathway were repressed and highly dispersed, suggesting a variable repression of PI3K/AKT genes across samples upon the somatic loss of PTEN (Figure 3.5B, red dots).



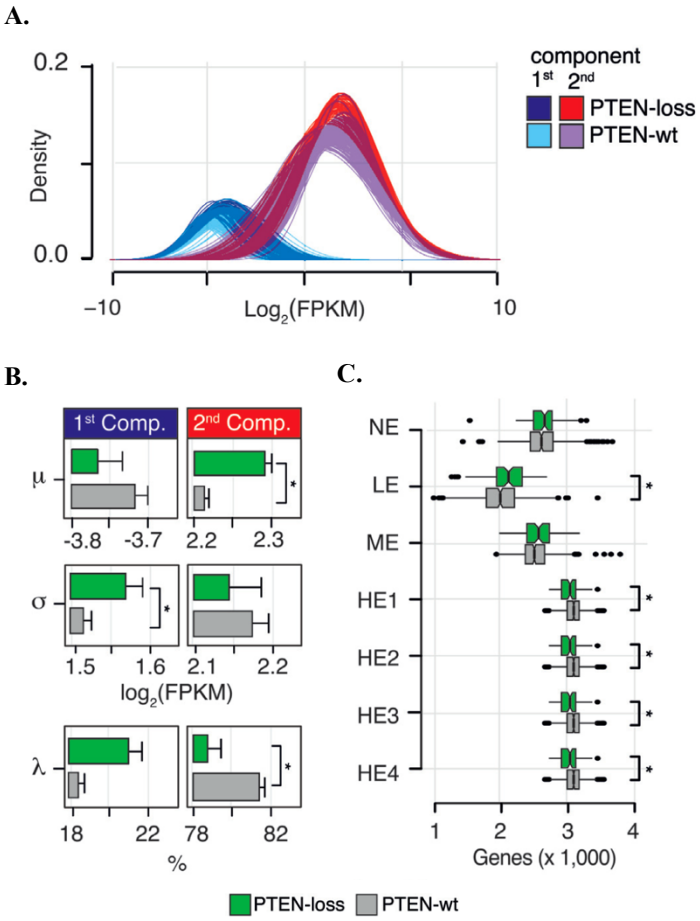
I - Figure 3. 5: The FC and dispersion landscape between PTEN-loss and PTEN-wt tumors. **A.** Scatter plot showing the log2 fold change (FC) and dispersion (D) values of all genes between PTEN-loss and PTEN wt samples. Grey lines represent the median values of FC and D. Dashed grey lines show the 25th and 75th percentile of the FC and D distributions and define four regions of expression changes (TL = top-left; TR = top-right; BL = bottom-left; BR = bottom-right). **B.** Boxplots depicting the percentage of genes of each KEGG gene set in the four regions of expression changes. Red dots represent genes in PI3K/AKT signaling pathway.

3.2 GSECA results on PTEN-loss PRAD tumors

We applied GSECA to the PRAD dataset, testing the 158 KEGG gene sets between PTEN-loss and PTEN-wt tumors. The first step of GSECA is the sample-specific FMM of gene expression levels distributions to identify the two subpopulations of lowly (first component of the mixture) and highly (second component of the mixture) expressed genes (Figure 3.6A, see Chapter 2, paragraph 2.2.1). To understand whether the FMM step captured the observed IH of the PRAD dataset (paragraph 3.1.2), we inspected the component parameters (i.e. mean μ, standard deviation σ and mixing proportion λ) obtained for the two cohorts (Figure 3.6B). We observed that, even if showing the same μ, the average σ of the first component (i.e. lowly expressed genes) was significantly higher in PTEN-loss samples as compared to PTEN-wt ones (*P*-value = 0.031, two-tailed Student's *t*-test), suggesting that PTEN-loss are more heterogeneous at low levels of expression as compared to PTEN-wt samples. Moreover, analyzing the second component, we found that the average μ was significantly higher in the PTEN-loss cohort (*P*-value = 0.003, two-tailed Student's *t*-test), with no differences in the σ parameter. Finally, comparing the mixing proportions λ, we observed a significantly lower number of genes assigned to the second component for PTEN-loss samples as compared to PTEN-wt ones (hence, a higher number was

assigned to the first component). The results of this analysis showed that the FMM step of GSECA correctly described the gene expression landscape of the PTEN-loss PRAD dataset, capturing the higher degree of IH for lowly expressed genes in PTEN-loss tumors (as previously observed in paragraph 3.1.2).

The second step of GSECA is the DD process, which groups the gene expression values into seven discrete ECs. To assess whether the DD step preserved the structure of the PRAD PTEN-loss dataset, we compared the distributions of the proportion of genes in each and expression class between PTEN-loss and PTEN-wt tumors. We found a significant increase of genes in the LE class (Bonferroni adjusted *P*-value = 0.016, two-sided Wilcoxon rank sum test) and reduction of genes in the four HE classes (Bonferroni adjusted *P*-value = 0.021, two-sided Wilcoxon rank sum test) for PTEN-loss as compared to PTEN-wt tumors. This result reflects the reduction of λ for PTEN-loss cohort detected by the FMM step, thus confirming the accuracy of the DD step in not altering the observed features of the original data.

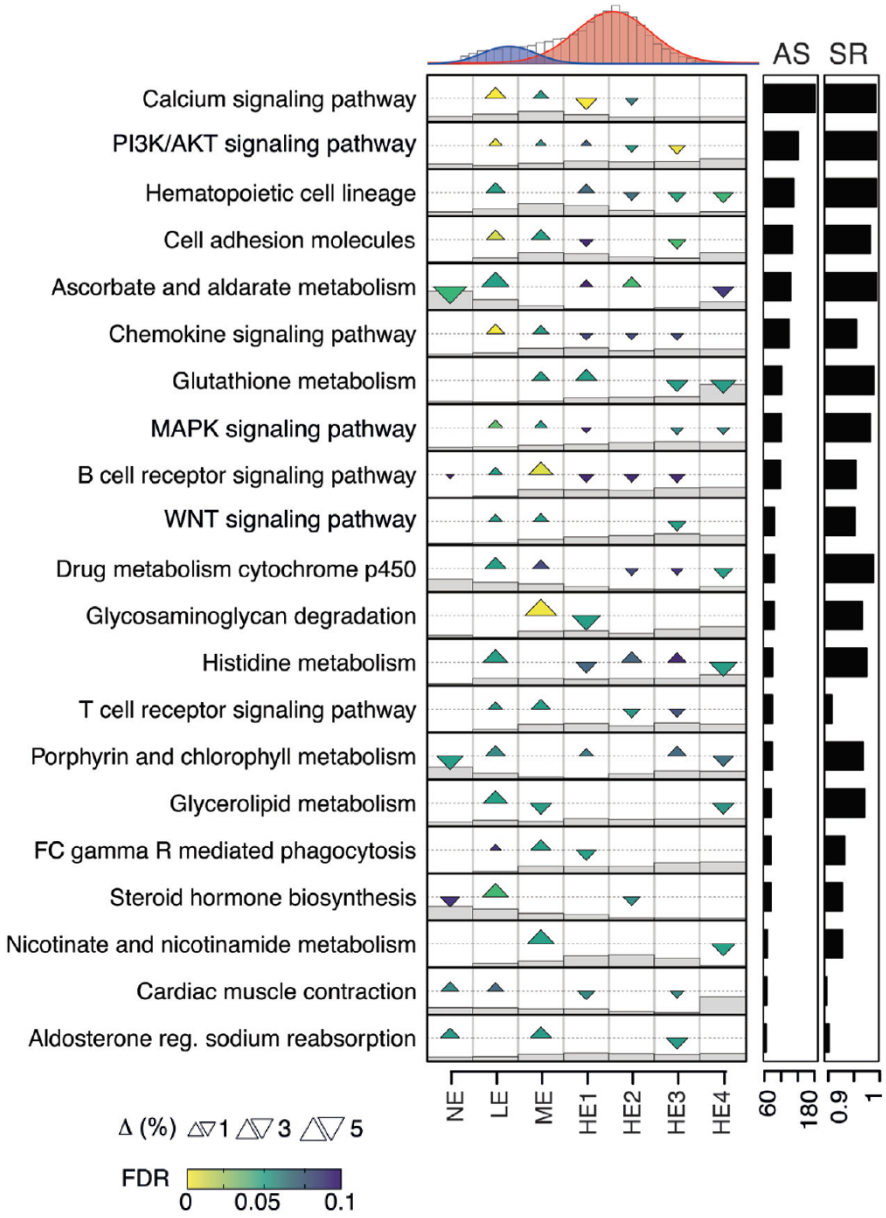


I - Figure 3. 6: GSECA FMM and DD steps on the PRAD PTEN loss dataset. **A.** Kernel density distributions of the two-component Gaussian mixtures obtained from the FMM step, for PTEN-loss and PTEN-wt tumors. **B.** Comparison of the component parameters (i.e. mean μ, standard deviation σ and mixing proportion λ) defined by the FMM between PTEN-loss and PTEN-wt tumors. Barplots depict the mean and standard deviation of each parameter, for PTEN-loss and PTEN-wt samples. Statistical test with a *P*-value < 0.05 are considered as significant (*, Student's *t*-test). **C.** Boxplot distributions representing the number of genes in each EC for PTEN-loss and PTEN-wt samples. Statistical test with a Bonferroni adjusted *P*-value < 0.05 are considered as significant (*, two-tailed Wilcoxon rank sum test).

The final step of GSECA is the statistical framework for the identification of AGSs (see Chapter 2, paragraph 2.2.1). For each of the 158 KEGG gene sets, GSECA compared the fraction of genes in each EC between PTEN-loss and PTEN-wt tumors, determining their enrichment or depletion, and summarized the results in the AS, providing for each gene set the empirical *P*-value (p_{emp}) and SR in order to avoid false discoveries and correct for unbalanced sample sizes. We found that 21 out of the 158 KEGG gene sets were significantly altered in PTEN-loss as compared to PTEN-wt samples ($AS \leq 0.01,\ p_{emp} \leq 0.001$ and $SR \geq 0.9$). The resulting AGSs were ranked on the basis of three significance metrics (i.e. AS, p_{emp} and SR) and reported in the EC map (Figure 3.7). Remarkably, GSECA identified the PI3K/AKT signaling pathway as the second top-ranked AGS, showing a significant increase in the LE, ME and HE1 classes and a significant decrease in the HE2 and HE3 classes (FDR < 0.1), thus supporting the presence of high IH for PI3K/AKT genes at low levels of expression (Figure 3.5B).

Among the remaining AGSs, GSECA identified five gene sets of signal transduction (i.e. calcium signaling, cytokine-cytokine receptor interaction, cell adhesion molecules (CAMs), MAPK and WNT signaling pathway) that are tightly connected with PI3K/AKT signaling pathway. In particular, PTEN silencing, and the subsequent alteration of PI3K/AKT pathway, impairs calcium signaling (53), alters epithelial CAMs and focal adhesion gene expression in prostate (54), alters MAPK (55) and WNT signaling cascade (56). Furthermore, GSECA detected the alteration of five immune-related processes (i.e. hematopoietic cell lineage, chemokine signaling pathway, B and T cell receptor signaling, FC gamma R mediated phagocytosis), supporting the role of PTEN in regulating the proliferation and differentiation of hematopoietic stem cell (57), controlling signaling and homeostasis in both B and T cells (58,59), and inhibiting FC gamma receptor signaling (60), as well as the role of PI3K/AKT pathway in the regulation of chemokine signaling during prostate tumorigenesis (61). Finally, GSECA highlighted the alteration of nine metabolic pathways, underlining the contribution of PTEN in metabolism control (62). GSECA also identified the alteration of cardiac muscle contraction and aldosterone-regulated sodium reabsorption gene sets. It is worth noting that the down-regulation of PTEN decreases heart muscle contractility (63) and the activation of PI3K/AKT pathway might be responsible for the alteration of aldosterone-mediated sodium transport in epithelial cells (64).

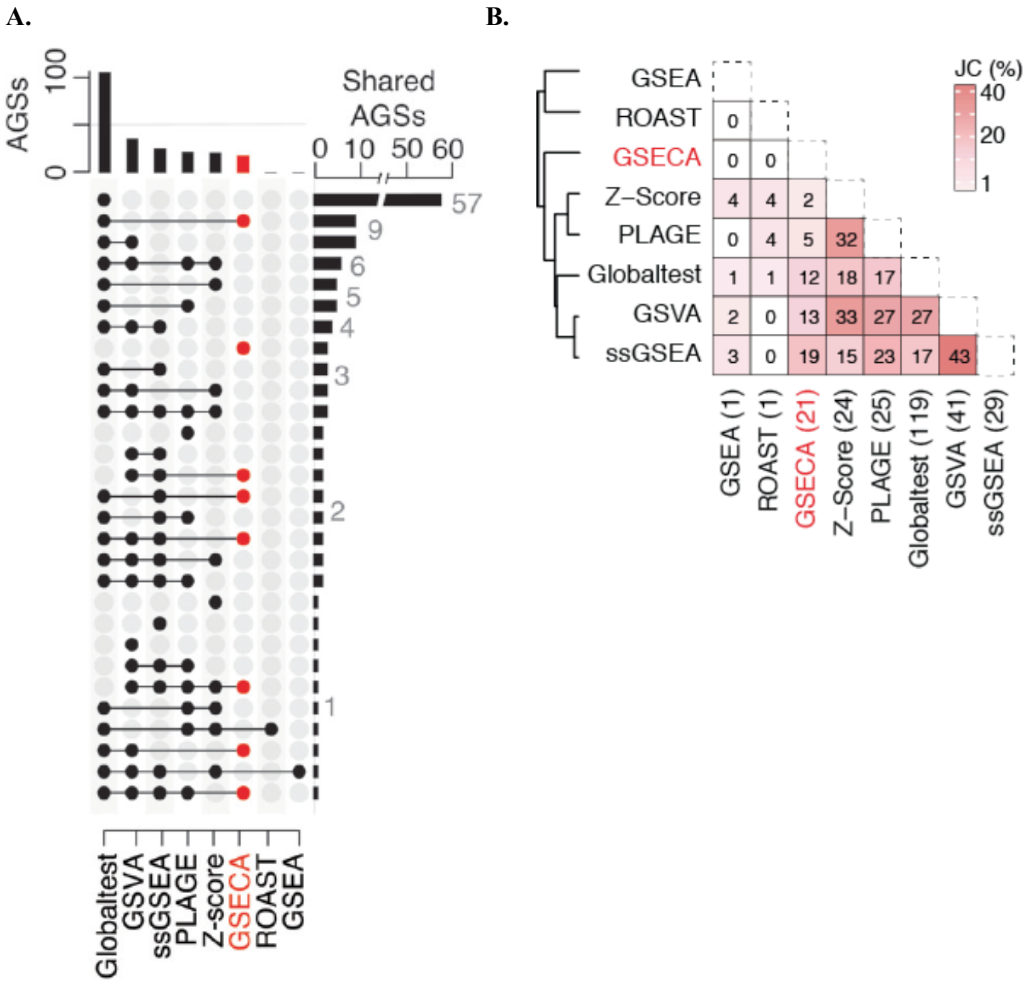
These results indicate that our method can successfully identify the alteration of biological processes (i.e. the expected alteration of PI3K/AKT signaling and related pathways) on real RNA-seq data characterized by a high degree of IH, as shown for the PTEN-loss PRAD dataset.



I - Figure 3. 7: GSECA results on the PRAD PTEN loss dataset. EC map for the AGSs identified by GSECA in PTEN-loss prostate adenocarcinoma. Triangles report the significant alteration of ECs. Direction of triangle depict enrichment (up) or depletion (down). Size of the triangles represent the proportion differences (%), colour report the FDR. Barplots on the right-side report AS and SR values, used to rank the AGSs.

3.3 Comparison with available GSA algorithms

We next compared the results of GSECA on the PRAD PTEN loss dataset with those of the selected GSA algorithms available from the bioinformatics community (see Chapter 2, paragraph 2.2.1). We found that Z-Score, PLAGE, ssGSEA identified an average of 20 AGSs, comparably to GSECA, whereas GSVA identified 41 AGSs and ROAST and GSEA detected only 1 AGS (adjusted *P*-value < 0.1, SR > 0.9, Figure 3.8A). To evaluate the concordance of results among methods, we measured the Jaccard Coefficient (JC, Figure 3.8B) between any couple of algorithms.



I - Figure 3. 8: Similarity of GSA methods' result on the PRAD PTEN-loss dataset. **A.** Overlap of AGSs in PRAD PTEN-loss samples identified by the GSA algorithms. **B.** Similarity matrix based on the Jaccard Coefficient (JC) of AGSs in PRAD PTEN-loss samples identified by the GSA algorithms.

We observed that the overall similarity among methods was low (mean JC = 12%) and that GSECA showed the highest agreement with ssGSEA (JC = 19 %). The poor concordance of the results of GSA methods has been previously reported as a consequence of the distinct statistical assumptions of each method (9). Hence, the number and type of algorithm-specific AGSs might reflect the ability of a method to handle the IH proper of the PRAD dataset.

Nevertheless, given this weak similarity, we validated the results obtained from each method using orthogonal analysis approaches (Figure 3.9). To do so, we first selected the ten top-ranked AGS identified by each method and hierarchically clustered their rankings. Next, for each gene set, we compared the GSA results with those obtained from three different analyses:

- Gene ontology (GO) analysis on the PTEN PPI interaction network retrieved from STRING(39).
- GO analysis on the list of differentially expressed genes between PTEN-loss and PTEN-wt samples, obtained with DESeq2 (24).

- Literature-based text mining analysis on published journal articles - available at the National Center for Biotechnology Information (NCBI) PubMed database (https://pubmed.ncbi.nlm.nih.gov/) - exploring the connection between PTEN and the identified AGSs. To do so, for each KEGG gene set, we inspected abstracts of published articles for the co-occurrence of keywords such as PTEN and the gene set nomenclature, using the R package RISMed (https://cran.r-project.org/web/packages/RISmed).

The hierarchical clustering of the top-ranked AGSs revealed the presence of five clusters, composed by three gene sets groups based on the methods of origin (Figure 3.9):

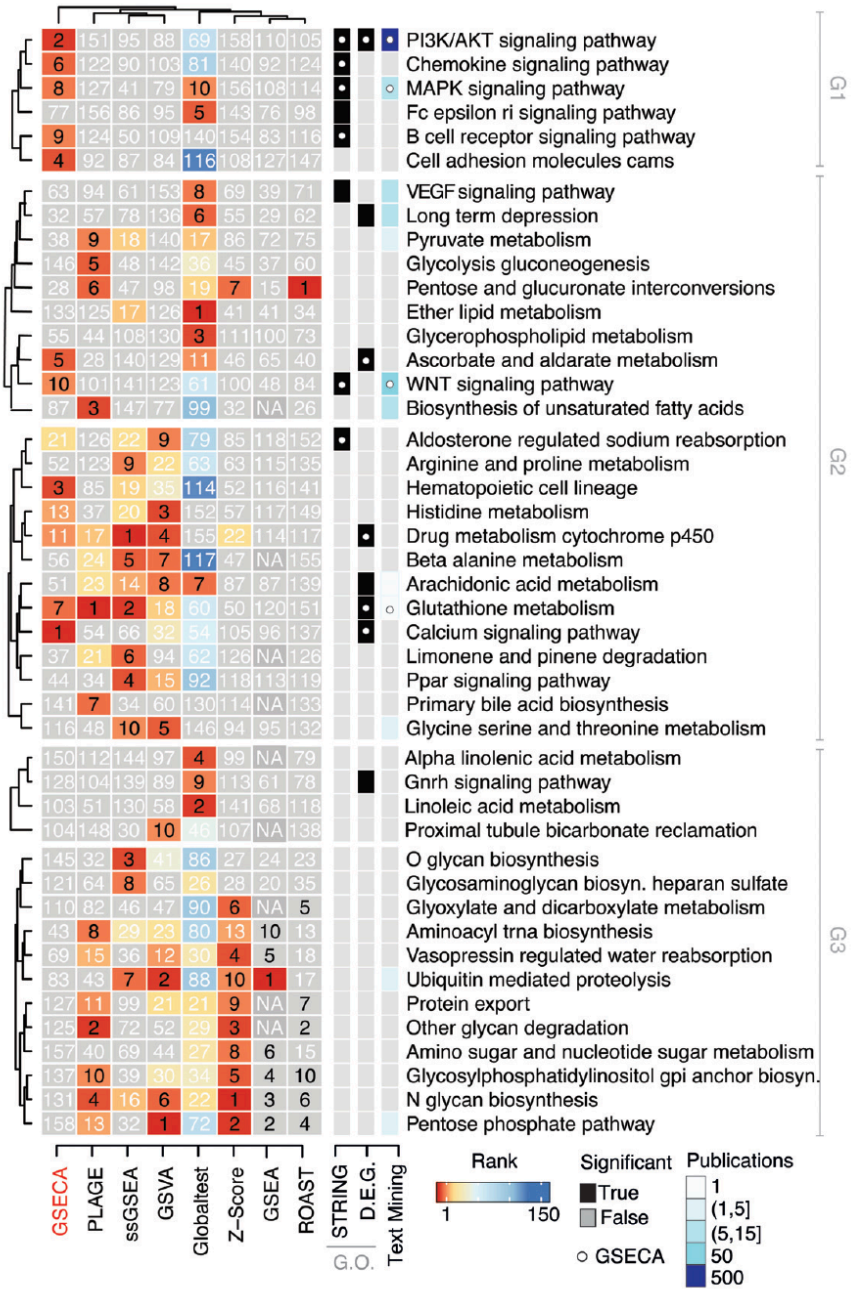
- G1: mainly GSECA and Globaltest;
- G2: generally resulting from all methods;
- G3: only GSA methods other than GSECA.

Notably, GSECA was the only method identifying the PI3K/AKT signaling pathway – i.e. an expected alteration in the PTEN-loss cohort, see paragraph 3.1 - among the top-ten AGSs. Moreover, we observed that the G1 group contains 75% of genes enriched from the STRING PPI gene ontology, and that five out of the eight gene sets enriched in the list of differentially expressed genes were found as significantly altered by GSECA (Figure 3.9). The GO analysis revealed enrichment of PI3K/AKT genes only when we considered significantly activated and repressed genes. The fact that GSECA was the only algorithm able to identify this gene set as top-ranked AGSs suggests its ability in identifying processes where genes are significantly altered in both directions rather than being either activated or repressed. Finally, four out of the ten top-ranked GSECA AGSs showed published evidence (≥9) for the interaction of PTEN with the gene set, with the highest number of articles (n = 499) supporting PTEN regulation of PI3K/AKT signaling pathway. In contrast, for the AGS identified by the other methods, we observed poor support from literature (mean publications = 4) and GO analysis, with no support from any of the three validation analyses for the G3 group (i.e., gene sets exclusively picked by methods other than GSECA).

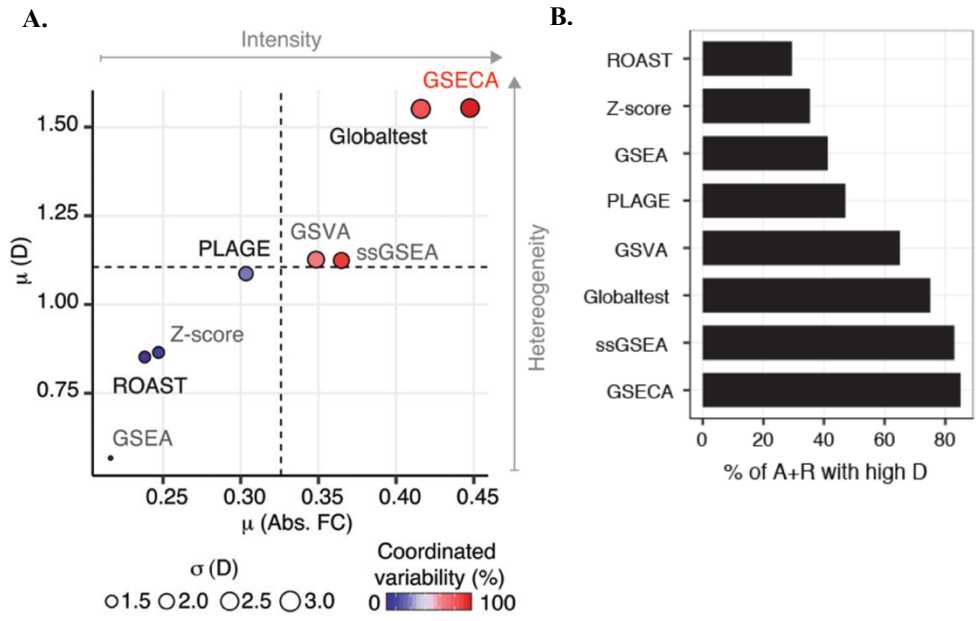
To gain further insights into the reasons why GSECA outperformed the other algorithms in detecting the alteration of PI3K/AKT signaling pathway upon PTEN loss, we analyzed the gene expression levels of the 20 top-ranked gene sets of each method independently from their statistical significance. As previously pointed out, a substantial fraction of PI3K/AKT genes is downregulated in a highly variable manner (see paragraph 3.1.2 and Figure 3.5B). Therefore, to detect its significant alteration, GSA methods must handle coordinated expression changes (i.e. activation or repression) of distinct genes in different samples even if they result in small FC differences in the whole population.

To evaluate this capability, we investigated the FC and dispersion (D) profile of the genes belonging to the selected gene sets (Figure 3.10A). In particular, for both FC and D, we calculated

the mean and standard deviation on the PRAD dataset, and next averaged them across genes. Moreover, to obtain a measure of the coordinated variability depicted by each GSA method, we calculated the proportion of gene sets composed by genes that were both significantly activated and repressed with a high degree of dispersion across samples (i.e. $\geq 75^{th}$ percentile of the dispersion distribution, Figure 3.10B).



I - Figure 3. 9: Evaluation of the AGSs identified by the GSA methods on the PRAD PTEN-loss dataset. Hierarchical clustering of the first ten top-ranked AGSs in PRAD PTEN-loss detected by GSA algorithms. Each cell reports the rank of the gene set of a specific method. The ranks of the top ten ranked gene sets are reported in black. Annotation heatmap (right) depicts gene sets identified by GO analysis performed considering the STRING PPI network and differentially expressed genes (i.e. DEG) in black and (ii) evidence coming from literature text mining in color key of blues.



I - Figure 3. 10: GSECA handles coordinated variability in AGSs. **A.** Scatter plot of the mean absolute FC (Abs. FC), and D averaged on the 20 top-ranked gene sets detected by each method. Dot size represents the average standard deviation (σ) of D for the 20 top-ranked gene sets. Color key depicts the percentage of the 20 top-ranked gene sets that contain both activated and repressed gene sets, namely coordinated variability. **B.** Bar plot depicting the percentage of the top 20 ranked gene set containing both activated (A) and repressed (R) genes with dispersion greater than the 75th percentile of the dispersion distribution.

We found that, overall, GSECA showed the best performance in characterizing gene sets that are more intensively activated or repressed (i.e. intensity and % of coordinated variability, Figure 3.10A-B) at all levels in a heterogeneous manner across samples (i.e. heterogeneity, Figure 3.10A). For this reason, GSECA was able to detect the altered modulation of PI3K/AKT signaling pathway that is composed of genes that are expressed at different levels (i.e. low and high) and are distinctly activated or repressed in different samples upon PTEN loss. The second best-method in this comparison was Globaltest. However, the highest type I error rate showed in simulation studies (see Chapter 2, paragraph 2.2.3), which resulted in a high number of significantly AGSs identified in the PRAD dataset (119 AGSs out of 158, Figure 3.8A), confers less robustness to its results. Remarkably, the remaining methods showed poor performance in handling datasets characterized by high degree of IH (Figure 3.10A).

Taken together, these results show that GSECA can detect functionally relevant altered biological processes under a phenotype of interest when considering more heterogeneous cohorts in contrast to other available methods.

Chapter 4

Pancancer analysis of PTEN somatic loss

Results

4.1 The pancancer dataset

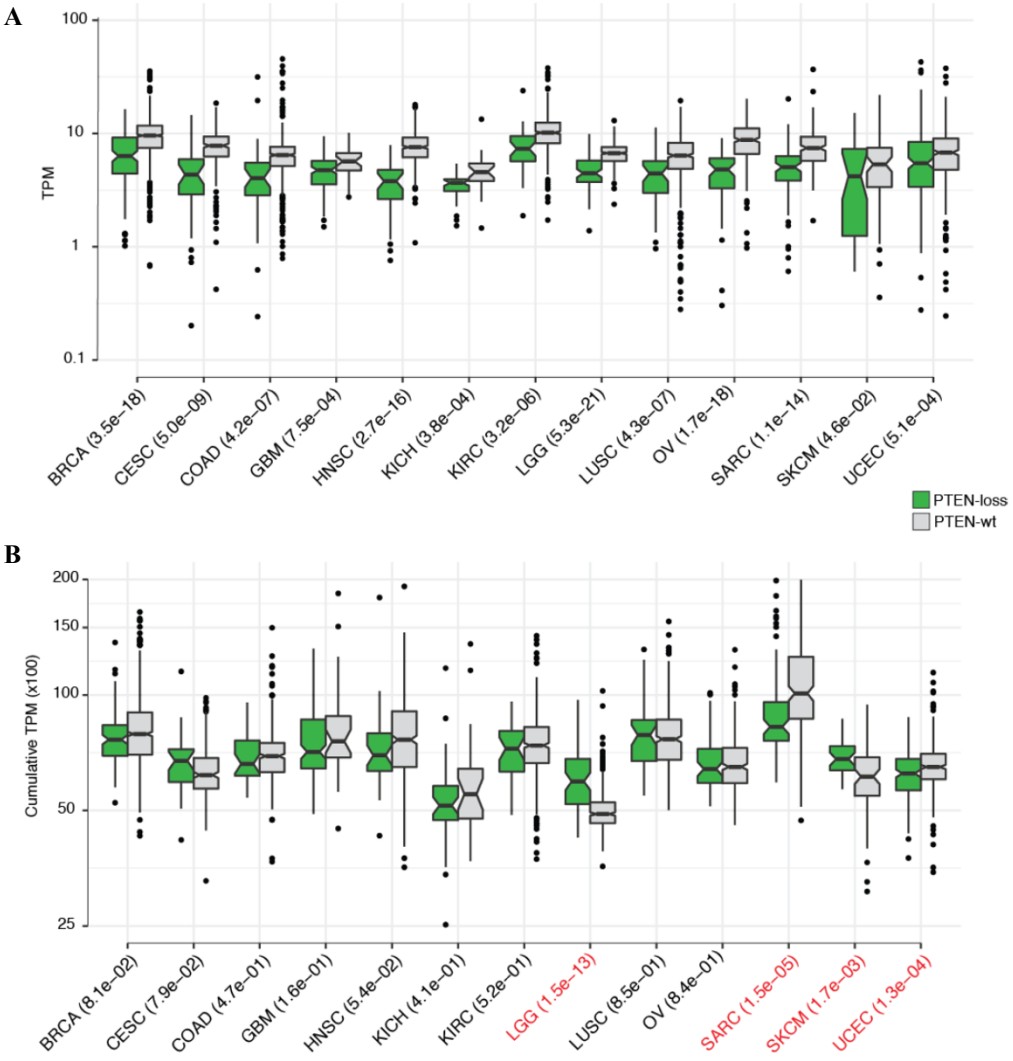
Somatic inactivation of PTEN occurs in a wide range of human cancers with various effects on each tissue (65). Therefore, we used GSECA to perform a comprehensive analysis of biological processes that are altered upon the somatic loss of PTEN across cancer types. To do so, we collected genomic data for 9944 samples of 31 cancer types available from TCGA. In particular, we downloaded data for somatic alterations and transcriptome profiling from the TCGA Data Matrix portal (Level 3, https://tcgadata.nci.nih.gov/tcga/dataAccessMatrix.htm). We identified samples harbouring PTEN somatic loss as described for the PRAD dataset (see Chapter 3, paragraph 3.1). Together with PRAD, we retained for subsequent analysis 13 cancer types for which we could identify at least 30 samples with somatic alterations of PTEN. As expected, we observed a significant reduction in the expression levels in PTEN-loss samples as compared to PTEN-wt ones for all cancer types (*P*-value < 0.05, one-tailed Wilcoxon rank sum test, Figure 4.1A). This lead to a significant alteration of the cumulative expression of PI3K/AKT signaling genes in 4 cancer types (*P*-value < 0.05, one-tailed Wilcoxon rank sum test, Figure 4.1B): low grade glioma (LGG), sarcoma (SARC), skin cutaneous melanoma (SKCM) and corpus endometrial carcinoma (UCEC).

4.2 GSECA results on pancancer PTEN-loss tumors

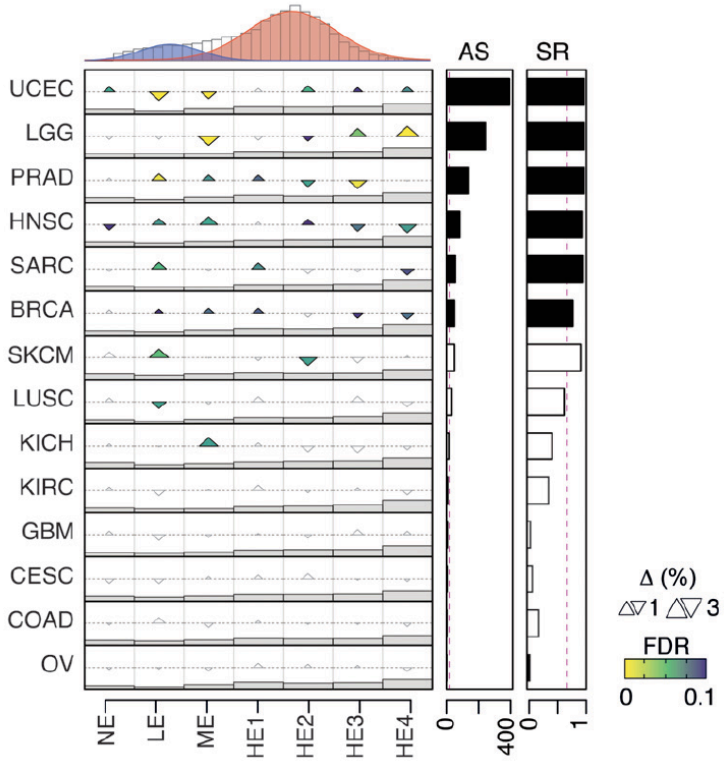
We next applied GSECA to the pancancer PTEN-loss dataset, testing the 158 KEGG human gene sets as described for the PRAD dataset (see Chapter 3, paragraph 3.1-2). We found that 10 out 13 cancer types showed at least one AGS ($AS \leq 0.05$, $p_{emp} \leq 0.05$ and $SR \geq 0.7$). Notably, six cancer types showed the significant alteration of the PI3K/AKT signaling pathway expression patterns (Figure 4.2): UCEC, LGG, HNSC, SARC, PRAD and breast cancer (BRCA). Indeed, GSECA AS showed a significant positive correlation with the extent of PI3K/AKT signaling pathway alteration, which was measured as the statistical difference in the cumulative expression of PI3K/AKT related genes in PTEN-loss tumors as compared to wild type samples (two-tailed Wilcoxon rank sum test, Figure 4.3A). To further assess whether the AS exploits the alteration of

PI3K/AKT signaling pathway, we employed a linear regression approach. In particular, for each of the 158 KEGG gene sets, we modeled the distribution of the AS (summarized by the median and inter-quartile range) as a function of six regressors given by a linear combination of the following variables:

1) the number of PTEN-loss and PTEN-wt samples;
2) the median Pearson's correlation coefficient of pairwise comparison of expression profiles in the two cohorts;
3) the statistically significant difference of PTEN expression levels between PTEN-loss and PTEN-wt tumors (i.e. one-tailed Wilcoxon rank sum P-value);
4) the statistically significant difference of PI3K/AKT signaling pathway cumulative expression levels between PTEN-loss and PTEN-wt tumors (i.e. two-tailed Wilcoxon rank sum P-value).



I - Figure 4. 1: The Pancancer PTEN loss dataset. **A.** Boxplot distributions showing PTEN normalized gene expression levels for PTEN-loss and PTEN-wt stratification in each of 13 cancer type for which at least 30 samples with somatic alterations of PTEN were found. **B.** Boxplot distribution showing the cumulative expression levels of genes in the PI3K/AKT signaling pathway for PTEN-loss and PTEN-wt samples in each of the 13 analyzed caner types.

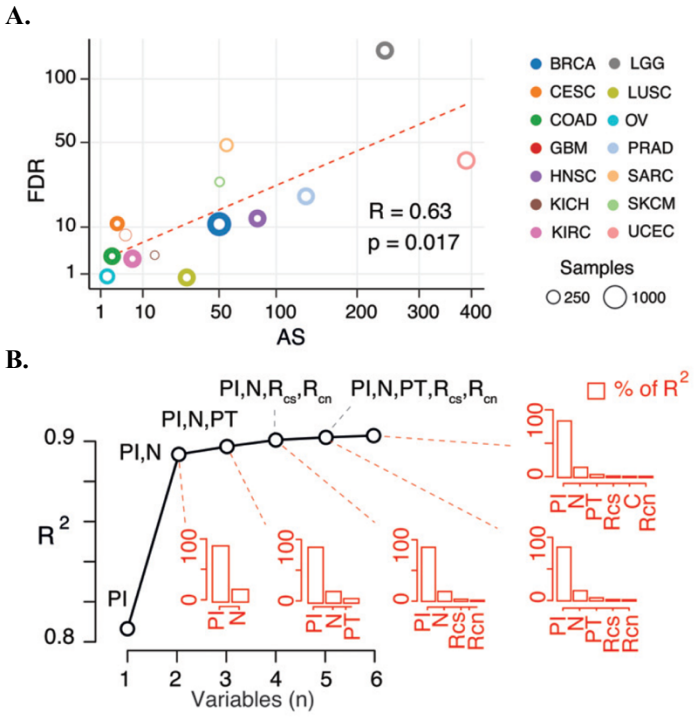


I - Figure 4. 2: Pancancer alteration of PI3K/AKT signaling pathway identified by GSECA. GSECA EC map showing the pancancer alteration of PI3K/AKT signaling pathway in tumors harbouring somatic loss of PTEN.

We then performed an exhaustive search for the best subsets of regressors for predicting the variability of the AS distributions using a branch-and-bound algorithm (66), as implemented in the *regsubsets* function of the R 'leaps' package (https://cran.r-project.org/web/packages/leaps/), using different number of variables (from 1 to 6) and reporting the best model in terms of coefficient of determination R^2. Finally, we calculated the relative importance of the model variables by splitting the coefficient of determination R^2 into the contribution of each regressor with the average over ordering method (67), as implemented in the R 'relaimpo' package (https://cran.r-project.org/web/packages/relaimpo/). We found that the alteration of PI3K/AKT signaling pathway gave the best fitting of the AS distributions in terms of coefficient of determination when using one predictor (R2 = 0.81, Figure 4.3B). Increasing the model complexity led to a closer fitting between the AS and the predictors (Figure 4.3B). Moreover, considering the relative importance of the regressors, we found that the alteration of the cumulative expression of PI3K/AKT signaling was the most critical regressor in all the linear models, accounting for 80 % of explained variance when considering all variables (i.e. n = 6, *P*-value = 0.009, Figure 4.3B). These results indicate that the GSECA AS consistently recapitulates the extent of PI3K/AKT signaling cascade alteration.

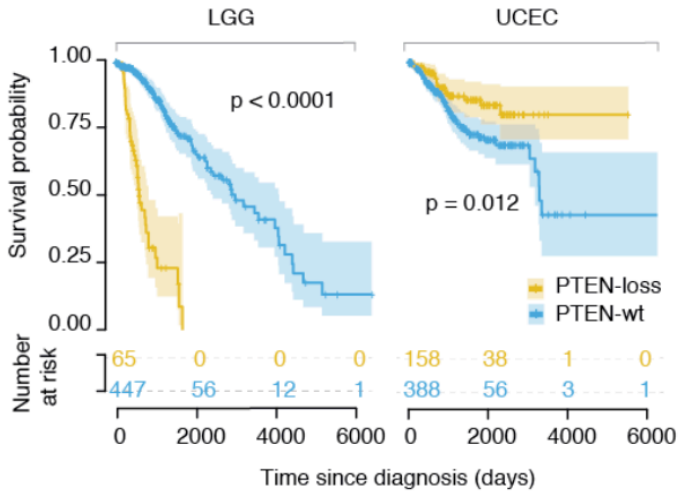
In addition to this, GSECA identified PI3K/AKT signaling pathway as altered in two cancer types (i.e. UCEC and LGG) for which the alteration of PI3K/AKT signature is known to impact on patient survival in positive and negative way (68,69). To state the validity of this result, we performed survival analysis on UCEC and LGG, comparing the disease-free survival probabilities

of PTEN-loss and PTEN-wt tumors. To do so, we downloaded clinical data from the GDC data portal (https://gdc.cancer.gov/) and defined disease-free survival time as the interval between the date of treatment and disease progression, as defined by biochemical or clinical recurrence, or until the end of follow-up (70).



I - Figure 4. 3: Correlation between the AS and the extent of PI3K/AKT signaling pathway alteration across cancer types. **A.** Scatter plot showing the correlation of GSECA AS (i.e. -10*log10(AS)) and the alteration of PI3K/AKT signaling pathway in PTEN-loss as compared to PTEN-wt tumors measured in terms of the adjusted P-value (i.e. -10*log10(FDR)) across cancer types. The size of the colored circles shows the number of samples, while the inner white circles the number of PTEN loss samples. **B.** Coefficients of determination (R2) of the linear regression model at the increasing of model complexity (i.e. the number of regressors in the model). PI=PI3K/AKT signaling pathway alteration; PT=PTEN downregulation; N=number of samples; RCS=correlation of PTEN-loss samples; RCN=correlation of PTEN-wt samples. Bar plots in red show the relative importance of each predictor to the R2 measured by the linear regression model of all variables.

We performed Kaplan–Meier estimation of the survival probabilities for the two groups of samples (i.e. PTEN-loss and PTEN-wt) and compared the resulting survival curves using the log-rank test (as implemented in the *ggsurvplot* function of the R 'survminer' package https://cran.r-project.org/web/packages/survminer, Figure 4.4). As expected, we observed a significant reduction of the disease-free survival probability for PTEN-loss LGG tumors as compared to PTEN-wt ones (*P*-value < 0.0001, log-rank test, Figure 4.4), and a significant increase of the disease-free survival probability for PTEN-loss LGG tumors as compared to PTEN-wt ones (*P*-value = 0.012, log-rank test, Figure 4.4). This result reinforced the robustness of GSECA predictions.



I - Figure 4. 4: Survival analysis of PTEN-loss LGG and UCEC tumors. Kaplan-Meier overall survival curves with the 'number at risk' table for LGG and UCEC PTEN-loss and PTEN-wt samples.

4.3 The somatic loss of PTEN impacts on immune-related processes

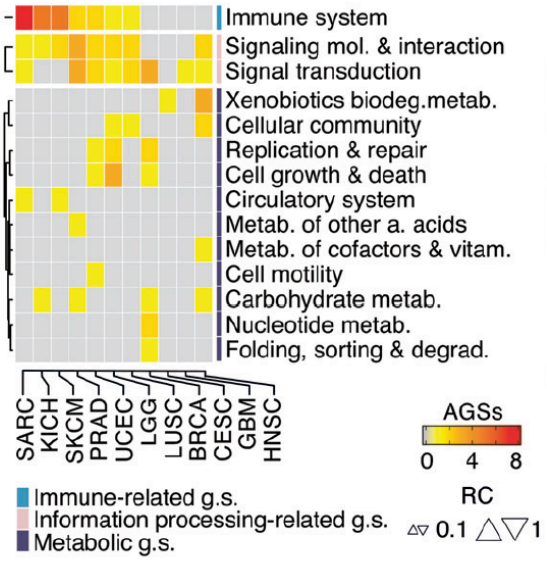
To gain functional insights on the cancer-specific regulation of PTEN, we next inspected the 10 top-ranked AGSs in each cancer type and hierarchically clustered them at an intermediate KEGG gene set category level (52) (Figure 4.5). We found that:

1) gene sets associated with metabolic pathways were distinctly altered across cancer types (Figure 4.5, bottom cluster);
2) information-processing gene sets (e.g. cell signaling activity) were altered in several cancer types (Figure 4.5, middle cluster);
3) immune system gene sets were altered in the majority of tissues (Figure 4.5, top cluster).

In particular, SARC, KICH and SKCM showed the highest number of immune-related AGSs, being hematopoietic cell lineage, chemokine and T cell receptor signaling pathways the most altered gene sets across cancer types. These results highlight the association between the loss of PTEN and the alteration of immune cell infiltrates, which has been recently noted (71). Given these observations, we further investigated the relationship between PTEN-loss and the immune-related transcriptional response across cancer types by integrating the results from two distinct analyses (Figure 4.6A):

- Comparison of the immune-related AGSs identified by the other selected GSA algorithms (see Chapter 2, paragraph 2.2.1).
- Evaluation of changes in the tumor immune microenvironment (TIME): we collected data about cellular composition of the immune infiltrates for TCGA tumors of 14 cancer types, available from (72), using the relative number of immune cells (73) as a measure of immune composition. To evaluate whether the composition of each immune cell type was altered upon somatic loss of PTEN across cancer types, we employed a Student's t-test to

compare cell fractions between PTEN-loss and PTEN-wt samples, adjusting for multiple hypothesis testing with the Benjamini & Hochberg procedure. Finally, to summarize the degree of perturbation of immune cells, we combined, for each cancer type, the resulting P-values into an immune score (IS) using Fisher's method (21) (Figure 4.6B, see Chapter 2, paragraph 2.1.1).



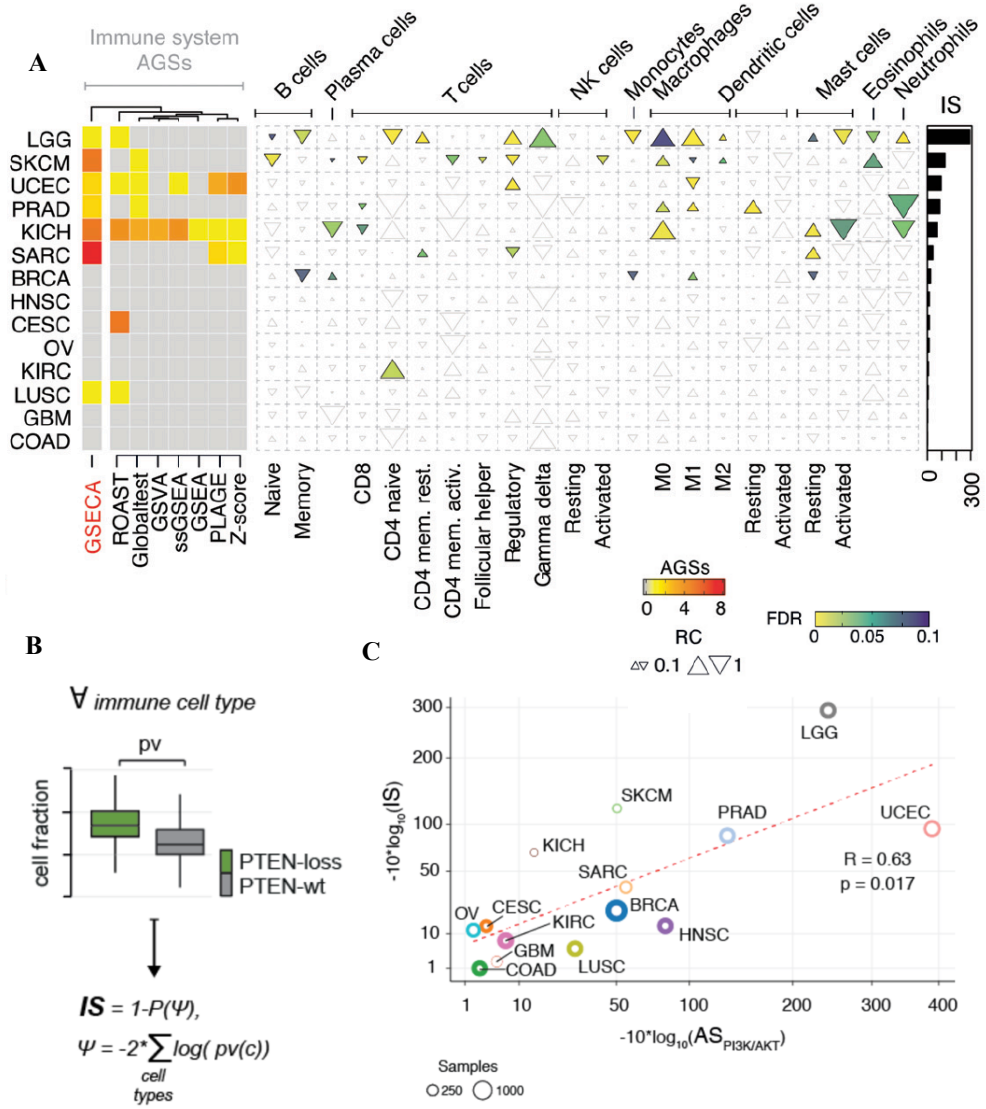
I - Figure 4. 5: Pancancer summary of AGSs. Heatmap showing hierarchical clustering of the altered classes of gene sets across cancer types. Classes are defined accordingly to intermediate the KEGG category *(52)*. Each cell reports the number of AGSs. The annotation heatmap indicates the KEGG superclass of biological processes.

Compared to the other GSA methods, GSECA detected the highest number of cancer types with a significant alteration of immune cell fractions (Figure 4.6A, left panel). Moreover, GSECA showed the highest positive correlation between the number of immune-related AGSs and the IS across GSA methods (Pearson's correlation coefficient R = 0.77, P-value = 0.003, Figure 4.6A, right panel). In particular, the AS resulted significantly positively correlated with the IS, highlighting the accuracy of GSECA results (Figure 4.6C). Taken together, these results indicate that GSECA was the most robust approach to highlight the link between PTEN loss and alteration of immune regulation by detecting the highest number of immune-related AGSs in the vast majority of cancer types with statistically significant changes in TIME composition.

4.4 Impact of PTEN loss on TIME of prostate adenocarcinoma

Emerging evidence has suggested that PTEN loss is an immunosuppressive event in prostate tumors (74). However, the connection between PTEN and the immune system is complex and involves both pro- and anti-tumorigenic immune responses depending on the cellular phenotype

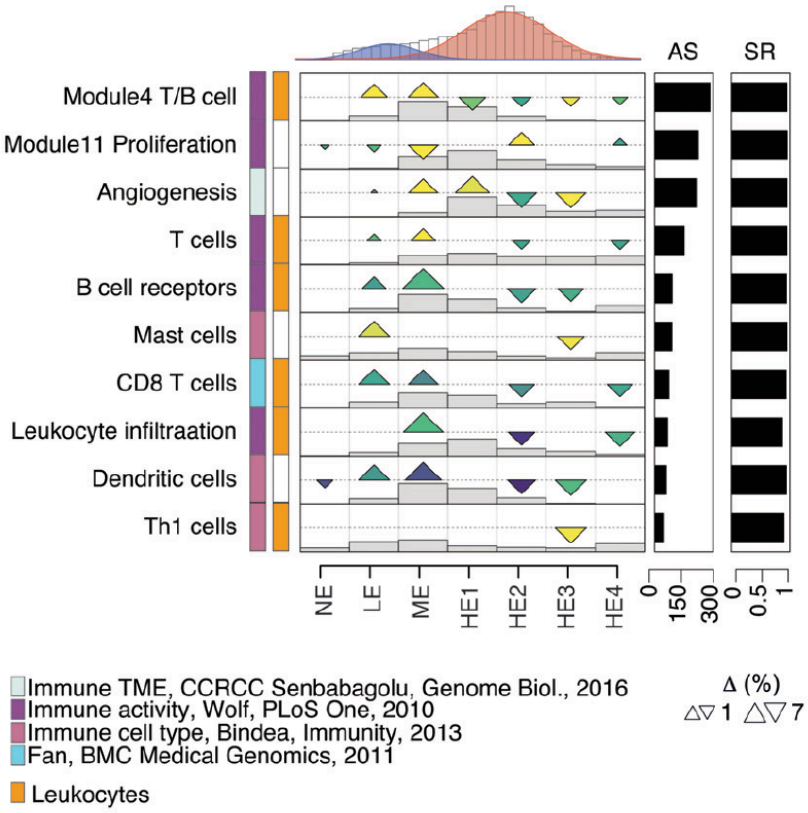
and the TIME (75). To assess the general applicability of GSECA we finally sought to investigate the impact of PTEN loss on TIME of PRAD samples.



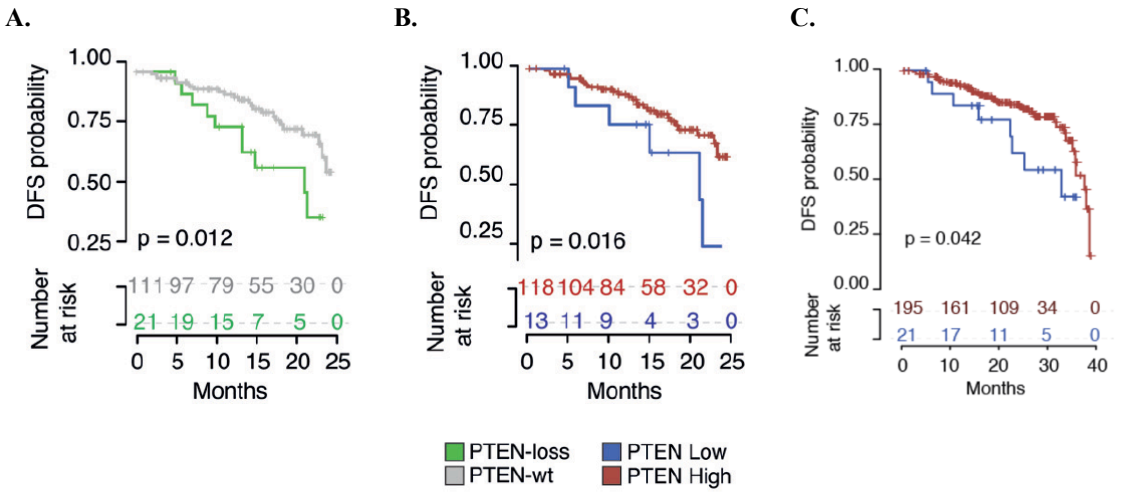
I - Figure 4. 6: Impact of PTEN loss on immune-related processes across cancer types. **A.** Heatmap on the left panel shows the number of immune-related gene sets that are altered upon the loss of PTEN across cancer types accordingly to GSECA and the other GSA methods. On the right panel, EC map-like heatmap depicts the statistically significant alteration of the immune cell population (i.e., TIME) across cancer types. The size of triangles the relative change of the percentage of tumor immune infiltrates between PTEN-loss and wild-type samples. Upper/lower vertexes of the triangles represent the increase/decrease of immune cells in PTEN-loss samples as compared to PTEN-wt tumors. The bar plot reports the IS for each cancer type. **B.** Evaluation of the immune score (IS). **C.** Pearson correlation coefficient of the AS measured for PI3K/AKT signaling pathway and the IS measured for all cancer types.

To do so, we ran GSECA on a collection of 102 expression signatures representative of different immune cell activities, states, and modes in tumor tissues (72). We found that 15 immune signatures were significantly altered upon PTEN loss ($AS \leq 0.05$, $p_{emp} \leq 0.01$ and $SR \geq 0.7$). Six of

the top ten AGSs characterized the state and activity of T and B cells, showing a general reduction of gene expression in the highly-expressed classes, and a reciprocal increase of genes in lowly expressed classed (Figure 4.7). These observations are consistent with previous data suggesting that PTEN loss prostate cancers are non-T cell inflamed, or 'cold', tumors (76). In particular, GSECA identified the decreased expression of genes representative of CD8 T cells, which was supported by the results of the TIME analysis (Figure 4.6D), as previously reported (74). Interestingly, the two top-ranked AGSs contained markers of lymphocyte activation (i.e. Module4 T/B cells) and cell proliferation (i.e. Module11 Proliferation), respectively. The combination of the down-regulation of the T/B cell module and the upregulation of the proliferation module has been strongly associated with decreased disease-free survival (DFS) in breast cancer patients (77). Since GSECA identified this same pattern, showing a reduction of genes in HE classes for the T/B cell module and an increase for the proliferation module (Figure 4.7), we assessed the impact of PTEN loss on DFS in prostate cancer patients by performing survival analysis (as described in 4.2). Remarkably, we observed a statistically significant reduction of DFS in PTEN loss patients as compared to PTEN-wt ones in the first 24 months from the treatment (*P*-value = 0.012, log-rank test, Figure 4.8A).



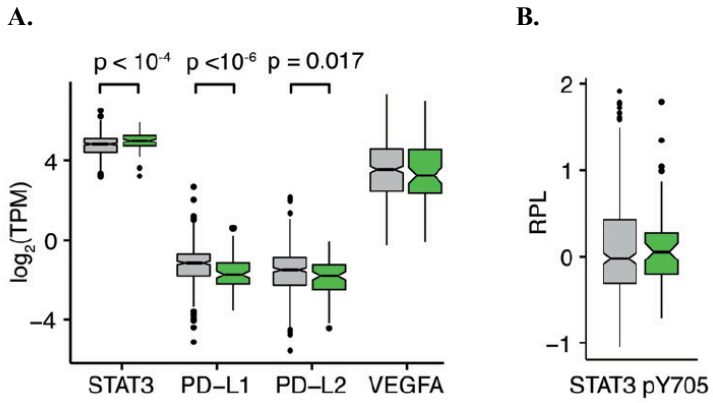
I - Figure 4. 7: GSECA analysis of TIME gene signatures in PTEN loss PRAD. GSECA EC map showing the altered immune expression signatures as a consequence of the somatic loss of PTEN in PRAD



I - Figure 4. 8: Impact of PTEN loss on DFS of PRAD patients. **A.** Disease-free survival (DFS) Kaplan-Mayer curves for PTEN-loss and PTEN-wt patients. **B.** DSF Kaplan-Mayer curves measured stratifying PRAD patients on the optimal PTEN expression level (i.e. TPM=3.56, maximally selected rank statistics=2.34) within two years from the initial treatment. **C.** DSF Kaplan-Mayer curves measured stratifying PRAD patients on the optimal PTEN expression level within three years from the initial treatment.

These data confirm the detrimental impact of PTEN loss on prostate cancer disease phenotype. Furthermore, since PTEN status determination impact on therapy management of prostate cancer patients (75), we wondered whether absolute PTEN expression levels could be prognostic of a shorter DFS. Using the maximally selected rank statistics approach (78), we found that patients with PTEN expression levels lower than 3.58 TPM had a statistically significant shorter DFS time in the first two years (Figure 4.8B), as well as three years (Figure 4.8C), from the initial treatment. These results highlight that not only the genomic status but also the expression levels of PTEN are associated with poor outcomes in patients with prostate cancer.

It has been shown that in *Pten*-null mice the activation of the Stat3 establishes an immunosuppressive TIME that contributes to tumor growth and chemoresistance (79). Therefore, to finally validate GSECA results, we compared the expression levels of STAT3 in PRAD PTEN loss tumors as compared to controls. We also evaluated the expression of the inhibitory immune checkpoint molecule PD-L1 and PD-L2 and the immune inhibitor VEGFA (72). We found that PRAD PTEN loss tumors significantly expressed STAT3 at higher levels and PD-L1 and PD-L2 at lower levels than PRAD PTEN wild-type samples (Figure 5A). Moreover, the level of phosphorylation of STAT3 was significantly higher in PRAD PTEN loss tumors as compared to controls (Figure 5B). These data support the establishment of an immunosuppressive TIME in human prostate cancers, which could be driven by the activation of STAT3, and validate the statistically significant associations found by GSECA.



I - Figure 4. 9: Immunosuppressive TIME in PTEN loss PRAD tumors. **A.** Boxplots showing expression distributions of PTEN normalized expression levels for PTEN-loss and PTEN-wt samples of four immune-response related genes. **B.** Boxplot distributions of the relative level of STAT3 phosphorylation for PTEN-loss and PTEN-wt PRAD samples.

Taken together, these results show the general applicability of GSECA in detecting biological processes that are altered in high-volume heterogeneous data sets. In particular, GSECA has proved highly accurate in associating the loss of PTEN to the alteration of PI3K/AKT signaling pathway and to the different regulation of immune-related processes across cancer types. In prostate cancer, GSECA detected the detrimental impact of PTEN loss on DFS of patients and the establishment of a 'cold' TIME through the down-regulation of lymphocytes signatures. Hence, our results support the emerging role of PTEN in immune system and therapy resistance (71,80,81)

Chapter 5

Discussion I

5.1 A novel approach to the GSA

In this work, we concentrated our efforts on extracting valuable biological knowledge from the huge amounts of genomic data that are now available to the scientific community. In particular, we focused on transcriptome profiling data and exploited computational techniques to resolve their high degree heterogeneity and complexity. Heterogeneity is a fundamental feature of information associated with complex phenotypes, which can arise from the subtle alterations of distinct genes in different patients rather than of a single gene (7), and IH affects the power of computational tools for the GSA in detecting these subtle changes.

With this in mind, we developed a novel approach to the GSA implemented in the GSECA algorithm. Starting from the global distribution of gene expression profiles coming from RNA-seq experiments, GSECA implements FMM to interpret their bimodality, which reflects the presence of two major subpopulations of transcripts in cells (14), and employs a DD procedure to reduce the gene expression measurements into a small list of categorical values, an approach that has been recently exploited in Big Data analysis to increase signal-to-noise ratio and improve the accuracy of machine learning algorithms (20,82,83). Finally, using these discretised expression values, GSECA implements a statistical procedure based on combined Fisher's exact tests to identify the relevant altered biological processes for the phenotype of interest.

5.2 GSECA best handles IH in both simulated and real data

To evaluate the performance of our method as compared to existing GSA tools, we employed both the controlled setting of *in silico* simulations as well as real high-volume RNA-seq datasets. In simulations, we modeled several conditions of differential gene expression, with a focus on the capability of handling IH. We observed that, compared to other seven 'state-of-art' algorithms, GSECA showed the lowest overall type I error rate and a high statistical power in detecting AGS, outperforming the other methods in treating heterogeneous RNA-seq data. Most importantly, GSECA displayed the highest F1 score – i.e. the best trade-off between sensitivity and specificity- among all methods to detect truly AGSs in the presence of IH in gene expression between samples,

as shown in the 'dispersion' simulation study (see Chapter 2). Therefore, GSECA can identify a smaller number of AGSs as compared to other GSA methods but with a higher sensitivity.

The predictions of GSECA were the most accurate ones when treating heterogeneous samples suggesting that its framework enhances the signal-to-noise ratio, and thus data interpretation. Interestingly, among the other methods, ssGSEA showed similar performance to GSECA in handling heterogeneity. Comparably to GSECA, this method treats each sample individually and collapses gene expression levels to a common scale using ranks (25). This finding confirms that the reduction of the large set of expression levels into a smaller range of values increases the power to detect truly AGSs in the presence of IH.

On real data, we used out method to identify the biologically relevant gene sets that are altered upon the somatic loss of PTEN, and the subsequent alteration of the PI3K/AKT signaling cascade in prostate cancer. GSECA correctly detected the alteration of the PI3K/AKT signaling pathway and related signal transduction gene sets, such as calcium signaling (53), epithelial CAMs (54), MAPK (55) and WNT signaling pathways (56). Interestingly, the FMM and the DD approaches captured the heterogeneity among the cohorts revealing a general decreased and widespread gene expression in prostate cancer due to the loss of PTEN that might underline the role of PTEN in regulating basal transcription through histones and chromatin remodelling (51). The comparative performance analysis of GSA methods in detecting the effect of PTEN silencing shows that GSECA was the only algorithm able to reveal the expected altered modulation of PI3K/AKT signaling pathway. Moreover, GSECA detected the altered regulation of processes where genes directly interact with PTEN and, thus, are influenced by the somatic loss of their interactor. This result indicates the ability of the method to spot functionally related AGSs. Importantly, GSECA highlights the alteration of gene sets composed by genes that are coordinately and heterogeneously modulated rather than being uniformly activated or repressed at different levels (i.e. low and high) in distinct samples (see Chapter 3), whereas other methods might suffer from this limitation. Together, these results indicate that GSECA boosts the signal-to-noise ratio in heterogeneous datasets, thus enabling the identification of the general biological mechanisms that are altered across samples.

5.3 PTEN and the immune system in cancer

The pancancer analysis of the effect of PTEN somatic loss generated a comprehensive assessment of its regulation across tissues. PTEN critically interconnects the canonical PI3K/AKT and the RAS/MEK/ERK pathway, which are the two dominant tumorigenic gene sets controlling cell survival and proliferation (65). Our data shows that the impact of PTEN silencing on cellular program regulation is proportional to the impaired modulation of the PI3K/AKT signaling cascade, with the stronger effect of gliomas, endometrial, head and neck, breast carcinomas, melanomas, and sarcomas. GSECA revealed a tissue-specific control of PTEN on metabolic processes, whereas

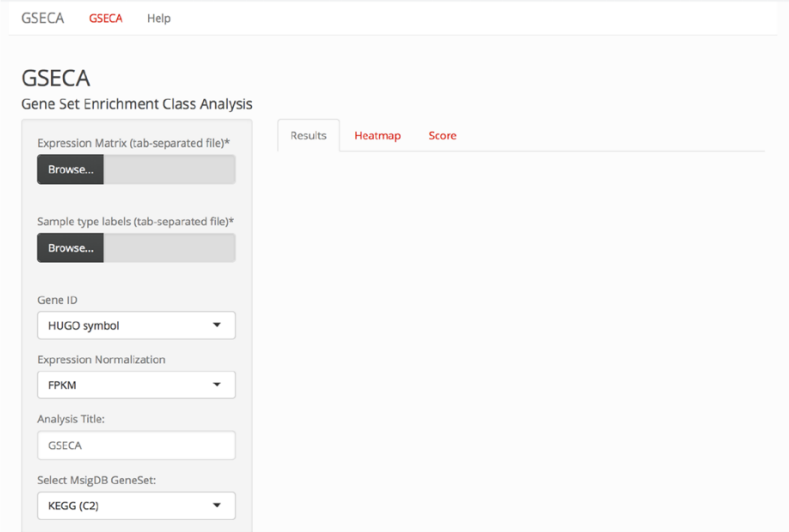
information-related processes, such as signal transduction, are more uniformly affected across tissues. Most importantly, GSECA correctly highlighted the role of PTEN in controlling immune-related processes in the majority of cancer types, particularly in those showing a significant alteration of the TIME composition. These data support the importance of PTEN in modulating the immune system (80) and therapy resistance. Using additional immune expression signatures, GSECA correctly highlights the immunosuppressive TIME of PTEN-loss prostate tumors (75), which could be driven by the significant activation of STAT3. Furthermore, GSECA results were pivotal to show the shorter of disease-free survival of these patients and to underline the biomarker potential of PTEN expression levels. These results validate previous findings in prostate mouse models (79), melanoma (81), breast (77) and provide indications that might be important for the clinical management of prostate cancer patients.

5.4 Conclusions and perspectives

In conclusion, our findings concordantly indicate that GSECA can improve the comprehensive identification of relevant biological processes that are altered in complex phenotypes. In particular, GSECA can detect functionally related and relevant altered cell mechanisms in a condition of interest considering more heterogeneous cohorts as compared to other available methods. By boosting signal-to-noise ratio, GSECA can successfully manage the heterogeneity of thousands of samples and provides useful insights on clinical and biological patterns proper of a phenotype. This encourages, as future work direction, its application to single-cell RNA-sequencing datasets, which typically consist in high-volume and intrinsically heterogeneous data (84).

5.5 Software availability

GSECA software is written in the R programming language. In order to facilitate its usage, GSECA is implemented as an R/Shiny application with a dedicated graphical interface, which is freely available from GitHub (https://github.com/matteocereda/GSECA). Pre-processed gene sets of biological pathways (7,52) and diseases (40) are provided.



I - Figure 5. 1: GSECA R/Shiny application. View of GSECA R/Shiny application. The input consists in two files: (i) Expression Matrix, which is the file containing the expression levels of each RNA-seq experiment. It is a tab-separated file arranged in matrix form, where rows are genes and columns are samples. The gene ID can be provided as HUGO symbol or ENSEMBL gene ID. (ii) Phenotype labels, which is the file containing the phenotype labels assigned to the samples in the Expression Matrix. It consists in an ordered list of labels and is provided as a tab-separated file. The output panel of GSECA app contains the results of the results of the analysis in tabular form and the EC map.

Part II

Dissecting the functional role of de novo DNA methylation during embryonic lineage specification

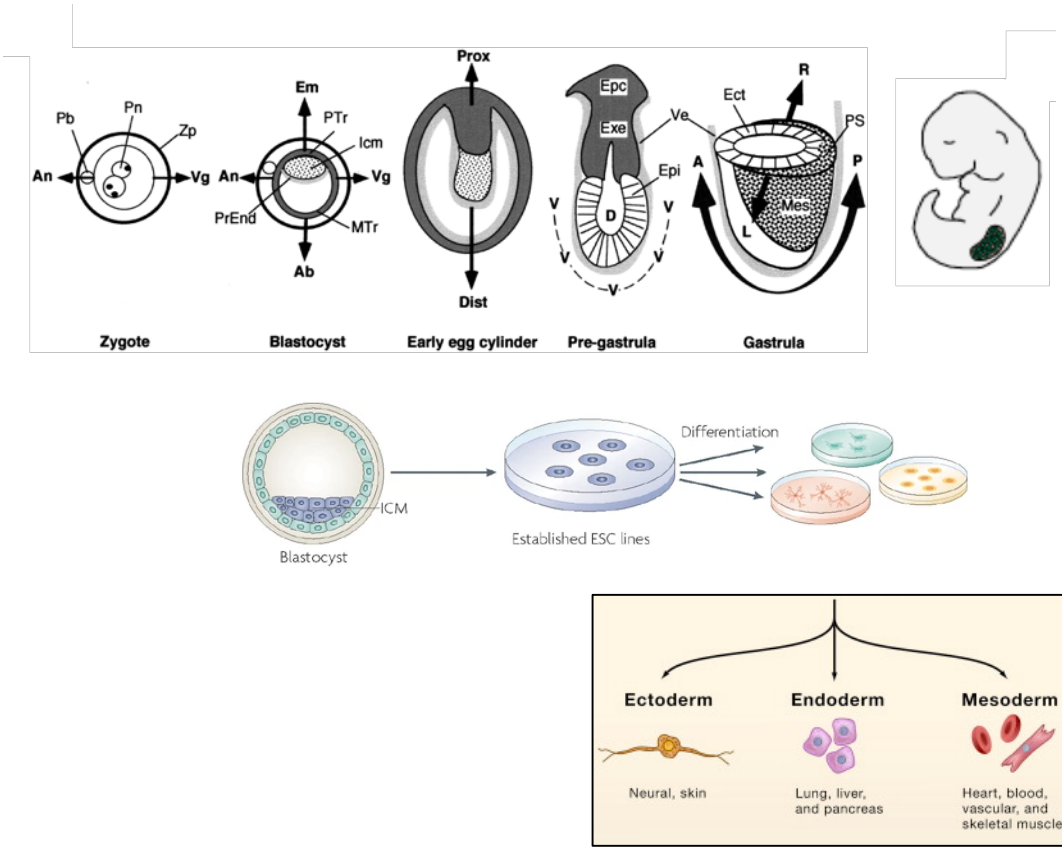
The data in the following chapters have been submitted for publication.

Chapter 1

Introduction II

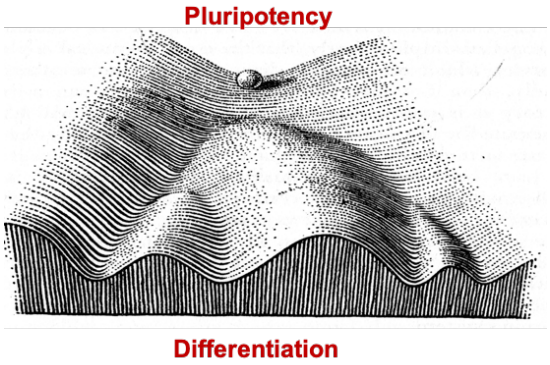
1.1 The epigenetic landscape

Cell differentiation during embryogenesis is a delicate process in which transcription and repression of specific genes must be tightly regulated. Embryonic stem cells (ESCs) are pluripotent cells derived from the inner cell mass (ICM) of the blastocyst that have the capacity to self-renew indefinitely (85) (86). Under proper stimuli, ESCs can differentiate into three primary germ layers - ectoderm, mesoderm and endoderm –, mimicking the process of gastrulation (87), and develop potentially into different tissues and organs (88-90) (Figure 1.1). The process of going from a pluripotent toward a differentiated state during embryonic development was firstly described with the metaphor of the epigenetic landscape by Conrad Hal Waddington (91). In this representation, pluripotent cells are depicted as balls rolling down a hill, which can run across specific allowed trajectories in order to reach their differentiated state and final tissue type. Hence, at various branching point, cells must take decisions about their fate, activating the expression of specific genes while repressing others (Figure 1.2). During this process, several coordinated molecular actors come into play, shaping the propagation of the epigenetic information – *i.e.* information responsible for stable and heritable changes in gene expression or cellular phenotype that is not propagated through the DNA sequence (92,93). The maintenance and self-renewal of the pluripotent ESC state requires a core network of transcription factors (TFs), including *Oct4*, *Sox2*, *Nanog* and *Klf4* among others (94,95), which is deconstructed and rewired in presence of differentiation stimuli to activate lineage-specific genes (93,94). Along with this, the chromatin – which represents the template for epigenetic regulation (93) - is extensively remodelled during development, involving the acquisition of biochemical modifications in its constituents – i.e. DNA methylation and histone protein modifications – as well as changes in the overall three-dimensional conformation, which result in the promotion of specific transcriptional programs at the time of lineage commitment (93).



II - Figure 1. 1: In vitro Embryonic Stem Cells differentiation. Schematic representation of the different stages of mouse embryonic development - from the zygote stage to gastrulation - and in vitro differentiation of pluripotent ESC derived from the ICM into progenitors of the three primary germ layers. – Adapted from (87,96,97).

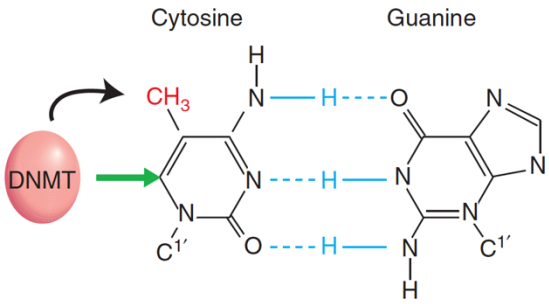
Such different layers of epigenetic modifications are deeply interconnected and dynamically regulated, thus conferring plasticity to the cells in response to external stimuli which can be inherited by cell progeny (93).



II - Figure 1. 2: The Epigenetic Landscape. Conrad Hal Waddington's view of the epigenetic landscape as a metaphor for embryonic development. To go from a pluripotent toward a differentiated state, cells (balls rolling down the valley) must take decisions about their fate at key branching point of their trajectories. - Adapted from (91).

1.2 DNA methylation machinery and mechanisms

DNA methylation (DNAme) is one of the best studied epigenetic modifications. It consists in the covalent modification of DNA by methylation of the fifth position of cytosine (5mC) (Figure 1.3). Appeared firstly in bacteria, it is fairly well conserved across species, including many plant, animal and fungal models (98), even though it has been lost in several eukaryotic lineages (*e.g.* the common model organism *Drosophila melanogaster* presents no 5mC) (99). In fact, in mammals, cytosine methylation occurs mainly at symmetrical CpG dinucleotides (*i.e.* Cytosine-phosphate-Guanine). As a consequence, mammalian genomes are CpG-depleted due to the intrinsic mutagenic properties of 5mC, which can spontaneously deaminate, yielding a C to T transition. This kind of mutations appear to be the most abundant point mutations in humans (100), and the evolutionary instability of CpG is confirmed by the fact that mammals have 4-to-5 fold underrepresentation of CpG dinucleotides than expected by their genome composition (99). Nevertheless, mammalian genomes are pervasively methylated: 70-80 % of all CpG sites exhibits 5mC in somatic tissues, and dysregulation of the DNAme landscape is typical of human cancers as well as many other diseases (71,100). The advent of new NGS-based technologies, such as whole-genome bisulfite sequencing (WGBS), enabled genome-wide profiling of DNAme at single base resolutions across organisms and cell types. Results from DNAme mapping studies revealed the presence of high 5mC levels in repetitive elements, pericentromeric satellite repeats, non-repetitive intergenic DNA sites and gene bodies (99,100).

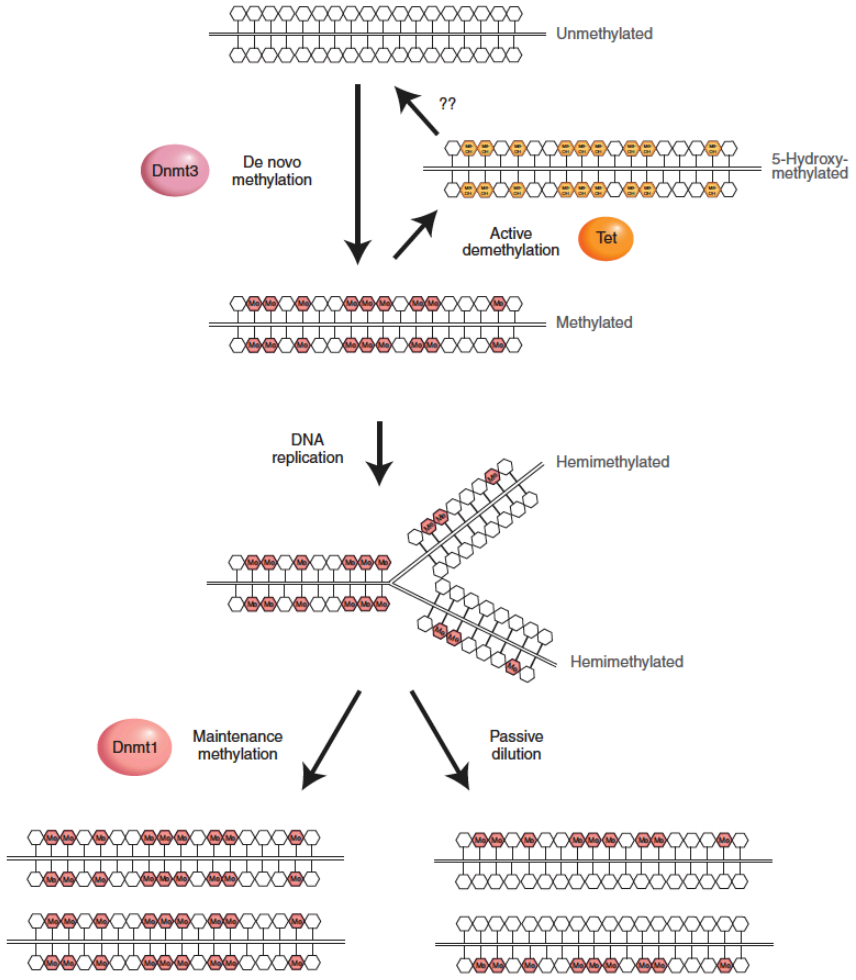


II - Figure 1. 3: Cytosine methylation in DNA. Covalent modification of DNA cytosine residues by addition of a methyl group (CH_3) in the fifth position of the pyrimidine ring – Adapted from (100)

Exceptions to this global trend of methylations of mammalian genomes is represented by GC-rich regions called CpG islands (CGIs). CGIs are genomic regions of ~ 1kb - accounting for less than 10% of all CpG sites - that are refractory to DNAme. Around 60 % of all human promoters have CGIs, and these are particularly enriched at transcription start sites of housekeeping and developmental regulator genes (98).

The whole process of DNAme consists of three phases: establishment, maintenance and erasure (Figure 1.4). Each of these steps involves a set of writer and eraser proteins (99):

- *Establishment*. During mammalian development, dynamic changes of DNAme are observed. Following fertilization, the genome undergoes genome-wide DNA demethylation, and the acquisition of *de novo* DNAme occurs next at the time of embryo implantation (100). Three major DNA methyltransferases (DNMTs) are responsible for the establishment of *de novo* DNAme patterns: DNMT3A, DNMT3B and DNMT3L. DNMT3A and DNMT3B both possess a highly conserved catalytic domain (MTase domain) and two conserved chromatin reading domains, namely the ATRX-DNMT3-DNMT3L (ADD) and the "proline-tryptophan-tryptophan-proline" (PWWP) domains, which accomplish distinct functions (99,100). Conversely, DNMT3L is catalytically inactive and acts in complex with DNMT3A/B to stimulate their activity, with peculiar roles in germ cells (101).
- *Maintenance*. Historically, DNAme has been viewed as the archetypal mechanism of cell memory and epigenetic inheritance given its stability and propagation through multiple cell division (100,102). In fact, the patterns of 5mC are inherited by cell progeny through replication via the activity of the DNA methyltransferase DNMT1 in concert with two core proteins, the proliferating cell nuclear antigen (PCNA) and the E3 ubiquitin-protein ligase UHRF1 (99). During DNA replication, UHRF1 recognizes hemi-methylated CpGs at replication fork, binding specifically to the methylated strand, and recruits DNMT1 to methylate the daughter strand (98), thus ensuring the re-establishment of symmetrical CpG methylation after mitosis.
- *Erasure*. The two waves of global DNA demethylation that occur in germline and early embryogenesis are driven by both passive and active processes (100). Passive demethylation occurs due to the lack of DNAme maintenance through cell replication, in absence or inhibition of DNMT1 activity. This leads to replication-dependent dilution of 5mC levels, as observed post-fertilization to the zygote maternal genome (100). Conversely, active demethylation is mediated by the activity of the ten-eleven translocation (TET) methylcytosine dioxygenases family of enzymes (TET1, TET2 and TET3), which can iteratively oxidise 5mC to 5-hydroxymethylcytosine (5hmC), 5-formyl cytosine (5fC) and 5-carboxylcytosine (5caC) (100). The conversion of 5mC to 5hmC blocks the DNAme maintenance mechanism, given that 5hmC is no more recognized by DNMT1/UHRF1 during DNA replication, thus leading to loss of the modification in cell progeny. Moreover, the subsequent oxidative products 5fC and 5caC can be recognized and cleaved by thymine DNA glycosylase (TDG), followed by correct base-pair restoration by the base-excision repair (BER) pathway (100).



II - Figure 1. 4: DNAme machinery. The establishment of de novo 5mC patterns is mediated by DNMT3A and DNMT3B alone or in complex with DNMT3L, while DNMT1 is responsible for DNAme maintenance through cell replication. Demethylation can occur in both passive and active ways. Passive dilution happens when 5mC is not maintained through cell division, while active demethylation is mediated by the TET methylcytosine dioxygenases family of enzymes – TET1, TET2 and TET3, which convert 5mC to 5hmC, 5fC and 5caC, leading to both dilution across cell replication and demethylation by TDG and base-excision repair system. - Adapted from (100).

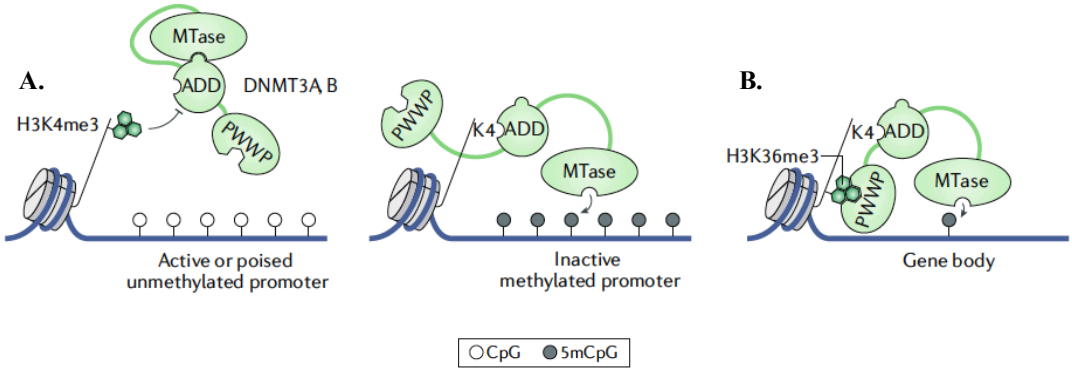
1.3 DNA methylation in the regulation of gene expression

Since its discovery, DNA methylation has been associated with transcriptional repression, observed as a negative correlation between 5mC levels at promoters and the transcription of related genes (100). However, the underlying molecular mechanisms that lead to gene silencing are not obvious and not yet completely clarified, and the modification might serve distinct functions at different genomic regions.

The deposition of 5mC by the DNMT3 family is mediated by the interaction with histone modifications (Figure 1.5). Specifically, in promoters of actively transcribed genes, the enrichment of tri-methylation marks on histone H3 Lys4 (H3K4) prevents the binding of de novo methyltransferases via their ADD domain, which additionally auto-inhibits the *de novo* methylation activity by binding the MTase domain (Figure 1.5A). Conversely, in absence of H3K4me3, the ADD domain can bind H3K4 tails, thus allowing the MTase domain to deposit new

5mC marks (Figure 1.5A) (99). DNA methylation at promoter elements can block gene activation in different ways (99). The presence of the 5mC mark interferes with the binding of several transcription factors that are sensitive to CpG-containing sequence motifs, thus preventing the activation of their target genes (99,100). Moreover, promoter silencing can be mediated by the recruitment by the DNMTs of chromatin remodelling complexes that promote heterochromatin formation (99). This can either occur via the interaction of *de novo* DNMTs with such complexes (e.g. the chromatin remodelling protein LSH (103) and histone H3K9 methyltransferase complex G9A (104)), or via the recognition of 5mC-containg sequences by the methyl-CpG-biding (MDB) family of proteins, which directly interact with histone deacetylase (HDAC) and nucleosome remodelling (NuRD) complexes to promote transcriptional repression (99,100). The involvement of DNAme in promoting stable transcriptional repression regulates various cellular processes in somatic tissues, such as the X chromosome inactivation, genomic imprinting, repression of germline-specific genes and of transposable elements, all of which occurring by 5mC silencing of their associated CpG-rich promoters and/or CGI (99).

In contrast to promoter elements, the gene bodies of actively transcribed genes show enrichment for 5mC marks (99). The deposition of intragenic DNAme is mediated by an epigenetic crosstalk mechanism involving DNMT3B, the histone methyltransferase SETD2 and the trimethylation of lysine 36 on histone H3 (H3K36me3) (105,106). Specifically, during transcriptional elongation, SETD2 is recruited by the RNA polymerase II (Pol II) for the deposition of H3K36me3 marks, which subsequently enable the recruitment of DNMT3B via the PWWP domain (105) (Figure 1.5B). The positive correlation between transcription and gene body DNAme has been linked to the involvement of 5mC in the regulation of co-transcriptional alternative splicing events (107) and, more recently, in agreement with the repressive nature of the 5mC mark, to the inhibition of intragenic cryptic promoters in cooperation with H3K36me3 and SETD2, thus ensuring the fidelity of the transcriptional machinery (106).



II - Figure 1. 5: Molecular mechanisms of de novo DNAme. Models of the molecular interactions related to the deposition of de novo 5mC marks at promoters (**A**) and gene bodies (**B**). – Adapted from (99).

1.4 Epigenetic reprogramming during development

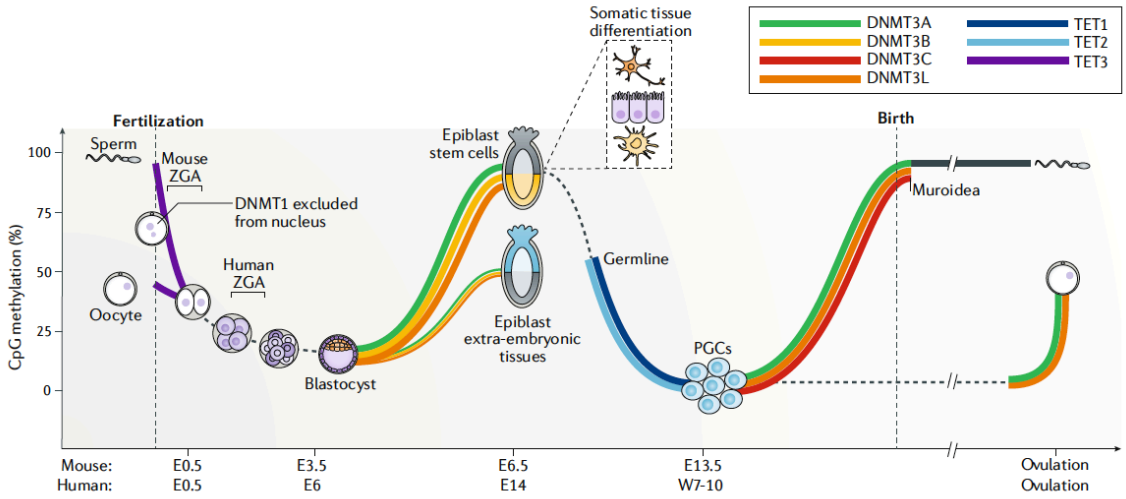
There are two major waves of epigenetic reprogramming occurring during development: following fertilization and during the development of germline progenitors (primordial germ cells – PGCs) (99). Early embryonic and germline DNA demethylation is established genome-wide in two phases, involving both active mechanisms, via the activity of the TET family of enzymes, and passive replication-dependent dilution (Figure 1.6) (99). In the germline, passive demethylation is followed by TET1 and TET2 hydroxylation activity, mostly targeting imprinting and germ cell identity genes. In post-fertilization embryo, the genome undergoes gamete-specific loss of DNAme, with the paternal genome being rapidly demethylated by TET3 in the first place, followed by replication-dependent dilution of both parental genomes until the blastocyst stage (even though evidences have been found for active TET3-mediated demethylation also for the maternal genome, to a lower extent) (99,100). However, in both embryonic and germline reprogramming, a small but relevant portion of DNA escapes the global DNA demethylation process. In the early embryo, this notoriously happens to genomic regions known as imprinting control regions (ICR), in which parent-of-origin DNAme patterns are retained to force allele-specific expression of related genes (*i.e.* the inheritance process of genomic imprinting) (99). These regions are able to resist reprogramming thanks to the sequence specific recruitment of KAP1, which protects the locus from DNA demethylation (99). Besides classical stable imprinting of ICR, recent evidences showed transient imprinting phenomena in early embryo for various loci, which mostly maintain maternal patterns of DNAme until the blastocyst stage, that, even if lost following implantation, can have a lifelong impact (99,108). In addition to genomic imprinting, the remaining portion of CpGs that escapes DNAme reprogramming is related to repeats - especially evolutionary young and potentially active families of transposons -, both in pre-implantation embryos and germline development (in which genomic imprinting does not occur in order to allow the subsequent establishment of sex-specific patterns of DNAme) (99,100).

Following the reprogramming process, *de novo* patterns of DNAme are established via the activity of the *de novo* DNA methyltransferases (Figure 1.6). In PGCs, sex-specific patterns of DNAme are established mostly by DNMT3A in cooperation with DNMT3L, which is essential for gametogenesis (101). In the developing embryo, during the transition from the pre-implantation blastocyst stage (E3.5) – where the ICM is composed by naïve pluripotent stem cells – to the post-implantation embryo (E6.5) – where most of the ICM is composed by epiblast stem cells – an strong increase in the expression levels of DNMT3A and DNMT3B is observed, together with a rapid increase in the global levels of 5mC, which here are established at levels that will eventually persist in somatic tissues (99). Epiblast stem cells are still pluripotent, thus the establishment of *de novo* DNAme patterns at this stage could act as a mechanism of epigenetic memory that 'primes' epiblast cells and, following the exit from pluripotency, is propagated through later developmental stages until somatic tissue differentiation, thus playing a fundamental role in the regulation of cell

fate specification. Hence, the epiblast stage represents a critical window in which most of *de novo* DNAme occurs, highlighting a crucial function for the *de novo* DNA methyltransferases at this specific stage.

1.5 Aim of the project

Over the past decades, the results of several studies have shown that the correct establishment of de novo DNAme patterns is essential for mammalian development (108-111). In mouse, *Dnmt3a* knockout results in postnatal lethality, while *Dnmt3b* knockout results in embryonic lethality, indicating distinct biological functions played by the *de novo* DNMTs (112,113). Biochemical and structural evidence indicated that DNMT3A and DNMT3B exhibit preferences for flanking sequences (114). In ESCs DNMT3A has been shown mainly to methylate shores of bivalent CpG island canyons (115,116), while DNMT3B preferentially binds to the gene body of active genes (105,106). However, the molecular targets and the mechanisms that determine the specificity of the *de novo* DNAme machinery during cell fate specification are not yet completely elucidated.



II - Figure 1. 6: Epigenetic reprogramming during development. Schematic representation of the waves of DNAme reprogramming happening in the developing embryo and during the establishment of PGCs and gametogenesis – Adapted from (99).

In this work, we investigated the role of the *de novo* DNMTs in controlling lineage-fate decision during mouse early development. Using a combination of *in vitro* stem cell differentiation models, loss of function experiments and high-throughput multi-omics approaches – WGBS, ChIP-, bulk-, and single-cell-RNA sequencing -, we demonstrated that DNMT3B, but not DNMT3A, dependent methylation is essential for the correct specification of the meso-endodermal lineages. Our results showed that, in the transition from the naïve to primed pluripotency, DNMT3B activity is directed

towards regulatory regions associated with key developmental transcription factors, acting as an epigenetic priming that ensures flawless commitment at later stages. We found that the differentiation into meso-endodermal progenitors is impaired in DNMT3B knockout (3BKO) cells, which are redirected towards neuro-ectodermal lineages. Finally, we demonstrated that the impaired meso-endodermal induction of 3BKO cells can be rescued by silencing *Sox2*, a master regulator of neuronal differentiation.

Chapter 2

Lack of Dnmt3b impairs meso-endodermal lineage commitment in Embryoid Bodies

Results

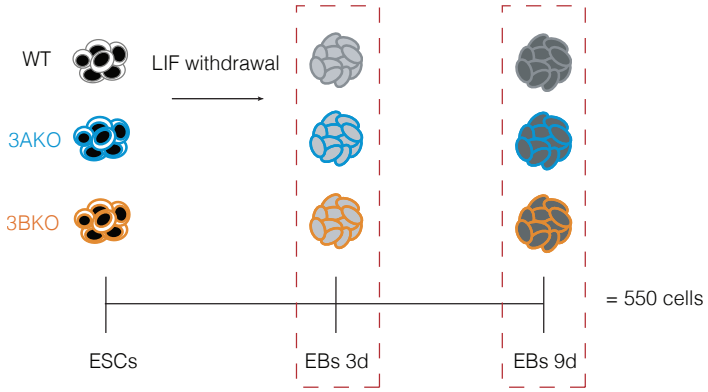
2.1 Differentiation of mESC in Embryoid Bodies

To dissect the function of *de novo* DNAme in early stages of development, we took advantage of previously generated *Dnmt3a* and *Dnmt3b* homozygous knockout cell lines from mouse embryonic stem cell (ESCs) line E14 (106). We maintained ESC in the undifferentiated and pluripotent state by means of culture medium supplemented with leukemia inhibitory factor (LIF) (85,86). To study early stages of development as well as to investigate the impact of a gene knockout *in vitro*, we differentiated *Dnmt3a* -/- (3AKO), *Dnmt3b* -/- (3BKO) and wild type (WT) ESCs into three-dimensional aggregates of cells called embryoid bodies (EBs) through a withdrawal of LIF (see Chapter 7, paragraph 7.1.3). ESCs within EBs spontaneously undergo differentiation and cell specification along the three primary germ layer lineages, reproducing *in vitro* the process of gastrulation and the molecular events associated with early embryogenesis (101,117). In fact, upon LIF withdrawal - and in agreement with what happens *in vivo* during gastrulation -, *Dnmt3a* and *Dnmt3b* show an increase in their expression levels (EBs 3d), which remain high until 9 days of EBs differentiation (EBs 9d), after which they start to decrease (101). Therefore, EBs differentiation of mouse ESCs represent a suitable *in vitro* model for the investigation of the regulatory role played by *de novo* DNAme in the control of embryonic lineage fate decisions.

2.2 Single-cell RNA-seq profiling of differentiating EBs

ESCs differentiating into EBs are a heterogeneous system in which individual cells make decisions about their fate. Therefore, they lend themselves to single-cell transcriptomics analyses to understand their biology and dissect the molecular events underpinning lineage fate choices that happen along their differentiation path (118). To this aim, we collected a total of 550 cells at two time points – day 3 and day 9 of EBs differentiation – from both WT and mutant differentiating

ESCs and performed single-cell RNA-sequencing (scRNA-seq, Smart-seq2 protocol, Figure 2.1, see Chapter 7, paragraph 7.1.20) to profile gene expression. After quality controls and filtering of low quality cells (see Chapter 7, paragraph 7.2.1), we retained a total of 465 high-quality cells for downstream analysis, with a median of > 200,000 reads per cell assigned to annotated transcripts.



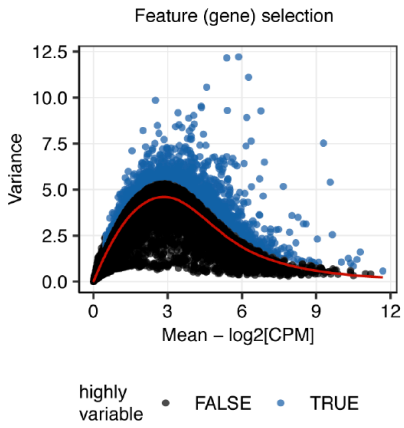
II - Figure 2. 1: Single-cell RNA-seq profiling of differentiating EBs. Schematic representation of EBs differentiation of mouse ESC. We differentiated WT, 3AKO and 3BKO ESC in EBs through LIF withdrawal and collected cells after 3 and 9 days of differentiation, for a total of 550 cells profiled by single-cell RNA-sequencing.

2.2.1 Cluster analysis of cell types

As first exploratory data analysis step, we performed dimensionality reduction of the single-cell gene expression profiles, in order to visualize in the reduced space the structure of our high-dimensional dataset. To do so, we initially performed a feature selection step to discard uninteresting or noisy features/genes, selecting the top 2000 highly-variable genes for the following analysis steps (Figure 2.2, see Chapter 7, paragraph 7.2.1).

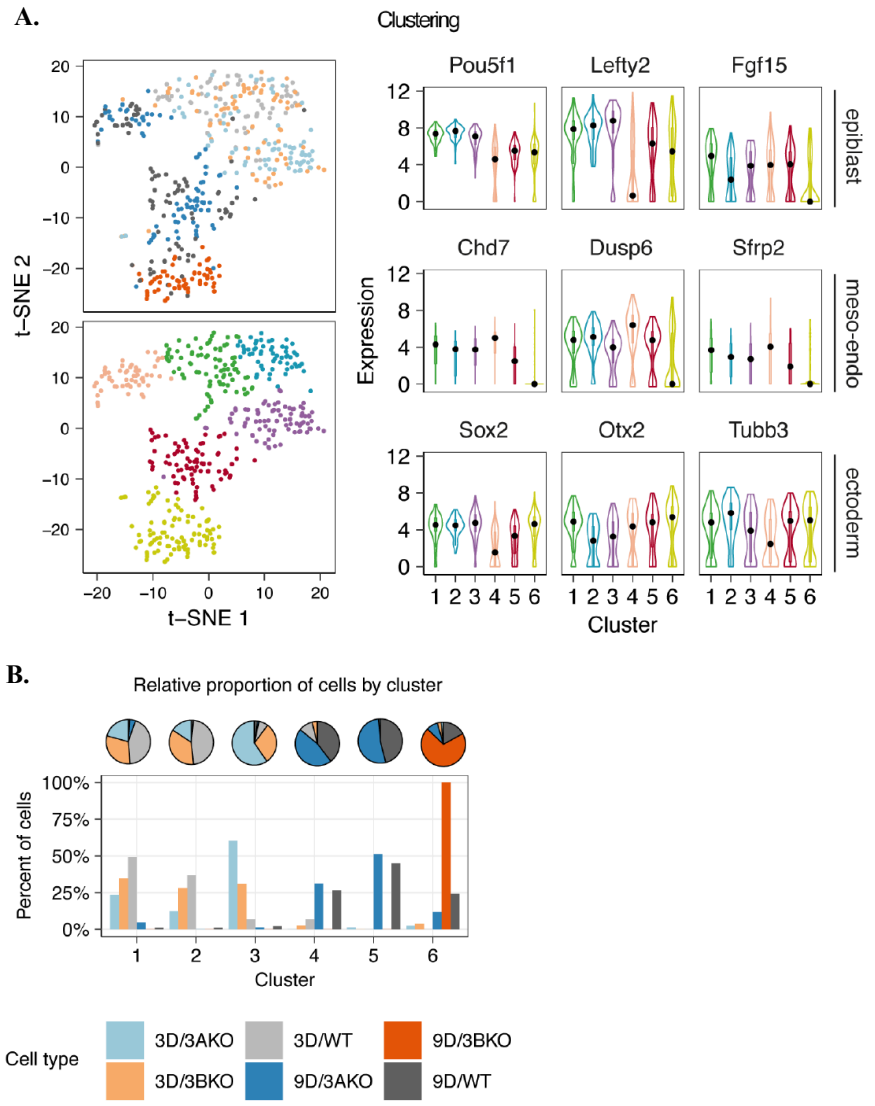
Next, we reduced the dimensionality of our dataset using Principal Component Analysis (PCA) followed by t-distributed Stochastic Neighbor Embedding (t-SNE) (Figure 2.3A, see Chapter 7, paragraph 7.2.1). Interestingly, the two first t-SNE components separated the two stages of differentiation and showed different patterns of segregation between WT and mutant cells (Figure 2.3): at day 3 of differentiation, the three populations are equally distributed in the 2D-reduced space, likely representing a more uniform cell state. In contrast, at day 9 of EBs differentiation, a clear separation emerges between 3AKO and 3BKO cells (Figure 2.2A, top-left panel). We next performed unsupervised clustering using a graph-based approach (Louvain clustering (119), see Chapter 7, paragraph 7.2.1) and identified six robust cell clusters (Figure 2.2A, bottom-left panel): three of them (i.e. cluster 1,2 and 3) were populated by cells at day 3 and three (i.e. cluster 4,5 and 6) by cells at day 9 of EBs differentiation, respectively. To understand the meaning of the obtained cell clusters, we identified cluster-specific marker genes using differential expression analysis, testing each cluster against all the remaining others (Figure 2.3A-right, see Chapter 7, paragraph

7.2.1). We observed that clusters at day 3 (i.e. cluster 1,2 and 3) are characterized by the expression of primed-pluripotency markers typical of the epiblast stage, such as *Lefty1*, *Lefty2*, *Pou3f1* and *Lef1*, whereas clusters at day 9 show the expression of lineage-specific markers, such as *Tubb3, Otx2, Sox2* (ectoderm), *Sox17, Krt19* (endoderm) and *T, Foxa2, Wnt8a* (mesoderm). The analysis of cell cluster composition reflected the differentiation stage and confirmed the segregation between 3AKO and 3BKO at day 9 (Figure 2.3B and Figure 2.4). In particular, while WT and 3AKO cells express markers of all lineages, the cell cluster populated by 3BKO cells (*i.e.* cluster 6) show a higher expression of pluripotency-associated markers as well as markers of the ectodermal lineage, thus suggesting an impairment of the differentiation trajectory toward meso-endoderm in the absence of *Dnmt3b*.

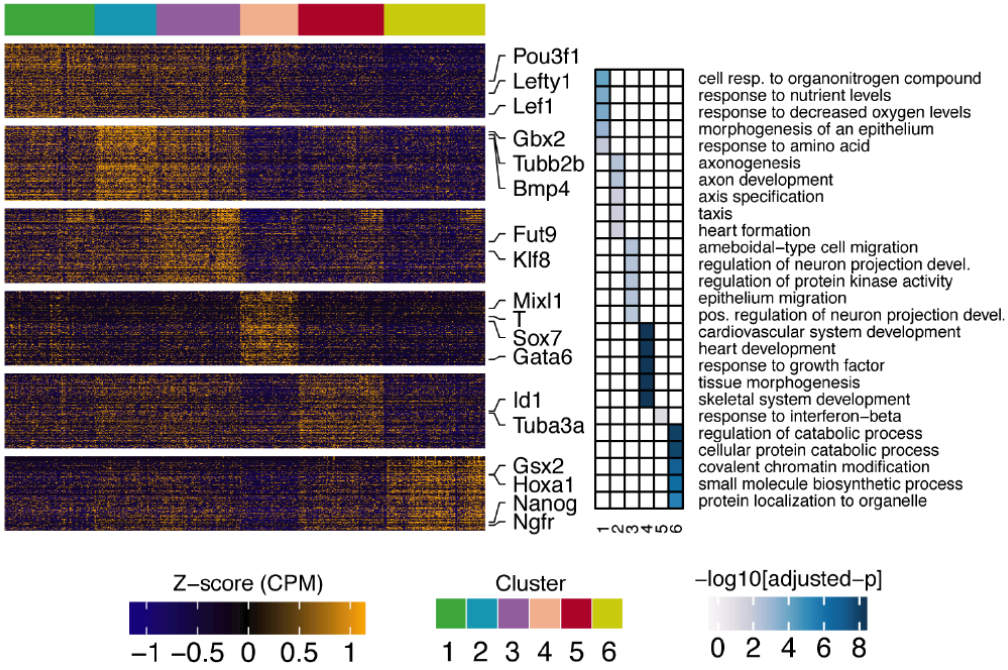


II - Figure 2. 2: Variance modeling for feature selection. Scatter plot showing the modelling of the per-gene mean-variance relationship used for feature selection (see Chapter 7, paragraph 7.2.1). Blue dots represent highly variable genes selected for downstream analysis (i.e. dimensionality reduction, clustering and differentiation trajectory reconstruction).

This hypothesis was further corroborated by the enrichment analysis of gene expression signatures for the identified clusters (Figure 2.4, see Chapter 7 paragraph 7.2.1). Indeed, we observed significant enrichment for gene ontology terms associated with tissues that are subsequently derived from mesoderm and endoderm (i.e. *heart development, skeletal system development*) in cell cluster 4, which is composed only by 3AKO and WT cells at day 9 of EBs differentiation, whereas cell clusters populated by 3BKO at day 9 (i.e. cell cluster 6) presented enrichment for pluripotency and neuro-ectodermal related GO terms, while not showing any significant enrichment for gene signatures related to meso-endodermal lineage commitment (Figure 2.4). The higher expression of pluripotency-associated genes in 3BKO cells (i.e. *Nanog*, Figure 2.4) is symptomatic of a differentiation defect associated with the loss of *Dnmt3b*.



II - Figure 2. 3: Cluster analysis of cell types. **A**. (left) t-SNE analysis of single-cell transcriptional profiles for WT, 3AKO and 3BKO cells, grouped into six cell clusters identified with the Louvain algorithm. (right) Gene expression levels distribution of epiblast (Pou5f1, Lefty2, Fgf15), meso-endodermal (Chd7, Dusp6, Sfrp2) and ectodermal markers (Sox2, Otx2, Tubb3) in the six identified cell clusters. **B.** (left) t-SNE analysis of single-cell transcriptional profiles for WT, 3AKO and 3BKO cells, grouped into six cell clusters identified with the Louvain algorithm. (right) Gene expression levels distribution of epiblast (Pou5f1, Lefty2, Fgf15), meso-endodermal (Chd7, Dusp6, Sfrp2) and ectodermal markers (Sox2, Otx2, Tubb3) in the six identified cell clusters.



II - Figure 2. 4: Gene expression signatures of cell clusters. (left) Heatmap showing expression patterns (scaled as Z-scores) for the top 200 marker genes identified in each cluster. (right) Heatmap showing enriched GO terms identified in each cluster markers list.

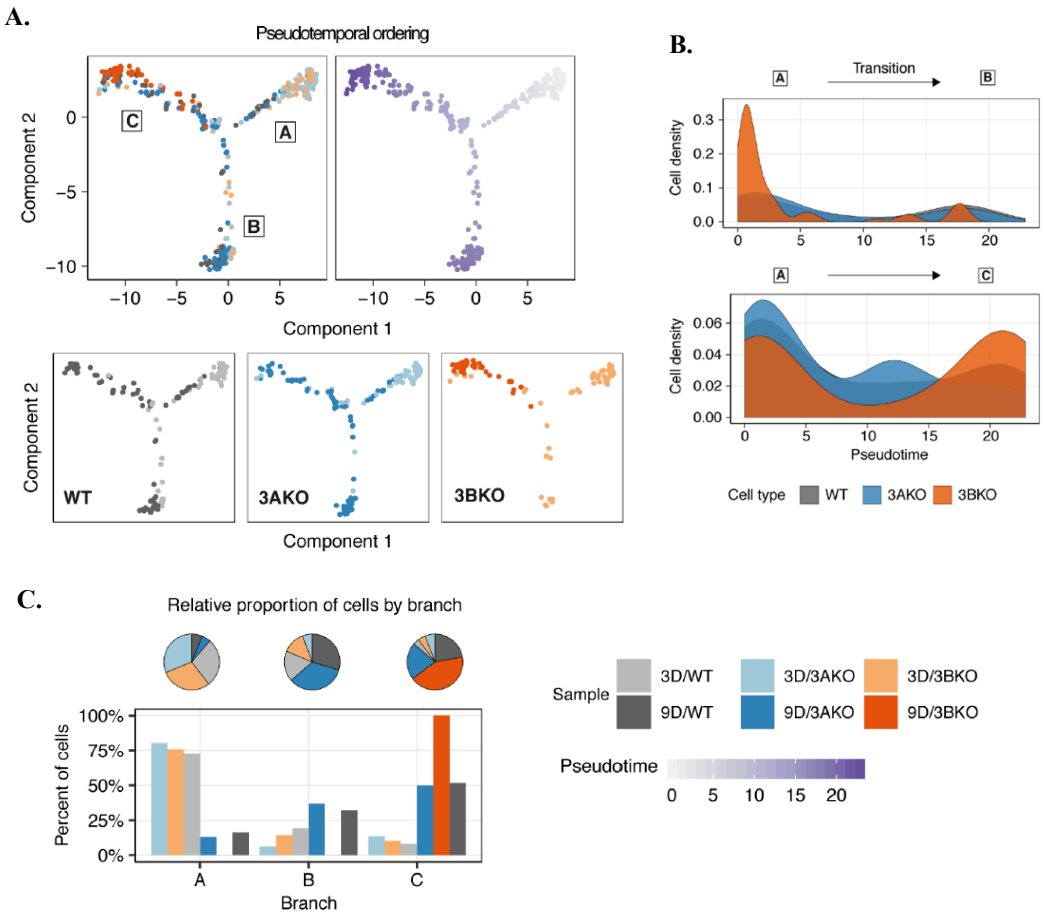
2.3 Impairment of meso-endoderm differentiation trajectory in 3BKO cells

2.3.1 Pseudo-temporal ordering of single-cell differentiation trajectories

To further inspect the functional impact of the loss of *Dnmt3a* and *Dnmt3b* during EBs differentiation, which is more like a continuous process of cells transforming over time, we performed pseudo-temporal ordering of single cell differentiation trajectories using reverse graph-embedding (119) (see Chapter 7 paragraph 7.2.1). The algorithm enabled the reconstruction of a tree connecting all the observed transcriptional cell states, revealing the existence of three major branches (A, B and C) ordered in pseudotime (Figure 2.5A top). Starting from branch A, equally populated by WT, 3AKO and 3BKO cells at day 3 (*i.e.* pluripotent/epiblast stage, Figure 2.5B), we observed that, at day 9, a comparable proportion of WT and 3AKO cells is found in both differentiation branches B and C, while the totality of 3BKO cells is redirected toward branch C (Figure 2.5A bottom, Figure 2.5B and Figure 2.5C).

2.3.2 Branch-dependent analysis of gene expression

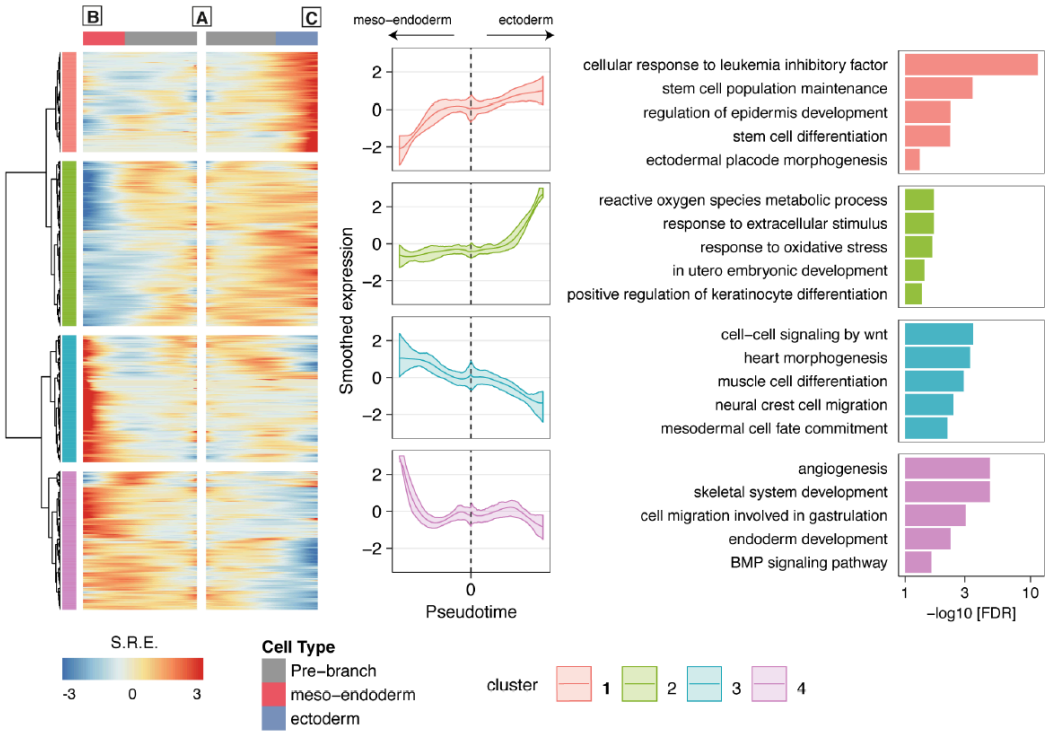
In order to characterize the reconstructed branches, we performed branch-dependent analysis of gene expression (BEAM) (120) (see Chapter 7 paragraph 7.2.1) and identified 1,467 genes ($FDR \leq 0.001$) whose expression is significantly dependent from the branching point (Figure 2.6-right).



II - Figure 2. 5: Pseudo temporal ordering of single cell differentiation trajectories in EBs. **A.** Pseudotime analysis of cell differentiation trajectories with DDRTree *(119)*. The analysis revealed 3 branches (A, B and C), ordered in pseudotime. (bottom) Visualization of the reconstructed differentiation trajectories for each source cell type (WT, 3AKO and 3BKO). **B.** Cell density plots showing the distributions of each source cell type (i.e. WT, 3AKO and 3BKO) along the reconstructed pseudotransition from branch A to branch B (top) and branch A to branch C (bottom). The plot shows the unbalancing in cell densities of 3BKO cells, which are all redirected to branch C at day 9. **C.** (top) Pie charts representing the composition of each branch of the reconstructed differentiation trajectories in terms of source cell types (WT, 3AKO and 3BKO at 3D and 9D), with respect to the total number of cells in each branch. (bottom) Bar plots depicting the relative proportion of WT, 3AKO and 3BKO cells from each time point (3D and 9D) found in each branch of the reconstructed differentiation trajectories, with respect to the total number of source cell type.

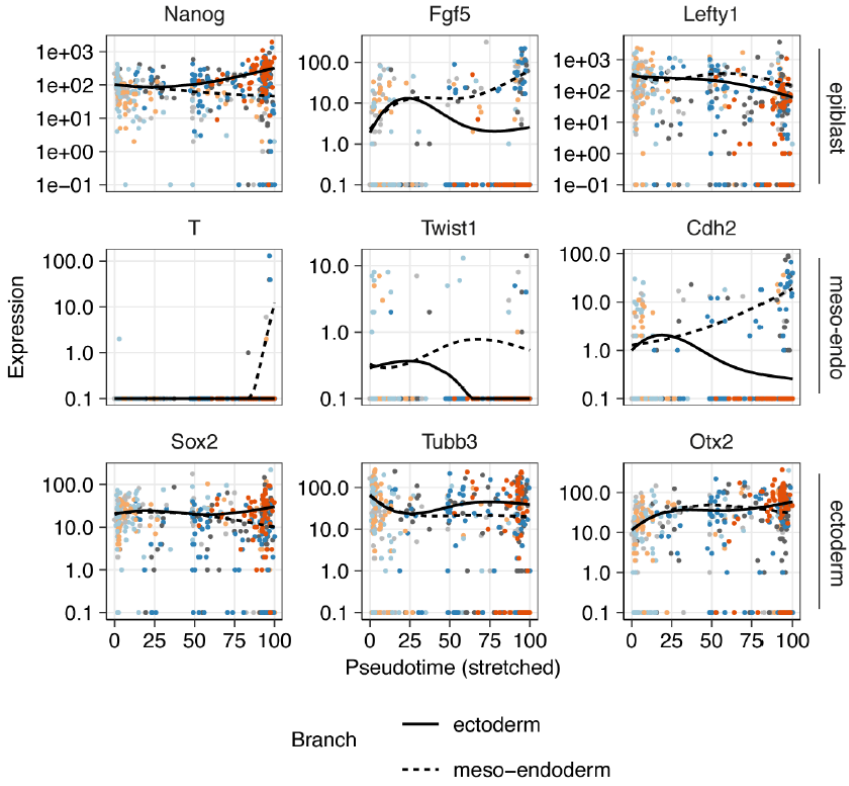
These genes were next clustered hierarchically into four clusters, reflecting distinct dynamical patterns of gene expression for the reconstructed branches (Figure 2.6-right): in particular, cluster 1 and 2 are characterized by genes that are upregulated when going from A to C and downregulated in the opposite branch (B), and show enrichment for gene ontology terms associated with neuro-ectodermal lineage (Figure 2.6-right). Among them, we observe the presence of well-known markers of neuro-ectodermal cell identity such as *Sox2*, *Tubb3* and *Otx2* (Figure 2.7). In contrast, cluster 3 and 4 are composed by genes that are upregulated when going from A to B and downregulated in the opposite branch (C), and show enrichment for gene ontology terms

associated with meso-endodermal lineage (Figure 2.6-right), with the noticeable presence of known meso-endodermal markers such as *T*, *Twist1* and *Cdh2* (Figure 2.7). These results indicate that the three reconstructed branches are associated to different stages of differentiation and cell fates, i.e. epiblast (A), meso-endodermal (B) and ectodermal (C) (Figure 2.5A and Figure 2.7), and the analysis of the reconstructed trajectories of each cell population confirmed the arrest of differentiation toward the meso-endodermal lineages for 3BKO cells, while 3AKO cells differentiate in both the meso-endodermal and ectodermal cell fates (Figure 2.5, 2.6 and 2.7). Taken together, these data reveal that *Dnmt3b*-dependent methylation is essential for the correct specification of the meso-endodermal lineages. In contrast, 3AKO cells are still able to form all the three germ layers, suggesting that the loss of *Dnmt3a*-dependent methylation has little functional consequences at this early stage of differentiation. The impaired differentiation trajectory towards meso-endoderm that we observed for 3BKO cells suggests a peculiar role for *Dnmt3b* in the epigenetic control of this specific cell fate, which is the focus of the following chapters.



II - Figure 2. 6: Branch-dependent analysis of gene expression. (left) Heatmap visualization of the results of BEAM (120) analysis for the 1,467 genes (FDR<0.001) whose expression is significantly dependent from the branching point, demonstrating the bifurcation of gene expression for the meso-endodermal and ectodermal lineage commitment. Columns are points in pseudotime, rows are genes, and the beginning of pseudotime is in the middle of the heatmap. The genes are hierarchically clustered into 4 clusters having similar lineage-dependent expression patterns. (middle) Summary plots depicting the average relative expression levels in the branched bifurcation for each cluster. The beginning of pseudotime is in the middle of the plot (time 0). The shaded area represents the interquartile range, while the top, middle and bottom lines represent the

first quartile, the mean and the third quartile, respectively. (right) Bar plots of selected gene ontology terms for enriched biological processes in each BEAM cluster.



II - Figure 2. 7: Branch-dependent expression of lineage marker genes. Gene expression levels of epiblast (*Nanog*, *Lefty1*, *Fgf5*), meso-endodermal (*T*, *Twist1*, *Cdh2*) and ectodermal markers (*Sox2*, *Otx2*, *Tubb3*) as a branch-dependent (i.e. ectoderm, meso-endoderm) function of pseudotime. The expression patterns of well-known lineage marker genes show the arrest of differentiation towards meso-endoderm for 3BKO cells.

Chapter 3

Loss of Dnmt3b does not affect the formation of EpiSCs, but is required for the differentiation towards meso-endodermal progenitors

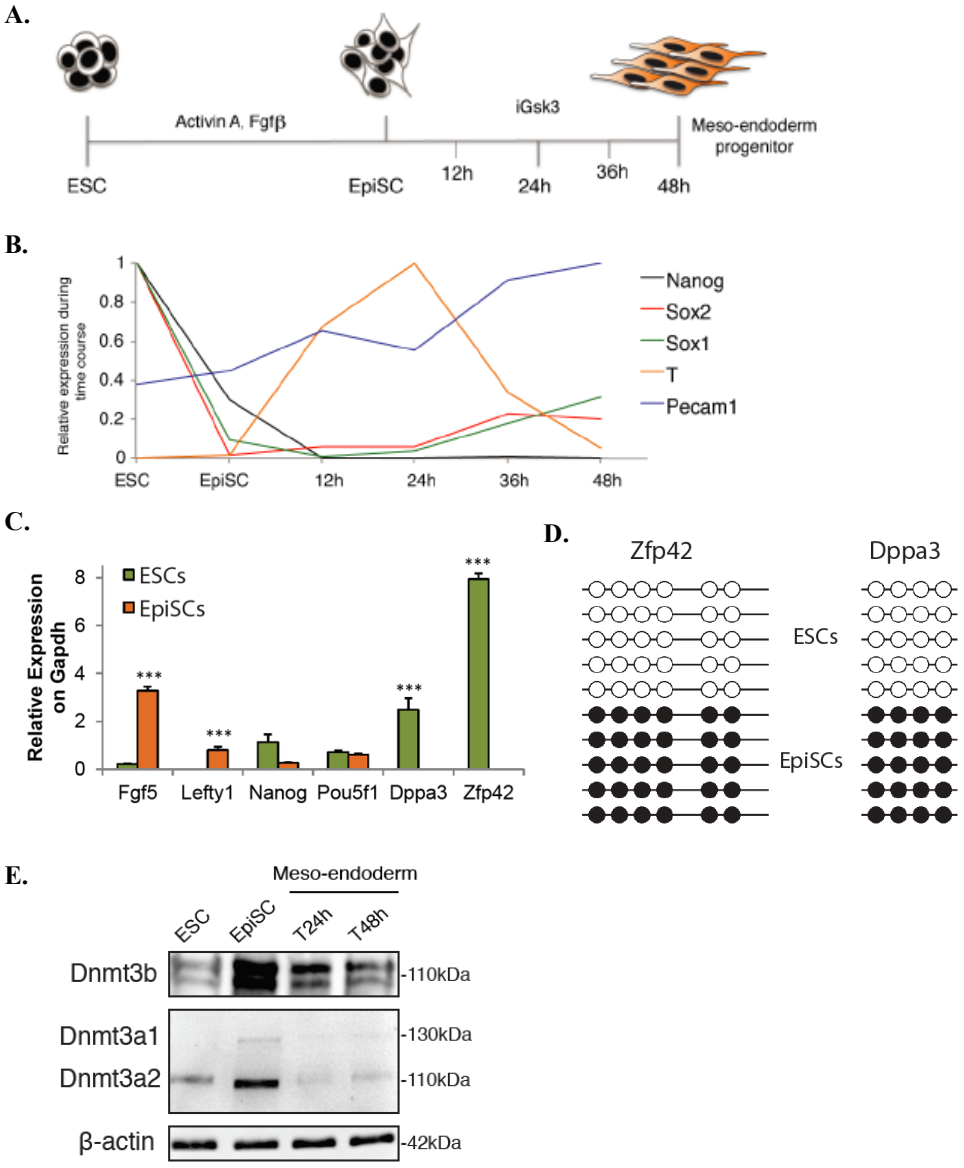
Results

3.1 Directed differentiation of ESCs towards meso-endodermal progenitors

In light of the results of scRNA-seq profiling of EBs, which showed an impairment of meso-endoderm differentiation for 3BKO cells (see Chapter 2), and to gain further insights into the functional role of *Dnmt3b*-dependent DNAme in resolving early lineage commitment, we switched to a different *in vitro* model involving two steps of directed differentiation (Figure 3.1A): first, upon *Activin/Nodal* pathway stimulation, ESCs are differentiated into a stable line of Epiblast Stem Cells (EpiSCs) (see Chapter 7, paragraph 7.1.4), reproducing the early post-implantation embryo state (E5.5), where most of the ICM is composed by 'primed' pluripotent stem cells, and a crucial step for the establishment of *de novo* DNAme patterns occurs (see Chapter 1). Next, by means of *WNT/TGF-β* pathway activation through inhibition of *Gsk3β* (94), EpiSCs are committed toward the meso-endodermal progenitor fate (see Chapter 7, paragraph 7.1.5).

To characterize our two-step differentiation model, we analyzed in WT cells the expression levels of stemness/pluripotency as well as lineage-specific marker genes by qRT-PCR. Indeed, ESCs and EpiSCs share key transcription factors at comparable levels, such as *Pou5f1,* which is instead downregulated upon meso-endodermal commitment (Figure 3.1B, C). Conversely, EpiSCs show a strong downregulation of the pluripotency markers *Zfp42*, *Klf4* and *Esrrb*, while upregulating transcripts like *Lefty1* and *Fgf5*, which are markers of the primed pluripotency stage (121) (Figure 3.1C). Moreover, as expected (121,122), the promoters of *Zfp42* and *Dppa3a* acquire de novo methylation in EpiSCs, while being hypo methylated in ESCs, thus confirming the epiblast identity of our induced EpiSCs (Figure 3.1D). At the time of meso-endodermal induction, we can observe (i) an increase in the expression levels of meso-endodermal markers such as *T* and *Pecam1* (Figure 3.1B), (ii) the downregulation of stemness marker genes, such as *Sox2* and *Nanog* (down-regulated

Figure 3.1B), and (ii) no expression of markers of the other lineages (i.e. the neuro-ectodermal marker *Sox1*, Figure 3.1B), thus confirming the accuracy of directed differentiation model.



II - Figure 3. 1: Directed differentiation of ESCs towards meso-endodermal progenitors. **A.** Schematic representation of the two-step differentiation model: from ESC to EpiSC and from EpiSC to Meso-endodermal progenitors, indicating the time points of RNA collection and the molecules used for the differentiation induction. **B.** qRT-PCR analysis of selected markers genes for the ESC (*Nanog*, *Sox2*), Meso-endoderm (*T*, *Pecam1*) and Ectodermal (*Sox1*, *Sox2*) lineages. **C.** qRT-PCR analysis of selected marker genes characterizing the ESC to EpiSC transition. *** = *p*-value < 0.001, *t*-test. **D.** Bisulfite analysis of *Zfp42* and *Dppa3* promoters' DNA me levels during the ESC to EpiSC transition. Each circle represents a CpG. White dot = unmethylated, black dot = methylated. The two promoters acquire *de novo* DNAme in EpiSC. **E.** Western blot analysis of *Dnmt3a* (both isoforms 1 and 2) and *Dnmt3b* expression during the ESC-EpiSC-Meso-endoderm differentiation time course.

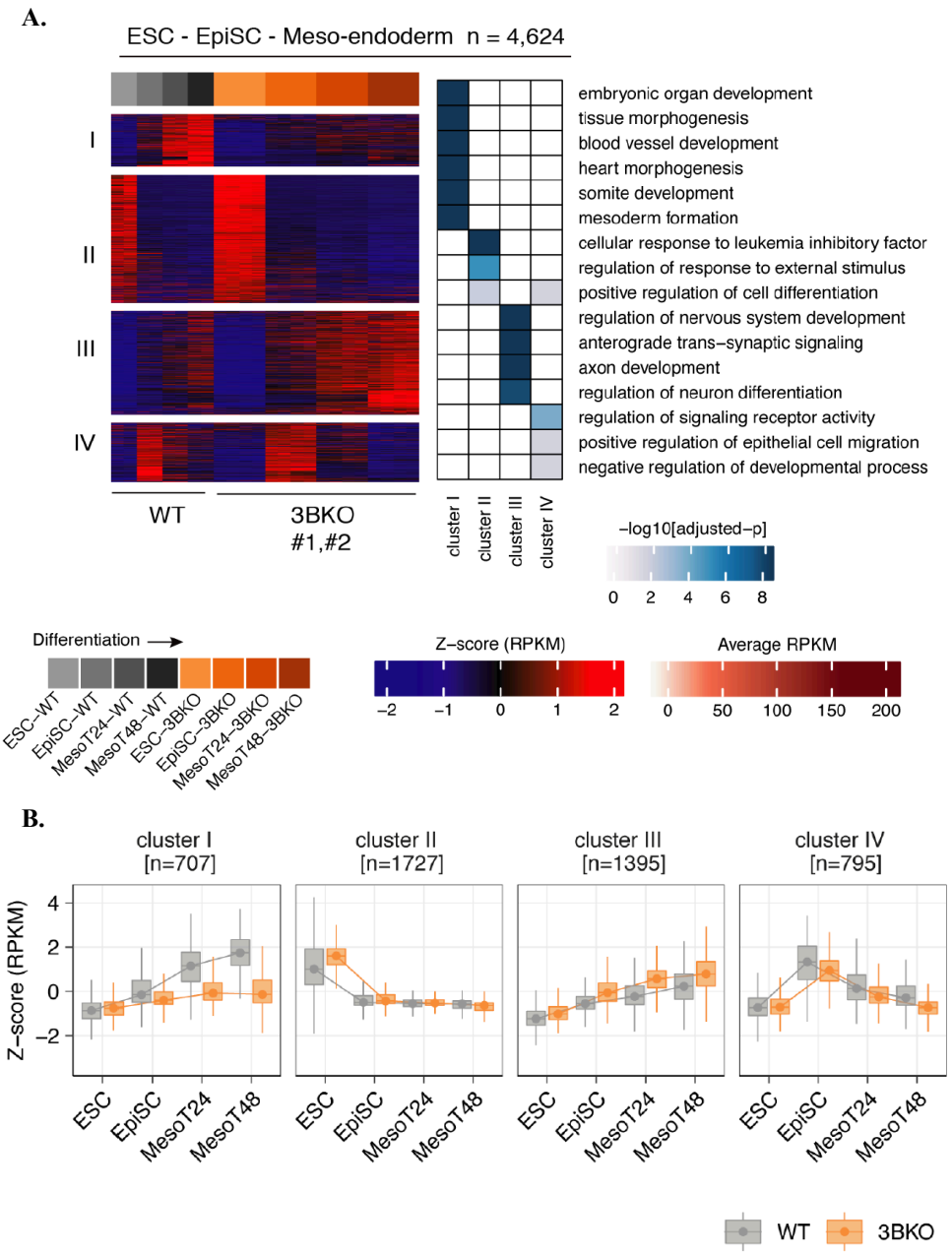
Interestingly, the two *de novo* DNA methyltransferases display - as it is thoroughly described (121,122) - a significant upregulation in EpiSCs formation (Figure 3.1E), but show an opposite profile during meso-endoderm commitment: in fact, both *Dnmt3a* isoforms are rapidly downregulated, while *Dnmt3b* remains expressed (Figure 3.1E), thus supporting its observed role in this specific differentiation (see Chapter 2).

3.2 Transcriptome profiling of WT and 3BKO cells in the ESC-EpiSC-Meso-endoderm transition

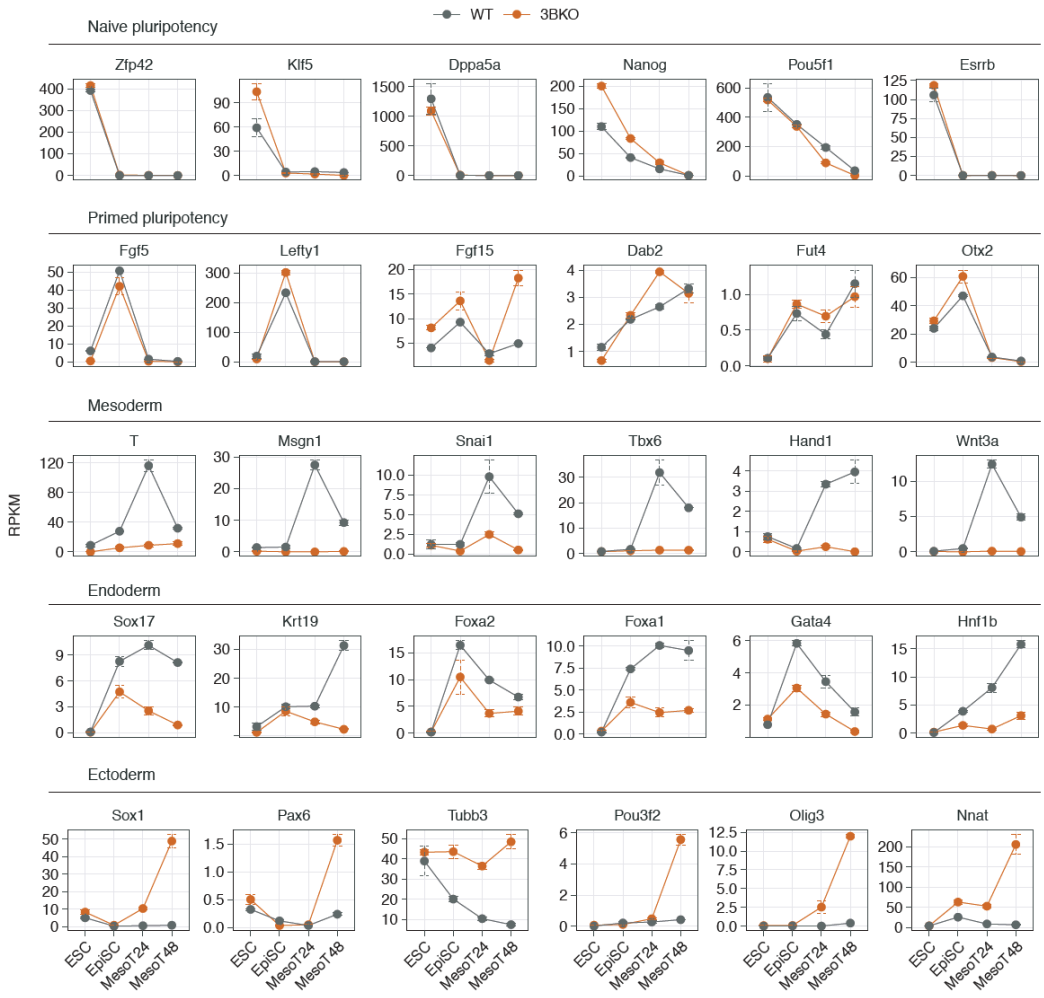
To investigate the effects of *Dnmt3b* loss in the transcriptional program that controls meso-endoderm specification, we performed gene expression profiling using bulk RNA sequencing (RNA-seq) on WT and two independent 3BKO clones (106) committed to meso-endoderm with our two-step differentiation approach. This analysis revealed 4,624 differentially expressed genes arising over the differentiation time course (*i.e.* in any sample group, edgeR (31) ANOVA-like test for any difference, $|logFC| \geq 1.5$ and $FDR \leq 0.001$, see Chapter 7 paragraph 7.2.2), which we grouped into 4 gene clusters recapitulating distinct patterns of gene expression dynamics for WT and 3BKO cells (Figure 3.2). Cluster II and IV are populated by genes that show a similar expression pattern in both WT and 3BKO cells. In particular, cluster II is characterized by genes that are stably downregulated after the exit from the ESC-pluripotent state, shows enrichment for gene ontology terms associated with pluripotency (e.g. *cellular response to LIF*, Figure 3.2) and the presence of well-known Naïve pluripotency markers such as *Nanog*, *Esrrb*, *Dppa3a* and *Zfp42* (Figure 3.3). Therefore, this gene cluster defines the ESC ground state, from which both WT and 3BKO cells are induced to exit when receiving the differentiation stimuli. Cluster IV is defined by genes that are upregulated at the exit from the ESC-pluripotent state, reach their peak at EpiSC stage and are downregulated after the meso-endodermal induction; it shows enrichment for signaling pathways associated with the epiblast stage (e.g. *positive regulation of epithelial cell migration*, Figure 3.2) as well as the presence of primed pluripotency markers such as *Fgf5*, *Lefty1*, *Otx2* and *Fgf15* (121) (Figure 3.3). Thus, this gene cluster defines the EpiSC stage, where both WT and 3BKO cells express key epiblast markers and TFs at comparable levels (Figure 3.2-3.3, see paragraph 3).

On the other hand, cluster I and III are populated by genes that show a divergent expression pattern between WT and 3BKO cells. Specifically, cluster I is composed by genes that are upregulated in WT cells after meso-endodermal commitment and downregulated in 3BKO cells with respect to WT, and it shows enrichment for gene ontology terms associated with mesoderm formation (e.g. *somite development*, *mesoderm formation*, Figure 3.2) as well as the presence of key mesoderm (*T*, *Msgn1*, *Snai1*, *Hand1*, Figure 3.3) and endoderm (*Sox17*, *Krt19*, *Foxa1*, Figure 3.3) marker genes. Cluster III is composed by genes that are upregulated in 3BKO cells with respect to WT after meso-endodermal commitment, and it shows enrichment for gene ontology terms associated

with ectodermal commitment and neuronal development (e.g. *regulation of neuron differentiation*, *axon development*, Figure 3.2) and the presence of known ectodermal markers such as *Sox1*, *Pax6*, *Tubb3* and *Pou3f1* (Figure 3.3).



II - Figure 3. 2: Transcriptome profiling of WT and 3BKO cells in the ESC-EpiSC-Meso-endoderm differentiation. **A.** (left) RNA-seq heatmap showing the results of gene expression profiles clustering with K-means for WT and 3BKO cells during the complete differentiation time course (ESC-EpiSC-Meso-endoderm). Differentially expressed genes arising during the differentiation time course in any group were identified by ANOVA-like test with edgeR *(31)* ($|logFC| \geq 1.5$ and $FDR \leq 0.05$). Rows are genes, columns are samples and the scaled expression level (Z-score) is plotted. (right) Heatmap showing selected gene ontology terms for enriched biological processes in each cluster. **B.** WT vs 3BKO gene expression dynamics for each of the reconstructed gene clusters over the differentiation time course. Boxplot show the gene expression levels distribution (Z-scores) at each time point. Line plots show the median trend.

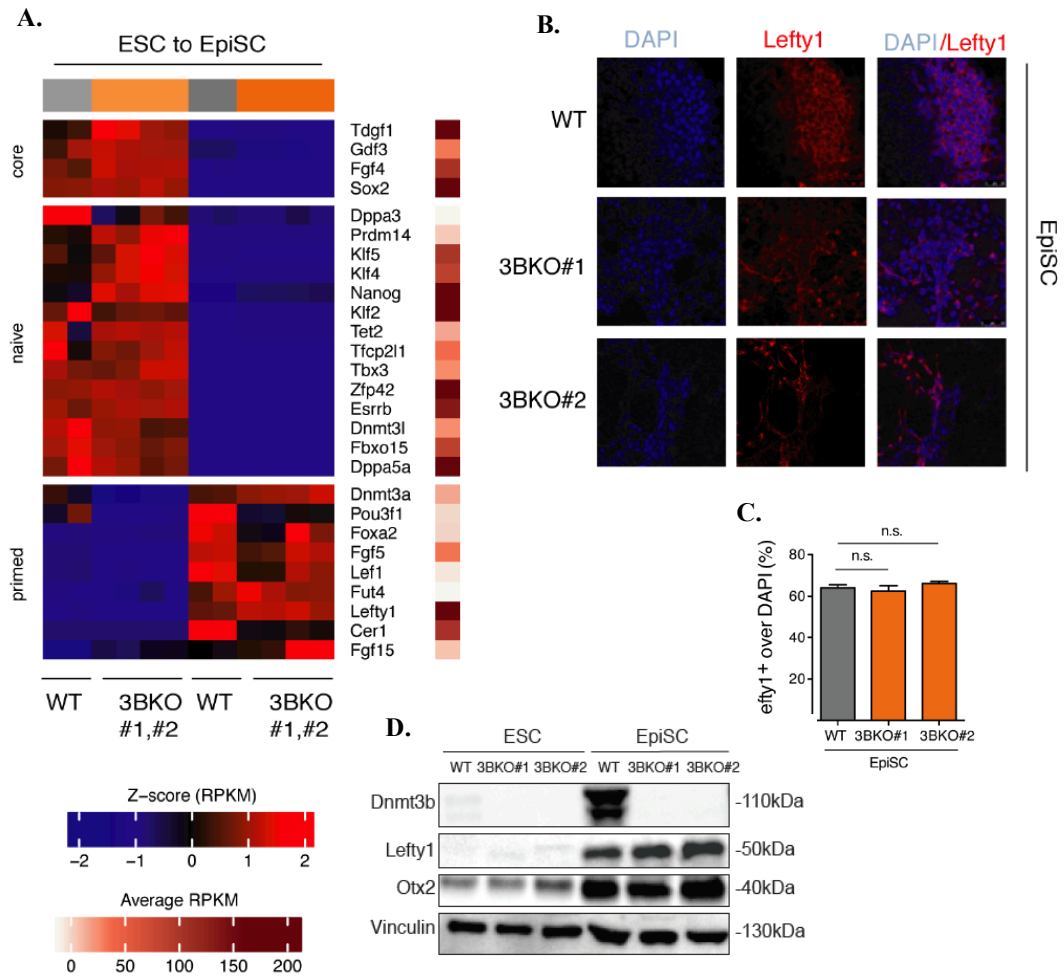


II - Figure 3. 3: Gene expression time-course of stage specific marker genes. RNA-seq gene expression time-course for stage-specific pluripotency (Naïve, Primed) and germ layers (Mesoderm, Endoderm, Ectoderm) markers genes. Dots represent normalized RPKM values, averaged by replicates/condition. Error bars represent standard errors. The analysis showed that both WT and 3BKO cells, upon differentiation induction, downregulate naïve pluripotency markers and upregulate primed pluripotency markers, thus reaching the epiblast stage. In contrast, a divergent trajectory emerges for WT and 3BKO cells once committed to meso-endodermal progenitors, with the redirection of 3BKO cells towards the neuro-ectodermal transcriptional program.

Overall, this analysis recapitulates the results obtained in the EBs differentiation model (see Chapter 2), confirming that 3BKO cells compared to WT show an impaired expression of meso-endodermal genes (cluster II) and are redirected towards a neuro-ectodermal-like transcriptional program (cluster III).

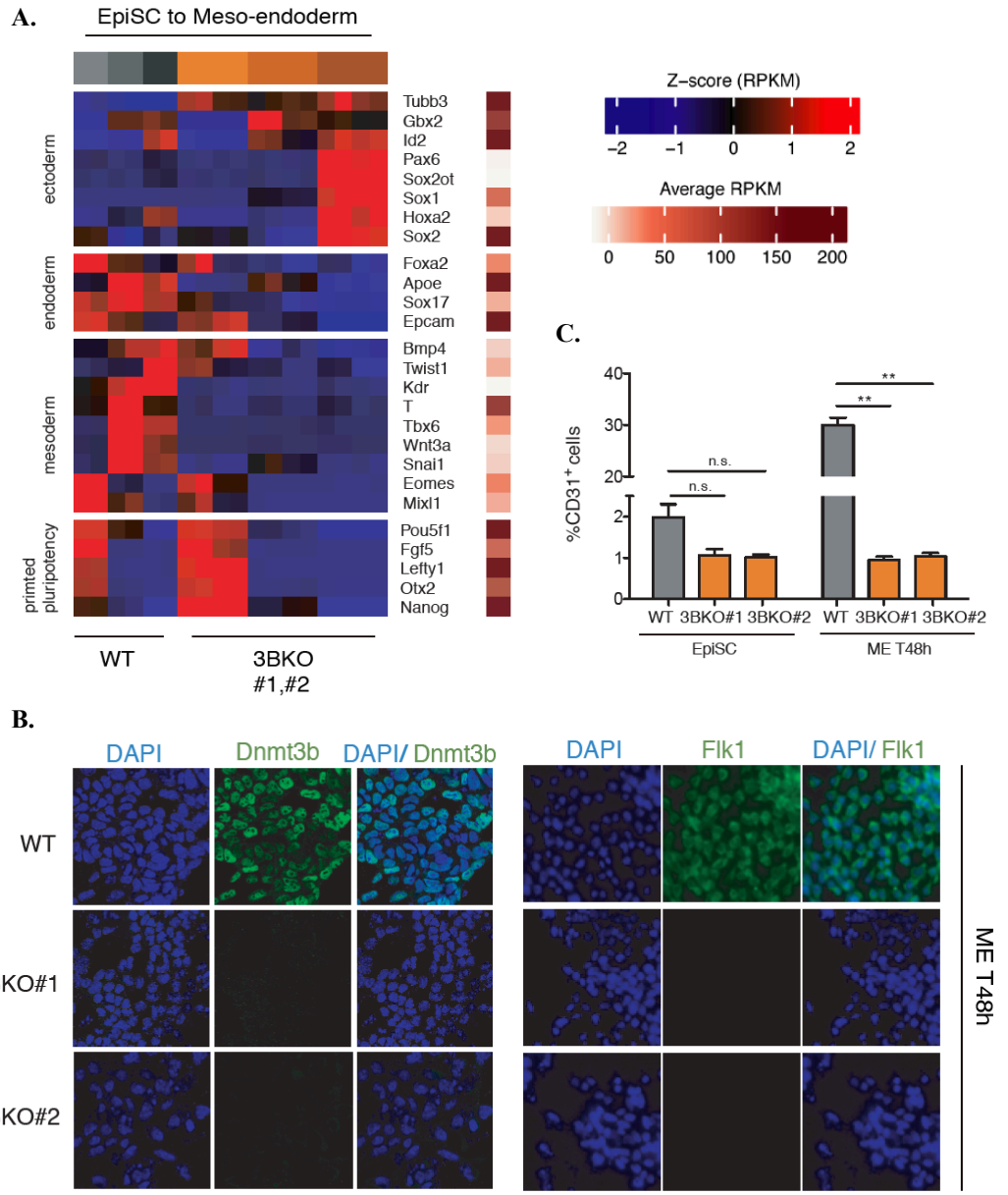
3.3 Loss of *Dnmt3b* does not affect the induction of EpiSCs, while impairs the formation of meso-endodermal progenitors

In order to deeply understand the previous result, we analyzed separately the two-step differentiation. Despite RNA seq analyses of ESCs-EpiSCs transition identified 906 genes upregulated and 426 downregulated between WT and 3BKO, we observed a similar expression of core (e.g. *Pou5f1, Nanog, Sox2*), naive (e.g. *Klf4, Dppa3, Prdm14*) and primed (e.g. *Otx2, Lefty1, Dnmt3a*) pluripotency genes (Figure 3.4A), with no morphological major changes between WT and mutant cells at the EpiSC stage. Thus, even in presence of a significant transcriptional response, the loss of *Dnmt3b*-dependent *de novo* DNAme has a minor impact in the formation of the epiblast-like phenotype. These results were confirmed at the protein level by western blot analysis of *Otx2* (Figure 3.4D) and immunofluorescence of *Lefty1* epiblast marker genes (88,121) (Figure 3.4B-C), which are expressed at comparable levels in both WT and 3BKO cells.



II - Figure 3. 4: Loss of Dnmt3b does not affect the induction of EpiSCs. **A**. RNA-seq heatmap showing the expression (Z-scores) of selected pluripotency marker genes during the transition from ESC to EpiSC for WT and 3BKO cells. Core and naïve pluripotency markers are downregulated, while primed pluripotency markers are upregulated in both WT and 3BKO cells. **B**. Representative

IF images of WT and 3BKO cells with the primed pluripotency marker Lefty1. **C.** Quantification of Lefty1^{+} cells as percentages of DAPI for both WT and 3BKO cells. **D.** Western blot analysis of the primed pluripotency markers *Lefty1* and *Otx2* during the ESC to EpiSC transition for WT and 3BKO cells.



II - Figure 3. 5: Lack of Dnmt3b impairs the formation of meso-endodermal progenitors. **A.** RNA-seq heatmap showing the expression (Z-scores) of selected pluripotency and differentiation (mesoderm, endoderm and ectoderm) marker genes during the transition from EpiSC to Meso-endodermal progenitors for WT and 3BKO cells. Meso-endodermal markers are upregulated in WT cells and downregulated in 3BKO cells, which instead show higher expression of ectodermal markers. Primed pluripotency markers are downregulated in both WT and 3BKO cells. **B**. Representative IF images of WT and 3BKO cells with *Dnmt3b* and the meso-endodermal marker *Flk1/Kdr*. **C**. FACS analysis of the mesodermal surface marker CD31 in both WT and 3BKO cells. ** = p-value < 0.05, t-test.

In contrast, the analysis of the expression of well-known lineage marker genes confirmed the divergent differentiation trajectory between WT and 3BKO cells after meso-endodermal commitment (Figure 3.5). Specifically, while prime pluripotency marker genes such as *Lefty1*, *Fgf5* and *Nanog* are downregulated in both WT and 3BKO cells, only WT cells show the induction of key mesoderm (*T*, *Snai1*, *Eomes*) and endoderm (*Sox17*, *Foxa2*) marker genes and TFs, while 3BKO cells upregulate markers of ectodermal lineage commitment (*Sox1*, *Sox2*, *Tubb3*, *Pax6*) (94) (Figure 3.5A). These results were confirmed by IF analysis of the mesoderm marker *Flk1*/*Kdr*, which is expressed only in WT cells at 48 hours of meso-endoderm induction (Figure 3.5B), as well as by FACS analysis of the mesoderm surface marker *CD31*/*Pecam1*, which showed a significant reduction of $CD31^+$ cells in 3BKO samples as compared to WT ones (Figure 3.5C).
Taken together, these data show that the *Dnmt3b*-dependent de novo DNAme does not affect the differentiation of ESC in EpiSC, but is required for the correct specification of the meso-endodermal progenitors at the time of gastrulation, thus corroborating also in a directed differentiation model the phenotype observed in EBs (see Chapter 2).

Chapter 4

Dnmt3b-dependent de novo DNAme primes EpiSCs for lineage commitment at later stages

Results

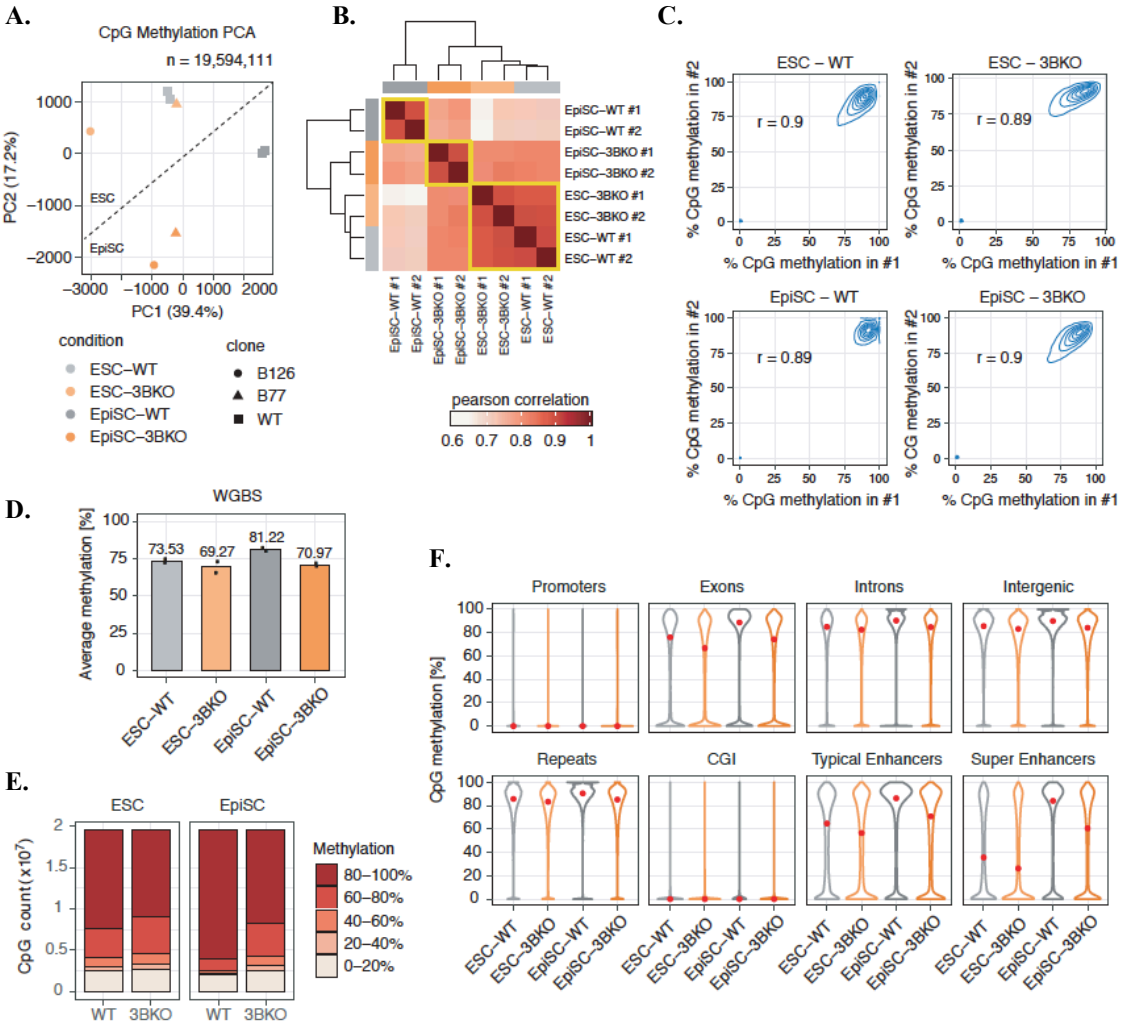
4.1 Methylome profiling by Whole Genome Bisulfite Sequencing

In order to obtain a detailed map of the DNAme landscape set up during the transition from naïve to primed pluripotency, as well as to investigate the effects of *Dnmt3b* loss on the establishment of *de novo* DNAme, we generated deep whole genome bisulfite sequencing (WGBS) data in WT and 3BKO cells at the two corresponding time points (*i.e.* ESC and EpiSC stage, two biological replicates for each condition, Pearson's $r \geq 0.89$ for all replicates, Figure 4.1 A, B, C), obtaining on average 450 millions paired-end reads and a mean coverage depth of ~ 23x per sample (see Chapter 7, paragraph 7.1.19 and 7.2.4), and covering a total of ~20 millions CpG sites in the mouse genome (minimum depth > 5x). The analysis of global DNAme levels showed a significant increase in DNA methylation in EpiSC as compared to ESC, recapitulating the in vivo dynamics, that is impaired in 3BKO cells (Figure 4.1 D, E). In fact, we identified a significant loss of global DNAme levels in 3BKO cells with respect WT at both the ESC and EpiSC stages (Figure 4.1D). Moreover, by PCA (Figure 4.1A) and clustering analysis of global methylation levels (hierarchical clustering by Pearson's correlation, Figure 4.1B), we observed a clear separation between 3BKO and WT cells, with a stronger effect visible at the EpiSC stage, where 3BKO cells fail to acquire proper *de novo* methylation, keeping their global DNAme levels close to the ones of naïve ESC (Figure 4.1D, E)

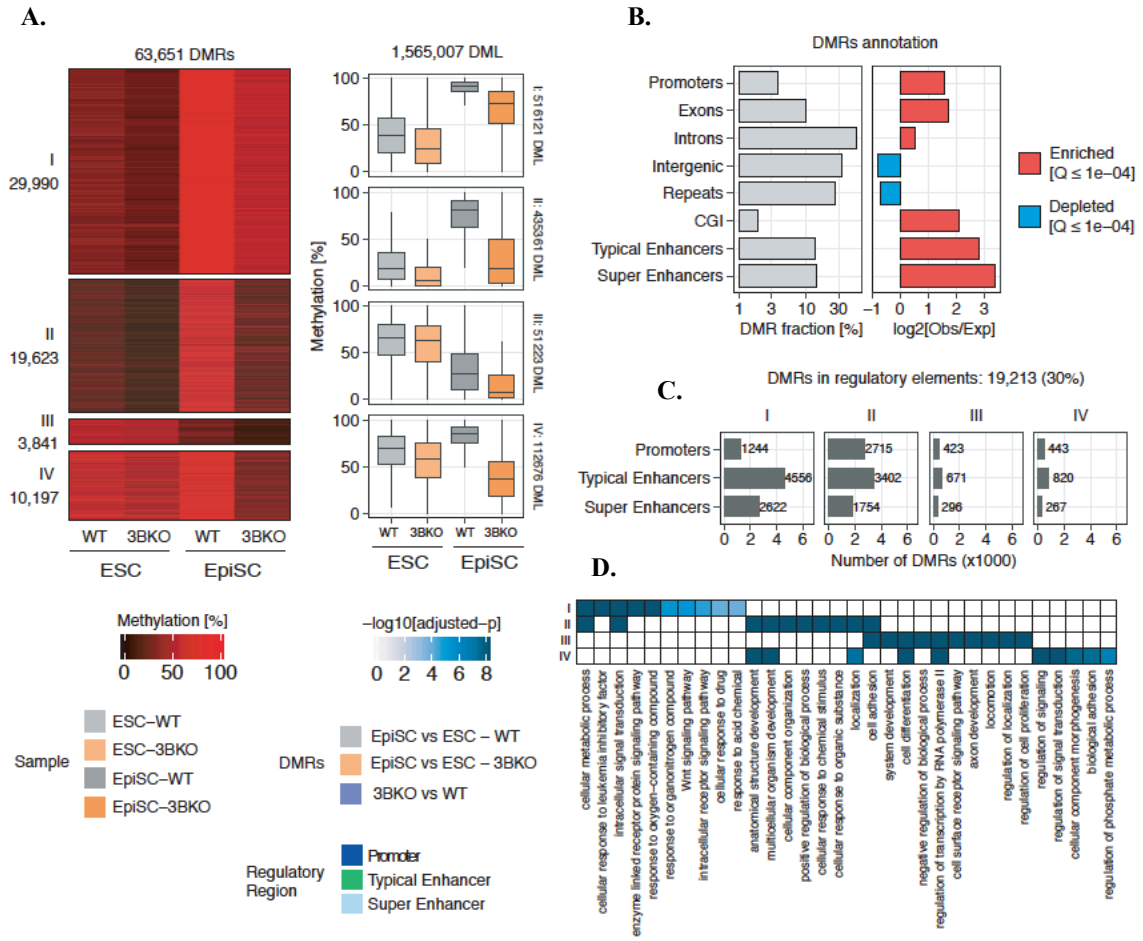
4.2 Lack of Dnmt3b impairs the *de novo* DNAme dynamics in EpiSCs

We next focused on the analysis of genomic regions with a dynamic DNAme sate during this transition. To do so, we performed differential methylation analysis and identified 63,651 differentially methylated regions (DMRs) that arise over the ESCs to EpiSCs induction

($methylation\ difference \geq 20\%, posterior\ p.value \leq 0.05$, see Chapter 7, paragraph 7.2.4), containing a total of 1,565,007 differentially methylated CpG loci (DML). The DMRs were next grouped into 4 major clusters showing distinct patterns of DNAme for WT and 3BKO cells (Figure 4.2A, see Chapter 7, paragraph 7.2.4). Interestingly, besides differences in basal levels of DNAme at the ESC stage, cluster I, II and IV – which account for ~ 93% of all the identified DMRs - are composed by regions that exhibit a gain of DNAme moving from the ESCs to the EpiSCs stage in WT cells, but fail to get properly methylated in 3BKO cells: therefore, these DMRs are *de novo* methylated by *Dnmt3b* during the ESCs to EpiSCs transition.



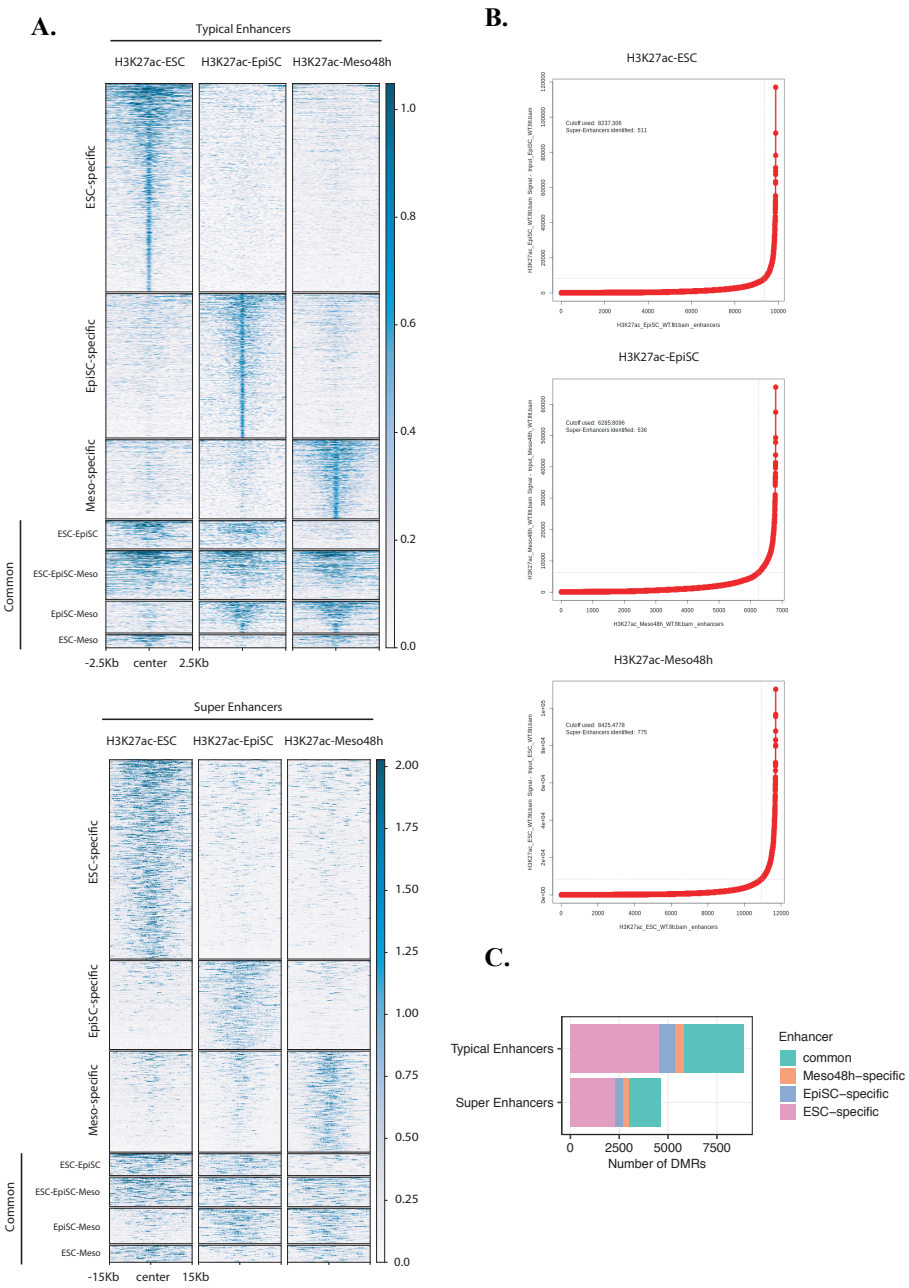
II - Figure 4. 1: Methylome profiling by Whole Genome Bisulfite Sequencing (WGBS). **A.** PCA of DNA methylome profiles obtained by WGBS, showing clustering of profiled samples by condition (3BKO, WT) and differentiation stage (ESC, EpiSC). **B.** Heatmap showing clustering of WGBS samples by Pearson correlation of single-base resolution CpG methylation levels. **C.** Density plot showing pairwise Pearson correlation between WGBS sample replicates for DNA methylation scores calculated in 500bp genomic windows. **D.** Bar plot showing global average CpG methylation levels in each sample group. **E.** Bar plot showing distribution of CpG methylation levels in each sample group.



II - Figure 4. 2: Identification of Dnmt3b-target DMRs in the transition from ESC to EpiSC. **A.** (left) WGBS heatmap showing DMRs arising during the transition from ESC to EpiSC in WT and 3BKO cells, clustered by K-means. Each cluster shows a distinct pattern of DNA methylation levels and dynamics. (right) Boxplot distributions of the DNA methylation levels for the differentially methylated CpG loci (DML) present in each DMR cluster (the total number in cluster is reported on top of each plot facet). **B.** Annotation of all identified DMRs to distinct genomic features. Pie chart shows the fraction of DMRs overlapping each feature. **C.** Annotation of DMRs to distinct genomic features, reported as (left) the percentage of DMRs overlapping each feature and (right) the log2-enrichment for each feature, calculated with the Genomic Association Test (GAT) software *(123)*. **D.** Heatmap showing selected gene ontology terms for enriched biological processes in each cluster. Gene set over-representation analysis was performed for genes associated with DMRs overlapping putative regulatory regions.

In contrast, cluster III is composed by a small set of regions (~ 6% of the total number of DMRs) that are demethylated moving from ESC to EpiSC, in both WT and 3BKO cells. Thus, these regions are likely related to the epiblast identity, which is not impaired upon *Dnmt3b*-loss (see Chapter 3).

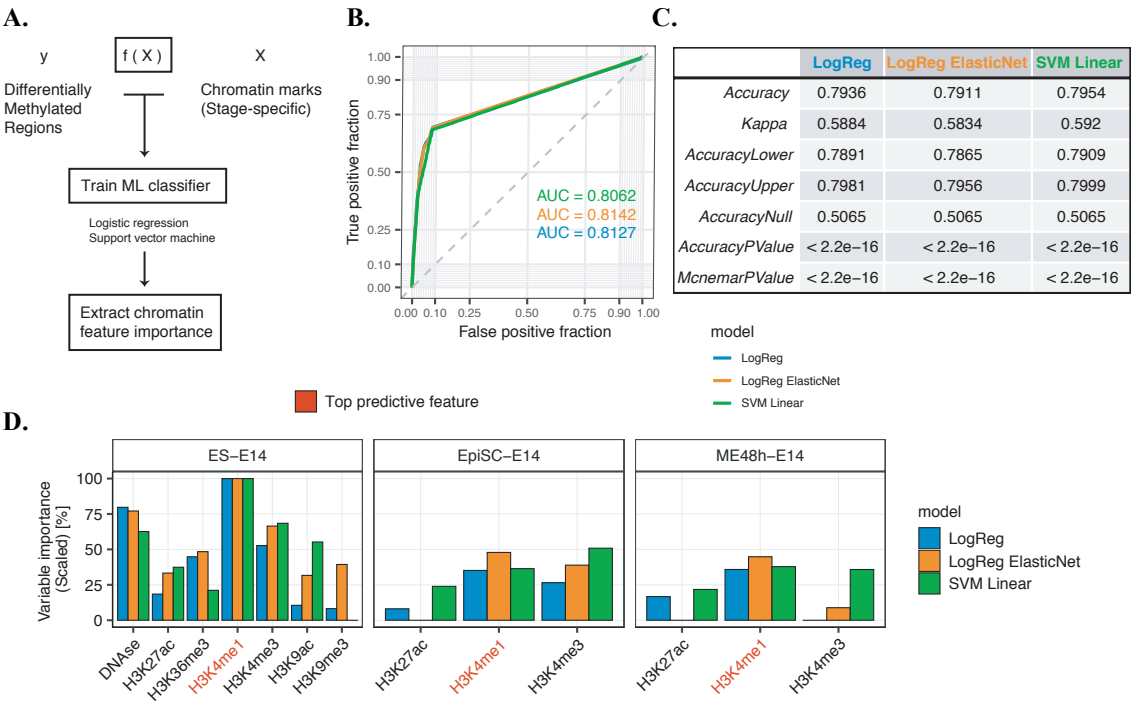
The genomic distribution of DMRs showed a significant over-representation at promoters (3.7 % of nucleotide overlap), CpG Islands (1.8 %) and exons (9.6 %) (Figure 4.2B, Q-value < $1e^{-4}$ as calculated by the GAT tool (123), as well as a high number of intronic (55.18 %) and intergenic (33.03 %) regions.



II - Figure 4. 3: Identification of differentiation-associated enhancers. **A**. Heatmaps showing H3K27ac ChIP-seq signals for WT cells over the differentiation time course, clustered in stage-specific or shared-by-stage regions, for typical (top) and super enhancer (bottom) regions. **B.** Ranking plots of H3K27ac ChIP-seq signals for WT cells at each time point of differentiation obtained from ROSE (124) (see Chapter 7, paragraph 7.2.3). Inflection points of the curves represent the cut-off for super enhancer definition. **C.** Bar plots representing the number of DMRs overlapping with either stage-specific or shared-by-stage (common) typical and super enhancer regions.

The elevated fraction of intergenic and intronic DMRs suggested a possible role for *Dnmt3b*-dependent *de novo* DNAme in regulating the activity of enhancer elements associated with differentiation. To assess this hypothesis, we took advantage of publicly available and newly generated ChIP-seq data for the histone marks H3K4me3, H3K4me1, and H3K27ac, to define a

complete set of putative regulatory elements arising over the differentiation time course (Figure 4.2C and Figure 4.3-4, see Chapter 7, paragraph 7.2.3). We found that 19,213 DMRs (30 % of total number of DMRs) overlapped with regulatory regions, either promoters or enhancers, that are active (H3K27ac) in any of the three stages of the differentiation (ESCs, EpiSCs or meso-endoderm) (Figure 4.2C and Figure 4.3C), with a significant enrichment at both typical (13.57 % of nucleotide overlap, log2[Obs./Exp.] ratio > 2 and Q-value < $1e^{-4}$) and super enhancer (14 % of nucleotide overlap, log2[Obs./Exp.] ratio > 3 and Q-value < $1e^{-4}$) elements (Figure 4.2B). Indeed, these were the two genomic features displaying the strongest changes overall in DNAme levels during the ESC to EpiSC transition (Figure 4.1F). Functional enrichment analysis of genes associated with each DMR cluster revealed a significant enrichment for gene ontology terms related to development and cell proliferation (i.e. *anatomical structure development, regulation of cell proliferation*), as well as for pathways associated with cell differentiation (i.e. *cellular response to LIF, WNT signaling pathways, cell differentiation*) (Figure 4.2D), in all the *Dnmt3b* target DMR clusters (i.e. cluster I, II, IV), thus confirming the crucial role played by *Dnmt3b*-dependent *de novo* DNAme in these early stages of embryonic development.



	LogReg	LogReg ElasticNet	SVM Linear
Accuracy	0.7936	0.7911	0.7954
Kappa	0.5884	0.5834	0.592
AccuracyLower	0.7891	0.7865	0.7909
AccuracyUpper	0.7981	0.7956	0.7999
AccuracyNull	0.5065	0.5065	0.5065
AccuracyPValue	< 2.2e−16	< 2.2e−16	< 2.2e−16
McnemarPValue	< 2.2e−16	< 2.2e−16	< 2.2e−16

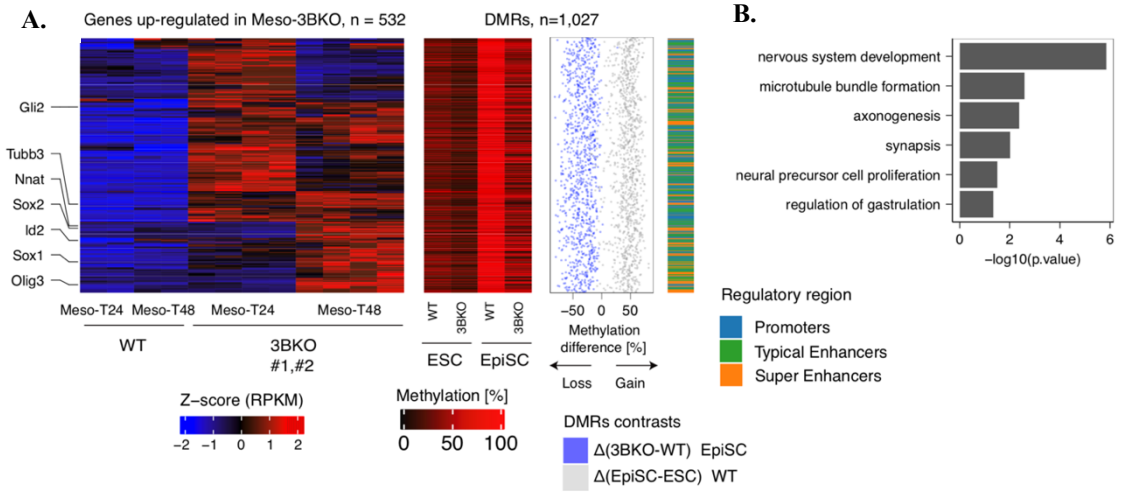
II - Figure 4. 4 Chromatin features predictive of DMRs occurrence. **A.** Schematic representation of the ML classification workflow employed to characterize chromatin features predictive of the occurrence of DMRs. **B.** Receiving Operating Characteristic (ROC) curves for the three trained ML models (LogReg=Logistic Regression; LogReg ElasticNet=Logistic Regression with Elastic Net regularization; SVM Linear=Support Vector Machine with Linear Kernel). AUC=Area Under the ROC Curve. **C.** Performance metrics for the three trained ML models, as calculated by the confusionMatrix function in the caret R packge (125). **D.** Chromatin feature importance (scaled in the range 0-100) for the prediction of DMRs occurrence according to the three trained ML models, reported for each differentiation stage (ESC, EpiSC, Meso-endoderm at 48h).

To further characterize the association between the identified DMRs and the chromatin context in which they emerge, we employed a machine learning (ML) classification approach, modeling the genomic occurrence of DMR as a function of various chromatin features obtained at different stages of differentiation (ESC, EpiSC and Meso-endoderm after 48h of differentiation, see Chapter 7 paragraph 7.2.5) (Figure 4.4A). Specifically, we used DNase hypersensitivity data mapping open chromatin regions and ChIP-seq data of histone modifications, both publicly available (for the ESC stage, retrieving ES-E14 data from ENCODE database) and in-house generated (H3K27ac, H3K4me1 and H3K4me3 for the EpiSC and Meso-endoderm differentiation). We trained three different ML classifiers (Logistic Regression, Elastic Net regularized Logistic Regression, Support Vector Machines with Linear Kernel), obtaining comparable predictive performances (computed using 10-fold cross validation, Figure 4.4B-C). The analysis of the most influential features predictive of DMRs occurrence revealed the H3K4me1 histone mark at the ESC stage as the most important variable overall for all the three classifiers (Figure 4.4D). Moreover, H3K4me1 was the most influential feature also at later stages of differentiation (Figure 4.4D). H3K4me1 defines both active and primed enhancer elements (126). Hence, this result confirms that the dynamic regulation of DNAme patterns at enhancers is of paramount importance at this developmental stage.

4.3 *Dnmt3b*-dependent DNAme acts as an epigenetic priming for lineage commitment

To identify the genes that are directly targeted by *Dnmt3b* during the ESC to EpiSC transition and, thus, responsible for the impaired meso-endodermal differentiation of 3BKO cells, we integrated the results from the joint multi-omics profiling of gene expression (RNA-seq), DNAme (WGBS) and histone modifications associated to active or primed regulatory regions (ChIP-seq of *H3K4me1-me3*, *H3K27ac*). We obtained a list of 532 Dnmt3b target genes, defined by (i) having at least one associated DMR directly targeted by Dnmt3b in EpiSC that overlaps a putative regulatory region and (ii) upregulation in gene expression upon meso-endodermal induction in 3BKO cells with respect to WT (Figure 4.5A). Remarkably, the identified list of targets showed significant enrichment for gene sets involved in neuro-ectodermal lineage commitment (Figure 4.5B), including key neuronal markers and TFs such as *Sox1*, *Sox2*, *Tubb3* and *Olig3* (94). This result suggests that the impairment of the DNAme landscape at the primed pluripotency stage is directly responsible for the neuro-ectodermal transcriptional program observed in 3BKO cells when committed to differentiation, and that Dnmt3b controls the correct formation of meso-endodermal lineages by silencing both promoters and enhancers of key neuro-ectodermal genes.
The neuro-ectodermal related genes need to be switched off in order to ensure the proper transcription of meso-endodermal genes at the time of lineage commitment. Therefore, at the primed pluripotency stage, the Dnmt3b-dependent DNAme acts as an epigenetic priming that ensures the proper establishment of lineage differentiation programs at the time of gastrulation.

II - Figure 4. 5: Dnmt3b-dependent DMRs overlapping putative regulatory regions. Genome browser view of WGBS data for four representative loci displaying Dnmt3b-dependent de novo methylation activity during the transition from ESC to EpiSC in enhancers (**A**) and promoter (**B**) elements. Typical and Super enhancer regions arising over the differentiation time course were defined using ROSE on H3K27ac signal (see Chapter 7, paragraph 7.2.3).



II - Figure 4. 6: Integrated analysis reveals key Dnmt3b-targeted transcription factors associated with neuro-ectodermal lineage commitment. **A.** Integrated RNA-seq, WGBS and regulatory regions heatmaps for Dnmt3b target genes. Dnmt3b target genes are defined as upregulated genes in 3BKO with respect to WT cells during meso-endodermal induction that have a Dnmt3b-target DMR within 100kb of their TSS overlapping a regulatory region (annotated promoters and/or enhancers defined by H3K27ac signal during the differentiation time course). **B.** Bar plots of top gene ontology terms for enriched biological processes in Dnmt3b target genes.

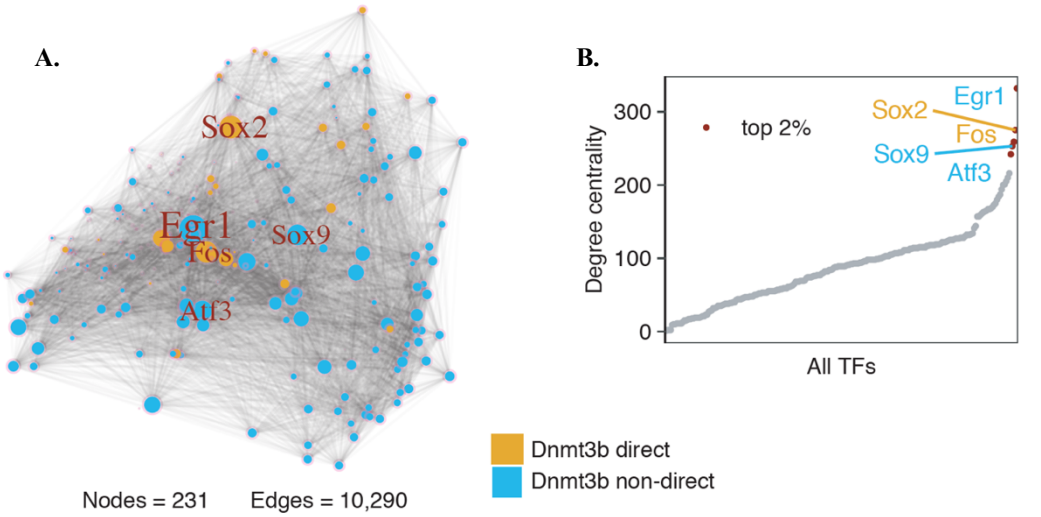
Chapter 5

Silencing the master regulator Sox2 rescues the impaired meso-endodermal induction of 3BKO cells

Results

5.1 Reconstruction of the Dnmt3b-dependent regulatory network

To further investigate the molecular targets of *Dnmt3b* that play a role in regulating the correct cell fate specification of EpiSC, we reconstructed the *Dnmt3b*-dependent network of transcription factors (TFs) that are dysregulated in meso-endodermal differentiation upon *Dnmt3b* loss. To do so, we exploited the results of our multi-omics integrated analysis (see Chapter 4) and combined them with known information about TFs targets publicly available from TRRUSTv2 (127) and ChEA3 (128) databases. (Figure 4.1A, see Chapter 7, paragraph 7.2.5). The network is composed by 231 nodes and 10,290 edges that summarize the regulatory interactions between TFs that are either (i) directly targeted by *Dnmt3b* or (ii) downstream regulated by *Dnmt3b* direct target TFs. Network's node prioritization by out-degree centrality (i.e. how many TFs are regulated by each TF) revealed the presence of *Sox2* - as direct target of *Dnmt3b* - among the top 2% central nodes (Figure 5.1B). *Sox2*, besides playing a role as a core pluripotency factor (95), is a well-known master regulator of neuronal differentiation (94), and results from previous studies showed its antagonistic activity with the mesodermal master regulator *T*/*Brachyury* to control neuronal-mesodermal lineage switching (89). For these reasons, it may be a crucial player that acts upstream of the regulatory cascade responsible for the observed neuro-ectodermal transcriptional phenotype in 3BKO cells.



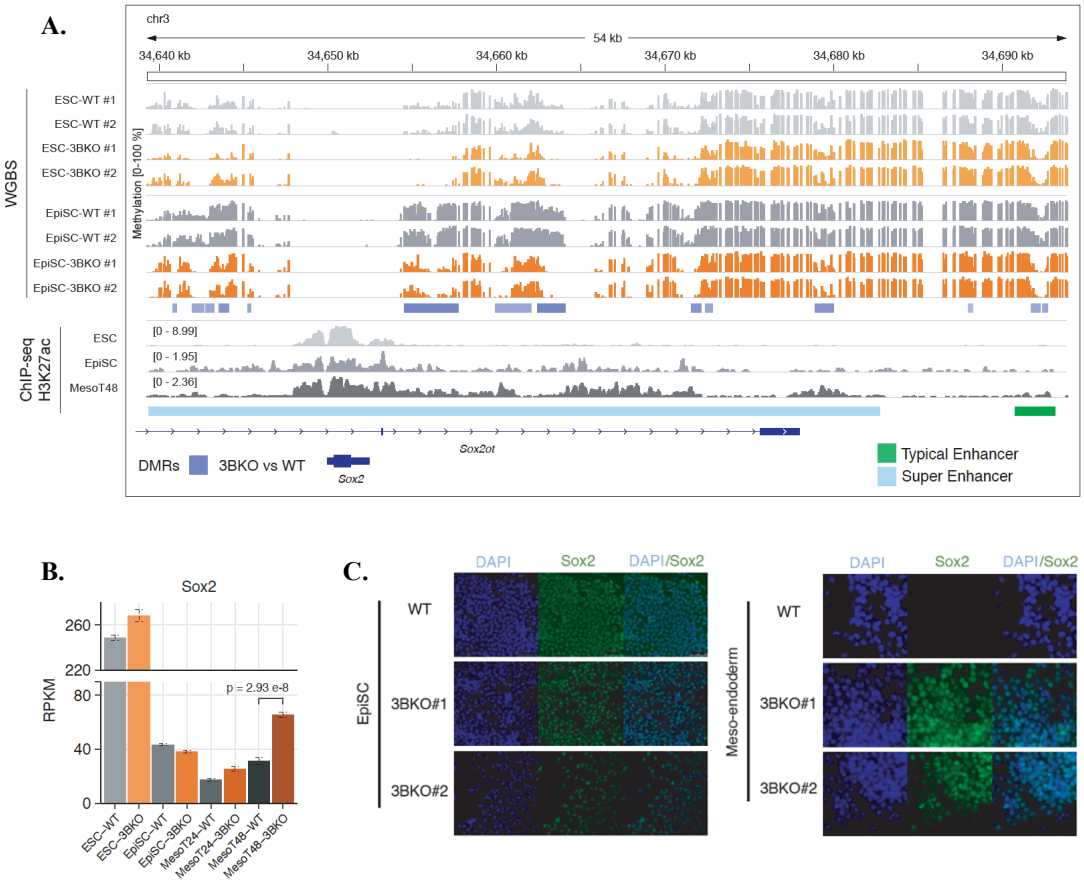
II - Figure 5. 1: Dnmt3b-dependent regulatory network. **A**. *Dnmt3b*-dependent regulatory network of transcription factors. Nodes are all the differentially expressed transcription factors between 3BKO and WT cells during meso-endodermal commitment. Yellow nodes are Dnmt3b direct target nodes, while blue nodes are indirect. Evidence of regulation between transcription factors was retrieved from TRRUSTv2 and ChEA3 databases. **B.** Node ranking on the basis of their out-degree centrality (i.e. number of target factors). Red dots represent the top 2%.

5.2 Silencing *Sox2* rescues the meso-endodermal phenotype in 3BKO cells

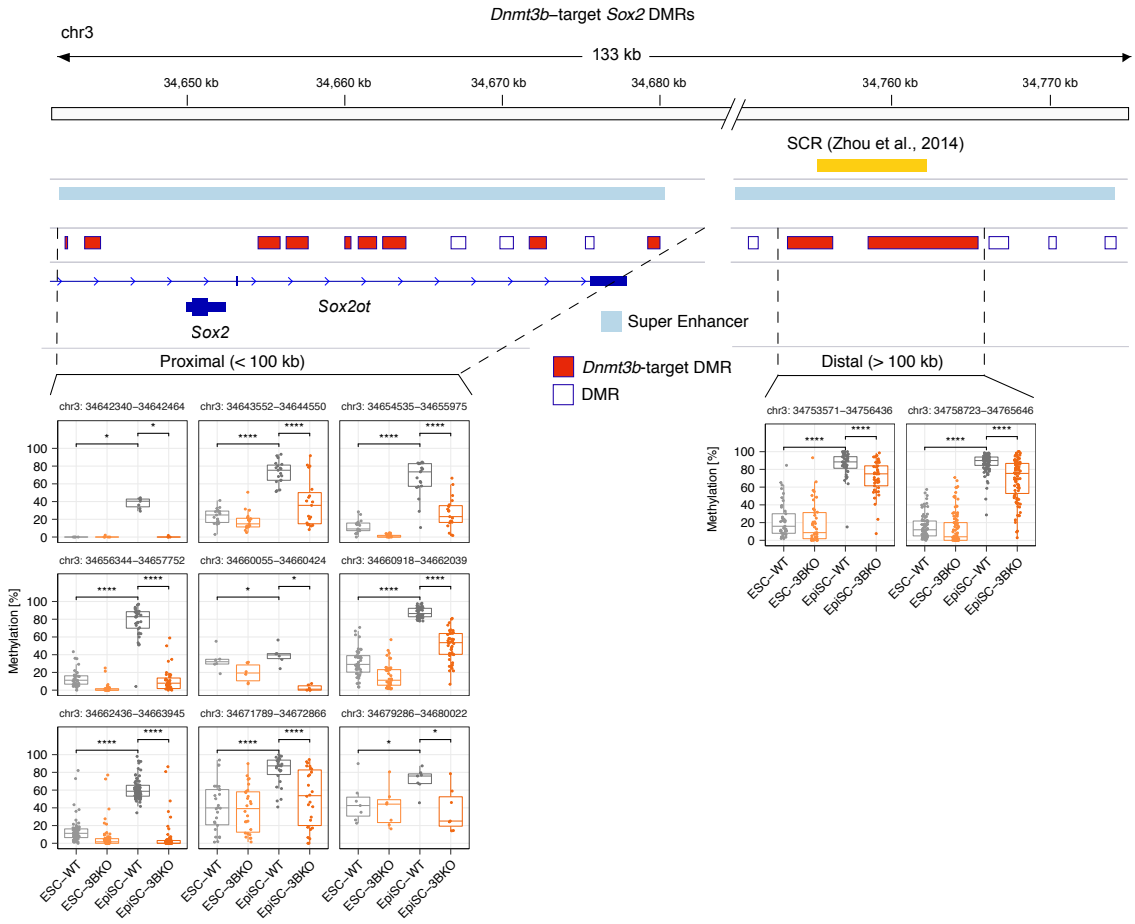
Once stated the evidence of *Sox2* as a direct *Dnmt3b* target upstream of the regulatory cascade that controls the aberrant neuro-ectodermal transcriptional program in 3BKO cells once committed to meso-endoderm (Figure 5A), and given its established role as a master regulator of neuronal differentiation, we asked whether its silencing could rescue the meso-endodermal phenotype in 3BKO cells. The *Sox2*-associated DMRs targeted by *Dnmt3b* are within 2 different enhancer loci located respectively at ~4 kb and ~12kb of *Sox2* TSS (Figure 5.2A, Figure 5.3), both of them overlapping the broad *Sox2* super-enhancer locus. Moreover, results from recent studies demonstrated the presence of a distal enhancer cluster, located at ~100kb of *Sox2* TSS, responsible for the regulation of *Sox2* expression specifically in mouse ESC (129). Hence, in addition to the proximal regions, we observed significant reduced DNAme levels in 3BKO cells at the EpiSC stage also at the distal *Sox2* control region (SCR), even if with a smaller effect size as compared to the DNAme loss observed in the distinct subunits of the proximal super-enhancer locus (Figure 5.3).

The hypo methylated state of the proximal regions in 3BKO cells at the EpiSC stage was confirmed by targeted bisulfite analysis, showing loss of methylation in majority of their CpG sites. The lack of *de novo* DNAme at these key *Sox2*-associated regulatory regions results in aberrant *Sox2* expression in 3BKO cells (Figure 5.2). Indeed, the upregulation of *Sox2* is observed at 48 hours of meso-endodermal induction, as shown by both RNA-seq and IF (Figure 5.2B-C).

To achieve depletion of *Sox2* expression during meso-endoderm induction, we took advantage of RNA interference techniques. At the EpiSC stage, we double-transfected 3BKO cells with 2 independent shRNA constructs against *Sox2* (shSox2-1, shSox2-2) and a scramble (sh-pLKO) shRNA for 6 hours, after which we induced meso-endoderm (Figure 5.3A, see Chapter 7 paragraph 7.1.8). Following 48h of differentiation, we observed a significant downregulation of *Sox2* expression for both *Sox2*- targeting constructs with respect to sh-pLKO at both the RNA (~ 50 % reduction observed by qRT-PCR) and the protein level by western blot analysis (in both 3BKO clones, Figure 5.3B, D).

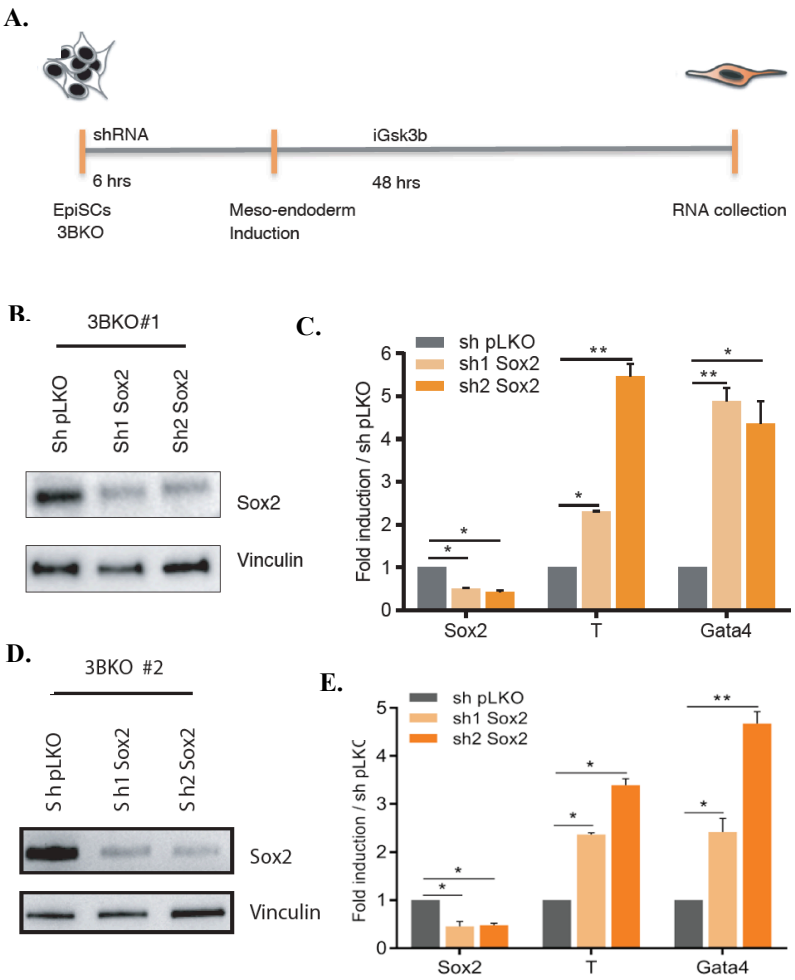


II - Figure 5. 2: Aberrant Sox2 methylation and expression in 3BKO cells. **A.** Genome browser view of the Sox2 locus showing (i) the DNA methylation profiles obtained by WGBS in WT and 3BKO cells at the ESC and EpiSC stage, and (ii) H3K27ac ChIP-seq signals in the WT differentiation time-course (ESC, EpiSC, Meso-endoderm at 48 hours), which define the Sox2 super-enhancers locus. Distinct Dnmt3b target DMRs are present in distinct subunits of the super-enhancer locus, which are de novo methylated in the ESC to EpiSC transition, but remain hypo-methylated in 3BKO cells. **B.** Gene expression levels of *Sox2* for WT and 3BKO over the differentiation time-course, obtained from RNA-seq and reported as normalized RPKM values. Significant up-regulation in 3BKO with respect to WT is observed at 48h of meso-endodermal induction (p-value from differential expression analysis, see Chapter 3). **C.** Representative ICC images of *Sox2* (green) protein expression in WT, 3BKO#1 and 3BKO#2 in EpiSCs (Top) and meso-endoderm (bottom) induced cells, nuclei stained by DAPI (blue). Scale bar 25 mm.



II - Figure 5. 3 Dnmt3b target Sox2 DMRs. Genome browser view of the DMRs associated to Sox2 during the ESC to EpiSC transition. Dnmt3b-target loci emerge in both the proximal (< 100 kb from TSS) super-enhancer region and distal (>100 kb from TSS) Sox2 control region (SCR), as defined by Zhou et. al *(129)*. Boxplots show distributions of methylation levels measured at individual CpG sites in each Sox2-related Dnmt3b-target DMR.

We next analysed by qRT-PCR the expression of selected meso-endoderm markers, observing a significant upregulation of *T* (mesoderm) and *Gata4* (endoderm) in *Sox2*-silenced 3BKO cells (Figure 5.3C, E), thus demonstrating a partial rescue of the meso-endodermal transcriptional program. Taken together, these results demonstrate the key role played by the down-modulation of *Sox2*, mediated by *Dnmt3b*-dependent methylation of its super-enhancer locus, for the proper establishment of the meso-endodermal fate at the time of lineage commitment.



II - Figure 5. 4: Silencing the master regulator Sox2 rescues the impaired meso-endodermal induction of 3BKO cells. **A.** Scheme of meso-endoderm rescue experiment by silencing *Sox2*. WB analysis of 3BKO#1 (**B**) and 3BKO#2 (**D**) cells silenced with two different shRNAs against *Sox2*. qRT-PCR analyses for ectoderm (Sox2), mesoderm (T) and endoderm (Gata4) genes in Sox2 silenced 3BKO#1 (**C**) and #2 (**E**) cells. Error bars represent standard deviation between three biological replicates (* = p-value < 0.05, ** = p-value < 0.01, *t*-test).

Chapter 6

Discussion II

6.1 The specificity of DNMT3B in early embryonic development

The establishment of *de novo* DNAme patterns at the exit from pluripotency is crucial for cell fate specification during early mammalian development (112,113). In this work, we dissected the specific contribution of the *de novo* DNMTs - DNMT3A and DNMT3B - in controlling lineage fate decisions during mouse early embryonic development, showing a predominant role for DNMT3B at this early developmental stage. The single cell transcriptomic analysis of the differentiation trajectories in EBs, a 3-dimensional unbiased differentiation model that recapitulates *in vitro* the lineage specification toward progenitors of the three primary germ layers, showed that the commitment toward the meso-endodermal fate is specifically impaired in 3BKO cells, whereas 3AKO knockout cells could differentiate in both ectoderm and meso-endoderm cells (Chapter 2). Our results are concordant with previously published *in vivo* studies showing that mice lacking either *Dnmt3a* or *Dnmt3b* manifest distinct defects and die at different stages of development (112,113). In particular, they showed that *Dnmt3b* mutant mice manifest early embryonic lethality, while *Dnmt3a* knock-out mice die after birth (112). Moreover, a more recent study has shown that, in humans, DNMT3A-dependent de novo DNAme is essential for the correct formation of motor neuron (111), which are terminally differentiated somatic cells formed later on during development and derived from ectoderm. Therefore, in agreement with our results, these data indicate that *Dnmt3b*-dependent *de novo* DNAme plays a more important role in the regulation of early embryonic development, while *Dnmt3a* targets are crucial for late development or after birth.

6.2 Target genomic loci of DNMT3B at the primed pluripotency stage

To further characterize the functional role of DNMT3B in resolving early lineage commitment, we induced the formation of *in vitro* EpiSCs, which resemble the primed pluripotent state of the early post-implantation embryo, and further differentiated them toward the meso-endodermal lineage. Remarkably, the lack of *Dnmt3b* does not affect the formation of EpiSCs, but prevents their subsequent commitment to meso-endodermal progenitors (Chapter 3). By WGBS profiling,

we demonstrated that *Dnmt3b* loss results in failures in the acquisition of *de novo* DNAme in a huge fraction of DMR established during the ESCs to EpiSCs transition, mostly targeting regulatory regions associated with key developmental genes (Chapter 4). The results from the multi-modal integrated analysis of WGBS, ChIP-seq of histone marks and RNA-seq, revealed that the activity of *Dnmt3b* in epiblast is responsible for the silencing of a number of regulatory regions associated with neuro-ectodermal marker genes, which need to be switched off to ensure proper differentiation into meso-endoderm cells, in agreement with recent findings obtained *in vivo* stating the ectodermal path as the default route of lineage differentiation (130). Therefore, the Dnmt3b-dependent de novo methylation participates in priming the epiblast to ensure flawless commitment at later stages of embryonic development.

6.3 DNMT3B ensures meso-endodermal specification by regulating *Sox2*

The reconstruction of the *Dnmt3b*-dependent network of TFs during this process highlighted the presence of *Sox2* as a direct target of Dnmt3b in epiblast stage (Chapter 5). Sox2 is a well-known master regulator of neuronal differentiation, which competes with the mesoderm marker *T*/*Brachyury* for neural/mesodermal lineage switching (89). Indeed, we observed that 3BKO EpiSCs show a significant reduction of methylation at their super-enhancer region, which should be decommissioned in the cells primed to differentiate into meso-endoderm (129,131). In agreement with these observations, we found that *Sox2* silencing in 3BKO at the EpiSC stage restores the expression of the meso-endoderm markers upon their further differentiation, thus rescuing the impaired meso-endodermal transcriptional program. In conclusion, our data provide a functional characterization of DNMT3B and its role in the regulation of cell differentiation during early stages of mouse embryonic development, demonstrating that the specific activity of DNMT3B is necessary to prime EpiSCs for their further differentiation into the meso-endodermal lineages, thus establishing the epigenetic inheritance needed for the selective activation of specific transcriptional programs at the time of lineage commitment.

Chapter 7

Materials and methods

7.1 Experimental procedures

7.1.1 Cell culture

Embryonic stem cells were generated and cultured as described previously (106). E14 mouse WT, DNMT3B $^{-/-}$ (cl.B77) and DNMT3B $^{-/-}$ (cl.B126) ES cells were cultivated in high-glucose DMEM (Euroclone) supplemented with 15% FBS (Millipore Corp., Billerica, MA, USA), 0.1 mmol/l nonessential amino acids (Invitrogen), 1 mmol/l sodium pyruvate (Invitrogen), 0.1 mmol/l β-mercaptoethanol, 1500 U/ml Leukemia Inhibitory Factor (LIF; Millipore), 25 U/ml penicillin, and 25 μg/ml streptomycin.

7.1.2 Dnmt3a -/- and Dnmt3b -/- generation

Dnmt3a$^{-/-}$ and Dnmt3b$^{-/-}$ ESCs generation were performed taking advantage from TALEN and CRISP-R/Cas9 technologies, respectively, as described in (106) and (132)

7.1.3 Embryoid body formation

To induce formation of EBs, ES cells were transferred using trypsin to low-attachment plates (CORNING) in Alpha-MEM (BE02-002F LONZA) supplemented with 10% KOSR (10828-028 GIBCO), 5% FBS (Millipore), 1% nonessential amino acids (Invitrogen), 1% sodium pyruvate (Invitrogen), 0.1 mmol/l β-mercaptoethanol, 25 U/ml penicillin and 25 μg/ml streptomycin. Medium was changed every 3 days.

7.1.4 EpiSCs induction from ESCs

Epiblast (EpiSCs) induction was modified from (88). Briefly, a single-cell suspension was seeded onto Geltrex (A1413202 GIBCO)-coated plates at a density of 10,000 cells cm-2 in N2B27

medium supplemented with 20ng/ml ActivinA (PHC9564 GIBCO) and 12 ng/ml bFGF (PHG0026 GIBCO). The cells were passaged 1:3 as small clumps using Collagenase IV (17104019 GIBCO). EpiSCs were collected for DNA and RNA analyses after 14 days of induction followed by daily medium changes.

N2B27 medium is composed by 50% advanced DMEM/F12 (12634028 GIBCO) and 50% Neurobasal medium (21103049 GIBCO), supplemented with 0.5% N2 Supplement (17502048 GIBCO), 1% B27 Supplement (17504044 GIBCO), 0.033% BSA solution (A9647 SIGMA), 50 uM β-mercaptoethanol (M3148 Sigma), 2mM Glutamax (35050038 GIBCO), 100U/ml penicillin and 100 ug/ml streptomycin (DE17-602E LONZA).

7.1.5 EpiSCs differentiation towards meso-endoderm (ME) lineage

For ME lineage specific differentiation, EpiSCs were plated as small clumps onto Geltrex-coated plates in EpiSCs medium for 24 hours. The day after, medium was replaced with N2NB27 medium consisted of 50% advanced DMEM/F12 (12634028 GIBCO) and 50% Neurobasal medium (21103049 GIBCO), supplemented with 0.5% N2 Supplement (17502048 GIBCO), 1% B27 supplement minus Vitamin A (12587010 GIBCO) and 3 uM iGSK3β (CHIR99021 SIGMA). Cells were fed daily until the end of differentiation.

7.1.6 FACS analysis

After 48 hours of differentiation toward ME lineage cells were dissociated. 1*10^6 cells were analysed for Annexin V-kit (Miltenyi Biotec) following the manufacturer's instruction. Shortly cells were wash twice in the Annexin Buffer and then incubated with Annexin V-FITC for 15' in the dark at room temperature. After that cells were washed once in Annexin Buffer and, immediately before FACS acquisition, PI was added.

7.1.7 Protein extraction and Western blotting

For total cell extracts, cells were resuspended in F-buffer (10mM TRIS-HCl pH 7.0, 50mM NaCl, 30mM Na-pyrophosphate, 50mM NaF, 1% Triton X-100, anti-proteases) and sonicated for 3 pulses. Extracts were quantified using bicinchoninic acid (BCA) assay (BCA protein assay kit; catalog no. 23225; Pierce) and were run on SDS-polyacrylamide gels at different percentages, transferred to nitrocellulose membranes and incubated with specific primary antibodies overnight.

7.1.8 shRNA Constructs

Custom shRNAs against *Sox2* were constructed using the TRC hairpin design tool (http://www.broadinstitute.org/rnai/public/seq/search), and designed to target the following sequences:

- 5′-ACCAATCCCATCCAAATTAAC-3′ (shRNA1)
- 5′-GCACAGTTTGAGATAAATAAA-3′ (shRNA2)

Hairpins were cloned into pLKO.1 vector (Addgene: 10878) and each construct was verified by sequencing.

7.1.9 Transfections

Transfections of mouse EpiSCs were performed using Lipofectamine 2000 Transfection Reagent (INVITROGEN) in accordance with the manufacturer's protocol using equal amounts of each plasmid in multiple transfections. For Sox2 knockdown, cells were transfected twice with 5 μg of the specific shRNA construct, and maintained in growth medium for 48 h.

7.1.10 Alkaline phosphatase (AP) staining and Immunostaining

ES cells and EpiSCs were fixed with 4% paraformaldehyde for 2 min and then stained with Vector® Red alkaline phosphatase substrate kit (SK-5100) according to the manufacturer's protocol. For immunostaining, cells were fixed with 4% paraformaldehyde for 10 min at room temperature. Permeabilization was performed in 0.1% Triton X-100 in PBS for 15 min, and then the cells were blocked in 2% BSA in PBS at room temperature for 2 hours. Cells were stained with primary antibodies at 4 °C overnight. Secondary antibody was applied for 1 hour at room temperature. Nuclei were stained with DAPI (D21490 INVITROGEN). Images were acquired using a Leica TCS SP5 Confocal microscope and LAS AF Lite software.

7.1.11 Antibodies

The following antibodies were used for western blotting: Dnmt3a (sc-365769, Santa Cruz), Dnmt3b (ab122932, Abcam), Dnmt3l (provided by Dr. S. Yamanaka, Kyoto University, Japan), Lefty1 (ab22569, Abcam), Otx2 (ab21990, Abcam), Sox2 (sc-365823, Santa Cruz), Prdm14 (MAB6175, R&D systems), Vinculin (SAB4200080, Sigma), β-Actin (A5441, Sigma). The following antibodies were used for immunostaining: Dnmt3b (ab122932, Abcam), Lefty1 (ab22569, Abcam), Sox2 (sc-365823, Santa Cruz), Flk1 (sc-6251, Santa Cruz).

7.1.12 Immunoprecipitation (IP)

Nuclear proteins from about 10 * 10^6 cells were incubated with 3 ¥ g of specific antibody overnight at 4° C. Immunocomplexes were incubated with protein-G- conjugated magnetic beads (DYNAL, Invitrogen) for 2 hours at 4° C. Samples were washed four times with digestion buffer supplemented with 0.1% NP-40 at RT. Proteins were eluted by incubating with 0.4M NaCl TE buffer for 30 min and were analyzed by western blotting.

7.1.13 Chromatin Immunoprecipitation (ChIP) assay

For ChIP experiments, approximately $2*10^7$ cells were cross-linked by addition of formaldehyde to 1% for 10 min at RT, quenched with 0.125 M glycine for 5 min at RT, and then washed twice in cold PBS. The cells were resuspended in Lysis Buffer 1 (50 mM Hepes-KOH pH 7.5, 140mM NaCl, 1mM EDTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100 and protease inhibitor) to disrupt the cell membrane and in Lysis Buffer 2 (10 mM Tris-HCl pH8.0, 200 mM NaCl, 1mM EDTA, 0.5 mM EGTA and protease inhibitor) to isolate nuclei. The isolated nuclei were then resuspended in SDS ChIP Buffer (20 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS and protease inhibitors). Extracts were sonicated using the BioruptorH Twin (Diagenode) for 2 runs of 10 cycles [30 sec ‘‘ON’’, 30 sec ‘‘OFF’’] at high power setting. Cell lysate was centrifuged at 12,000 g for 10 min at 4°C. The supernatant was diluted with ChIP Dilution Buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton) before immunoprecipitation step. Streptavidin beads (Dynabeads®Protein G, Life Technologies) were saturated with PBS/1% BSA and the samples were incubated with 2 ug of antibody overnight at 4°C on a rotator. Next day samples were incubated with saturated beads for two hours at 4°C on a rotator. Successively immunoprecipitated complexes were washed five times with RIPA buffer (50 mM Hepes-KOH pH7.6, 500mM LiCl, 1mM EDTA, 1% NP-40, 0,7% Na-Deoxycholate) at 4°C for 5 minutes each on a rotator. Elution Buffer was added and incubated at 65°C for 15 minutes. The de-crosslinking was performed at 65°C overnight. De-crosslinked DNA was purified using QIAQuick PCR Purification Kit (QIAGEN) according to the manufacturer’s instruction.

7.1.14 DNA extraction.

Genomic DNA was extracted from cells using the DNeasy Blood and Tissue kit (QIAGEN, 69506) following the manufacturer’s instructions.

7.1.15 DNA methylation analysis.

For DNA methylation analysis, 1 µg of genomic DNA was used for bisulfite conversion by using the EpiTect Conversion Kit (QIAGEN, 59104) according to the manufacturer's protocol. Converted DNA was eluted in 20 µl and 3 µl of converted DNA was used in a 50 µL PCR reaction (INVITROGEN, 12346). PCR products were purified and cloned into TOPO-TA vector (Invitrogen 450030), and positive clones were verified by sequencing. The bisulfite sequencing analysis of CpG methylation was performed using QUMA (http://quma.cdb.riken.jp/), as described in (133).

7.1.16 RNA qRT-PCR analysis

RNA was extracted using TRIzol reagent (Invitrogen), according to the manufacturer's protocol. Real-time PCR was performed using the SuperScript III Platinum One-Step Quantitative RT-PCR System (Invitrogen, cat.11732-020) following the manufacturer's instructions.

7.1.17 RNA-seq library preparation

Total RNA was isolated using TRIzol reagent (Invitrogen), according to the manufacturer's protocol. Quantity and quality of the starting RNA were checked by Qubit and Bioanalyzer (Agilent). 1 μg of total RNA was subjected to poly(A) selection, and libraries were prepared using the TruSeq RNA Sample Prep Kit (Illumina) following the manufacturer's instructions. Libraries were sequenced on Illumina NextSeq 500 System (single-end 75 bp reads)..

7.1.18 ChIP-seq library preparation

Starting from 10 ng of ChIP eluted sample, the library was produced for the genome wide analysis following the NEBNext® ChIP-Seq Library Prep Reagent Set for Illumina® (E 6240L NEB) manufacturer's instructions. Libraries were sequenced on Illumina NextSeq 500 System (single-end 75 bp reads).

7.1.19 Whole genome bisulfite-seq library preparation

The whole-genome Bisulfite-seq library was prepared starting from 5 μg of sonicated genomic DNA for 2 runs of 10 cycles [30 sec ''ON'', 30 sec ''OFF'']. In order to obtain 200 nt fragments. Sonicated DNA was then end-repaired twice, dA-tailed, and ligated to adapters, using the Illumina TruSeq DNA Sample Prep Kit, following manufacturer's instructions. Adapter-ligated DNA was loaded on an EGel Size select 2% agarose pre-cast gel (Invitrogen), and a fraction corresponding to fragments ranging from 25 0bp to 350bp was recovered. Purified DNA was then subjected to bisulfite conversion using the EpiTect Bisulfite Kit (QIAGEN) following the manufacturer's instruction except for the double desulfonation after the conversion. Bisulfite-converted DNA was finally enriched by 15 cycles of PCR using Pfu Turbo Cx. HotStart Taq (Agilent). Libraries were sequenced on Illumina Novaseq 6000 System, generating an average of 500 million 100bp paired-end reads and an average coverage depth of 30x per sample.

7.1.20 Single cell RNA-seq library preparation and sequencing

Full length single cell RNA-seq was performed using a modified version of the Smart-seq2 protocol (134). Briefly, individual cells are sorted into 96 well plates containing lysis buffer in presence of RNase inhibitor, dNTPs and OligodT. Reverse transcription of the polyadenylated

RNA will be performed with SuperScriptII and Template Switching Oligos. The resulting cDNA will be amplified with 25 cycles of PCR and libraries will be prepared for sequencing with standard NexteraXT Illumina protocol. Libraries were sequenced on Illumina NextSeq 500 System (single-end 75bp reads), reaching a median of ~ 578,000 generated reads per cell.

7.2 Bioinformatics data analysis

7.2.1 Single cell RNA-seq data analysis

Following quality controls (performed with FastQC v0.11.2), sequencing reads were processed with Trim Galore! v0.5.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) to perform quality and adapter trimming (parameters: *--stringency 3 –q 20*). Trimmed reads were next aligned to the mouse reference genome (UCSC mm9/NCBIm37) using HiSat2 v2.2.1 (135) with options: *-N 1 -L 20 -i S,1,0.5 -D 25 -R 5 --pen-noncansplice 20 --mp 1,0 --sp 3,0* and providing a list of known splice sites extracted from GENCODE (release M1 – NCBIM37 https://www.gencodegenes.org/mouse/release_M1.html and *hisat2_extract_splice_sites.py* script). Gene expression levels were quantified with featureCounts v1.6.1(136) (options: *-t exon -g gene_name*) using the GENCODE release M1 annotation.

The following criteria were applied to exclude low-quality cells from subsequent analyses:

- less than 100,000 assigned reads;
- less than 2,000 detected genes;
- more than 25% of reads assigned to mitochondrial genes

resulting in a total of 465 high-quality cells.

Gene expression counts were next analysed using Monocle2 (119). Only protein-coding genes were considered for downstream analysis. Read counts were size factors adjusted (*estimateSizeFactors* function), log-transformed and batch corrected (parameter *residualModelFormulaStr="~batch"* in the relevant Monocle2 functions). Variance modelling for feature selection was carried out using the scran R package (*trendVar* function, parameters*: parametric=T, method="loess"* and *decomposeVar* function). Dimensionality reduction (PCA + t-SNE) was next performed with the *reduceDimension* Monocle2 function, using the top 2,000 variable genes, the top 25 principal components and setting the perplexity equal to 20 (parameters: *max_components=2, reduction_method="tSNE", norm_method="log", num_dim=25, perplexity=20, residualModelFormulaStr="~batch"*).

Cell clustering was performed with the Louvain method implemented in the *clusterCells* Monocle2 function (parameters*: method="louvain"*). Cluster marker genes were identified by differential expression analysis, looking for upregulated genes in each cluster against the remaining cells with the *differentialGeneTest* function ($logFC \geq 0.5$ and $FDR \leq 0.01$). Gene set over-representation

analysis was performed for the top markers (logFC>0.8) list of each cluster with the gProfileR package (*gprofiler* function, parameters: *max_set_size=750*).

Pseudotime analysis was carried out using the DDRTree (119) algorithm, using the top 2,500 variable genes (*reduceDimension* function, parameters: *max_components=2*, *reduction_method="DDRTree",norm_method="log",residualModelFormulaStr="~batch"* and *orderCells* function). Branch-dependent analysis of gene expression was performed with the BEAM (120) methodology (*BEAM* function, parameters: *branch_point = 1, branch_states = c("B","C")*. The smoothed relative expression in each branch for the significantly branch-dependent genes ($qval \leq 0.001$) was hierarchically clustered (n=4) and visualized used the *plot_genes_branched_heatmap* function (parameters: *num_clusters = 4, branch_states = c("B","C")*) and as line plots using custom functions. Gene set over-representation analysis for each BEAM cluster was performed with the clusterProfiler R package.

7.2.2 RNA-seq data analysis

Following quality controls (performed with FastQC v0.11.2), sequencing reads were aligned to the mouse reference genome (mm10/GRCm38 Ensembl release 84) using HiSat2 v2.2.1 (135) with options: *-N 1 -L 20 -i S,1,0.5 -D 25 -R 5 --pen-noncansplice 20 --mp 1,0 --sp 3,0*. Pre-built indexes based on the Ensembl transcript annotation (release 84) for guided alignment to transcriptome were retrieved from the HiSat2 web site (https://cloud.biohpc.swmed.edu/index.php/s/grcm38_tran/download). Gene expression levels were quantified with featureCounts v1.6.1 (136) (options: *-t exon -g gene_name*) using the Ensembl release 84 transcript annotation (ftp://ftp.ensembl.org/pub/release-84/gtf/mus_musculus/Mus_musculus.GRCm38.84.gtf.gz). Multi-mapped reads were excluded from quantification.

Gene expression counts were next analysed using the edgeR (31) R package. Lowly expressed/detected genes (i.e. 1 RPKM in less than 2 samples) were filtered out, obtaining a total of 16,755 expressed genes for downstream analysis. Normalization factors were calculated using the trimmed mean of M-values (TMM) method (implemented in the *calcNormFactors* function) and RPKM were computed using normalized library sizes and gene lengths from the Ensembl release 84 annotation (*rpkm* function). Principal Component Analysis (PCA) was performed using the *prcomp* R function (parameters: *scale.=TRUE, center=TRUE*), using the top 2500 variable genes. Following dispersion estimation (*estimateDisp* function, *robust=TRUE*), an ANOVA-like test was implemented by fitting a Generalized Linear Model (GLM) to all sample groups (*glmFit* function) and performing Quasi-Likelihood F-test (*glmQLFTest* function) in order to identify the genes that were significantly varying during the differentiation time course (i.e. differentially expressed genes in any of the sample groups during the time course, using the ESC-WT condition as baseline in the design matrix formula). The resulting 4,624 genes ($|logFC| \geq 1$ and $FDR \leq$

0.001) were used for clustering of gene expression profiles with K-means (*kmeans* R function, parameters: *centers=4, iter.max=25, nstart=100*) followed by hierarchical clustering (parameters: *method="single", distance="euclidean"*). RPKM values were scaled as Z-scores across samples before computing distances. The optimal number of K-means clusters (n=4) was estimated using the within-cluster sum of squares methodology (*fviz_nbclust* function from the factoextra R package). Gene expression heatmaps were generated using the ComplexHeatmap R package. Gene set over-representation analysis was performed for each cluster with the gProfileR package (*gprofiler* function, parameters: *max_set_size=750*), using all the expressed genes as background. Differentially expressed genes between WT and 3BKO cells at each time point were obtained from the same GLM, comparing each contrast with the Quasi-Likelihood F-test ($|logFC| \geq 1$ and $FDR \leq 0.05$).

7.2.3 ChIP-seq data analysis

Following quality controls (performed with FastQC v0.11.2), sequencing reads were aligned to mouse reference genome (mm10/GRCm38) using Bowtie v2.3.4.1(137) (options: *-q --local*). Duplicated alignments (identified by Picard MarkDuplicates, https://broadinstitute.github.io/picard) and low-quality alignments/multi-mapping reads were excluded using SAMtools (138) (command: *samtools view –F1804 –q 30*). Coverage tracks were generated from filtered alignments using the deepTools (139) *bamCoverage* utility. IP and corresponding control (Input DNA) datasets were treated identically. Peak calling was performed using MACS v2.1.1 (140) (options: *callpeak -t=<IP> -c=<Input> -g mm --nomodel –extsize=<ES> --broad –q 0.05 --broad-cutoff 0.05 --fe-cutoff 1*). The read extension size (*ES*) was estimated by cross-correlation using the *phantompeakqualtools* package. Input-normalized ChIP-seq signals were obtained using the deepTools (139) *bamCompare* utility (options: *--extendReads=<ES> --scaleFactorsMethod readCount --binSize 10 --operation log2*). These processing steps were applied to all sample groups.

Identification of typical and super enhancer regions arising over the differentiation time course was performed with ROSE (141) (https://bitbucket.org/young_computation/rose.git), using H3K27ac signals and their Input DNA as control (*ROSE_main.py* script, parameters: *-g MM10 –t 2500*). Common and time-point specific differentiation enhancers were obtained using the *mergePeaks* utility from the HOMER suite (http://homer.ucsd.edu/homer). Signal profiles over peaks/genomic regions were obtained using the deepTools (139) *computeMatrix* utility and visualized using the *plotHeatmap* utility and/or custom R scripts.

7.2.4 WGBS data analysis

Following quality controls, sequencing reads were processed with Trim Galore! v0.5.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) to perform quality and adapter trimming (parameters: *--stringency 3 –q 20 --paired*). Trimmed reads were next aligned to the mouse reference genome (UCSC mm10/GRCm38) using Bismark v0.22.3 (142). The bisulfite-converted genome was created using the *bismark_genome_preparation* utility (parameters: *--genomic_composition --bowtie2*). Read mapping was performed with the *bismark* command (parameters: *--nucleotide_coverage*). Duplicated alignments were removed with the *deduplicate_bismark* utility and methylation calling was carried out using the *bismark_methylation_extractor* utility (parameters: *--ignore 1 --bedGraph --counts --gzip*). Genome-wide cytosine methylation reports with the top and bottom strand methylation evidence pooled into a single CpG dinucleotide entity were obtained using the *coverage2cytosine* utility (parameters: --zero_based --gzip --merge_CpG).

DMRs were identified using the DSS (143) R package, performing all pairwise comparisons between ESC and EpiSC for both WT and 3BKO samples, and between WT and 3BKO at matching time points. For each comparison, the *DMLtest* function was first run (parameters: *equal.disp=FALSE, smoothing=TRUE, smoothing.span=500*); next, differentially methylated loci were identified with the *callDML* function (parameters: *delta=0.1, p.threshold=0.001*); finally, DMRs were called using the *callDMR* function (parameters: *delta=0.2, p.threshold =0.05, minCG=5, dis.merge=100*). The resulting list of DMRs was combined into one DMR set, collapsing overlapping regions into a single DMR using BEDTools (144)

For further analysis, only CpG sites with coverage $\geq 5x$ in all samples were retained, and average DNA methylation levels for each DMR was calculated using the methylKit package (*regionCounts* and *percMethylation* functions). DMRs clustering was next performed using K-means, scaling region methylation scores as Z-scores before computing distances. The optimal number of K-means clusters (n=4) was estimated using the Gap Statistics methodology (*fviz_nbclust* function from the factoextra R package). DMRs annotation to genomic features and closest TSS was carried out using the *annotatePeaks.pl* script from the HOMER suite (parameters: *mm10 -annStats*). Gene set over-representation analysis was performed for DMR overlapping putative regulatory regions (defined by their H3K27ac signature) using the gProfileR package (*gprofiler* function, parameters: *max_set_size=750*).

7.2.5 ML classification analysis

To analyze the chromatin features predictive of DMRs, the genomic occurrence of DMRs was modelled as a function of various chromatin features. Specifically, the features in each DMR were considered to have a binary state defined by overlapping peaks in that region (0=no peak overlap; 1=peak overlap). As negative set, a random set of genomic regions was considered, matching the

size of the DMRs set (n=63,651) and excluding gaps and the DMRs set itself. Three ML algorithms were trained (Logistic Regression, Logistic Regression with Elastic Net regularization, Support Vector Machine with Linear Kernel) using the caret R package (125). To control for unbalanced sets, a down-sampling procedure was applied during training. Models performance was assessed using 10-fold cross validation. For Support Vector Machines, hyper-parameters tuning was performed, reporting the set of parameters with highest performance.

7.2.6 Integrated analysis

The Dnmt3b-target genes were defined by associating the Dnmt3b-target DMRs (within 100kb from the closest TSS and overlapping a putative regulatory region) to differentially expressed genes arising between 3BKO and WT samples during Meso-endoderm differentiation. The Dnmt3b-dependent transcription factors network was built integrating target-binding evidences from TRRUSTv2 (127) (https://www.grnpedia.org/trrust) and ChEA3 (128) (https://maayanlab.cloud/chea3, sets ARCHS4_Coexpression.gmt, ENCODE_ChIP-seq.gmt, Literature_ChIP-seq.gmt, ReMap_ChIP-seq.gmt, Enrichr_Queries.gmt, GTEx_Coexpression.gmt) and all the differentially expressed transcription factors arising between 3BKO and WT samples during Meso-endoderm differentiation, classifying nodes as direct or indirect Dnmt3b-target. Network metrics calculation and visualization was performed with the igraph R package (https://cran.r-project.org/web/packages/igraph/) (*degree* and *plot.igraph* functions).

7.2.7 Data availability

H3K27ac ChIP-seq for ESC were obtained from ENCODE (ENCFF001KFR, ENCFF001KFX, ENCFF001MXK, ENCFF001MXM, ENCFF001NNN, ENCFF001NNP). The datasets reported in this study are available in the Gene Expression Omnibus database with accession code GSEXXX.

Bibliography

1. Crick F. Central dogma of molecular biology. Nature. Nature Publishing Group; 1970 Aug 8;227(5258):561–3.

2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. Nature Publishing Group; 2009 Jan 1;10(1):57–63.

3. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. Nature Publishing Group; 2019 Nov;20(11):631–56.

4. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. Gibson G, editor. PLOS Genetics. Public Library of Science; 2007 Sep 28;3(9):e161.

5. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Jan 20;43(7):e47–7.

6. Anders S, Huber W. Differential expression analysis for sequence count data. Nat Prec. Nature Publishing Group; 2010 Apr 30;:1–1.

7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. National Academy of Sciences; 2005 Oct 25;102(43):15545–50.

8. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. Nat Rev Genet. Nature Publishing Group; 2016 Apr 12;17(6):353–64.

9. Tarca AL, Bhatti G, Romero R. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. Chen L, editor. PLOS ONE. Public Library of Science; 2013 Nov 15;8(11):e79217.

10. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007 Apr 15;23(8):980–7.

11. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2014 Jul;15(4):504–18.

12. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. BioMed Central; 2013 Dec 1;14(1):1–15.

13. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. BioMed Central; 2013 Dec 1;14(1):1–15.

14. Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Molecular Systems Biology. John Wiley & Sons, Ltd; 2011 Jan 1;7(1):497.

15. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England). 2003 Jan 1;4(2):249–64.

16. Geoffrey J McLachlan DP. Finite Mixture Models. Sons JW, editor. 2004;:419.

17. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). John Wiley & Sons, Ltd; 1977 Sep 1;39(1):1–22.

18. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013 Nov;14(6):671–83.

19. Discovery HULDMAK, 2002. Farhad H ussain, Chew Lim T an, M anoranjan Dash," Discretization: An Enabling Technique,."

20. Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: Opportunities and challenges. Neurocomputing. Elsevier; 2017 May 10;237:350–61.

21. Littell RC, Folks JL. Asymptotic Optimality of Fisher's Method of Combining Independent Tests. Journal of the American Statistical Association. Taylor & Francis Group; 2012 Apr 5;66(336):802–6.

22. Cereda M, Pozzoli U, Rot G, Juvan P, Schweitzer A, Clark T, et al. RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. Genome Biol. BioMed Central; 2014 Jan 31;15(1):R20–12.

23. Gambardella G, Cereda M, Benedetti L, Ciccarelli FD. MEGA-V: detection of variant gene sets in patient cohorts. Bioinformatics. 2017 Apr 15;33(8):1248–9.

24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. BioMed Central; 2014;15(12):550–21.

25. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS -driven cancers require TBK1. Nature. Nature Publishing Group; 2009 Nov 1;462(7269):108–12.

26. Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. Tucker-Kellogg G, editor. PLoS Comput Biol. 2008 Nov;4(11):e1000217.

27. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics. BioMed Central; 2005 Sep 12;6(1):225–11.

28. Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010 Sep 1;26(17):2176–82.

29. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004 Jan 1;20(1):93–9.

30. Geistlinger L, Csaba G, Zimmer R. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. BMC Bioinformatics. BioMed Central; 2016 Dec 1;17(1):1–11.

31. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

32. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. Brief Bioinform. 2016 May

17;17(3):393–407.

33. D'Agostino RB, Chase W, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial populations. American Statistician. 1988 Jan 1;42(3):198–202.

34. Butterwoths CRL, 1979. Information Retrieval. 2.

35. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative Genomic Profiling of Human Prostate Cancer. Cancer Cell. 2010 Jul 1;18(1):11–22.

36. Robinson D, Van Allen EM, Wu Y-M, Schultz N, Lonigro RJ, Mosquera J-M, et al. Integrative Clinical Genomics of Advanced Prostate Cancer. Cell. Cell Press; 2015 May 21;161(5):1215–28.

37. Song MS, Salmena L, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor. Nat Rev Mol Cell Biol. Nature Publishing Group; 2012 May 1;13(5):283–96.

38. Yuan TL, Cantley LC. PI3K pathway alterations in cancer: variations on a theme. Oncogene. Nature Publishing Group; 2008 Sep 18;27(41):5497–510.

39. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan 28;43(D1):D447–52.

40. Cereda M, Gambardella G, Benedetti L, Iannelli F, Patel D, Basso G, et al. Patients with genetically heterogeneous synchronous colorectal cancer carry rare damaging germline mutations in immune-related genes. Nat Commun. Nature Publishing Group; 2016 Jul 5;7(1):1–12.

41. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. Nature Publishing Group; 2009 Jul 1;4(7):1073–81.

42. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. Nature Publishing Group; 2010 Apr 1;7(4):248–9.

43. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. Nature Publishing Group; 2010 Aug 1;7(8):575–6.

44. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011 Jul 3;39(17):e118–8.

45. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. Cold Spring Harbor Lab; 2009 Sep;19(9):1553–61.

46. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Human Mutation. John Wiley & Sons, Ltd; 2013 Jan 1;34(1):57–65.

47. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. Cold Spring Harbor Lab; 2010 Jan;20(1):110–21.

48. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. Wasserman WW, editor. PLoS Comput Biol. Public Library of Science; 2010 Dec 2;6(12):e1001025.

49. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009 Jun 15;25(12):i54–62.

50. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Human Mutation. John Wiley & Sons, Ltd; 2016 Mar 1;37(3):235–41.

51. Chen ZH, Zhu M, Yang J, Liang H, He J, He S, et al. PTEN Interacts with Histone H1 and Controls Chromatin Condensation. Cell Reports. Cell Press; 2014 Sep 25;8(6):2003–14.

52. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D277–80.

53. Bononi A, Bonora M, Marchi S, Missiroli S, Poletti F, Giorgi C, et al. Identification of PTEN at the ER and MAMs and its regulation of Ca 2+ signaling and apoptosis in a protein phosphatase-dependent manner. Cell Death Differ. Nature Publishing Group; 2013 Dec 1;20(12):1631–43.

54. Wang M-H, Sun R, Zhou X-M, Zhang M-Y, Lu J-B, Yang Y, et al. Epithelial cell adhesion molecule overexpression regulates epithelial-mesenchymal transition, stemness and metastasis of nasopharyngeal carcinoma cells via the PTEN/AKT/mTOR pathway. Cell Death Dis. Nature Publishing Group; 2018 Jan 5;9(1):1–16.

55. Mulholland DJ, Kobayashi N, Ruscetti M, Zhi A, Tran LM, Huang J, et al. Pten Loss and RAS/MAPK Activation Cooperate to Promote EMT and Metastasis Initiated from Prostate Cancer Stem/Progenitor Cells. Cancer Res. American Association for Cancer Research; 2012 Apr 1;72(7):1878–89.

56. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. Oncogene. Nature Publishing Group; 2017 Mar 1;36(11):1461–73.

57. Hill R, Wu H. PTEN, stem cells, and cancer stem cells. Journal of Biological Chemistry. 2009 May 1;284(18):11755–9.

58. Suzuki A, Kaisho T, Ohishi M, Tsukio-Yamaguchi M, Tsubata T, Koni PA, et al. Critical Roles of Pten in B Cell Homeostasis and Immunoglobulin Class Switch Recombination. Journal of Experimental Medicine. 2003 Mar 3;197(5):657–67.

59. Newton RH, Turka LA. Regulation of T Cell Homeostasis and Responses by Pten. Front Immunol. Frontiers; 2012 Jun 15;3.

60. Cao X, Wei G, Fang H, Guo J, Weinstein M, Marsh CB, et al. The Inositol 3-Phosphatase PTEN Negatively Regulates Fcγ Receptor Signaling, but Supports Toll-Like Receptor 4 Signaling in Murine Peritoneal Macrophages. The Journal of Immunology. American Association of Immunologists; 2004 Apr 15;172(8):4851–7.

61. Garg R, Blando JM, Perez CJ, Abba MC, Benavides F, Kazanietz MG. Protein Kinase C Epsilon Cooperates with PTEN Loss for Prostate Tumorigenesis through the CXCL13-CXCR5 Pathway. Cell Reports. Cell Press; 2017 Apr 11;19(2):375–88.

62. Ortega-Molina A, Serrano M. PTEN in cancer, metabolism, and aging. Trends in Endocrinology & Metabolism. Elsevier Current Trends; 2013 Apr;24(4):184–9.

63. Crackower MA, Oudit GY, Kozieradzki I, Sarao R, Sun H, Sasaki T, et al. Regulation of myocardial contractility and cell size by distinct PI3K-PTEN signaling pathways. Cell. 2002 Sep 20;110(6):737–49.

64. Soundararajan R, Pearce D, Ziera T. The role of the ENaC-regulatory complex in aldosterone-mediated sodium transport. Molecular and Cellular Endocrinology. Elsevier; 2012 Mar 24;350(2):242–7.

65. Milella M, Falcone I, Conciatori F, Cesta Incani U, Del Curatolo A, Inzerilli N, et al. PTEN: Multiple Functions in Human Malignant Tumors. Front Oncol. Frontiers; 2015;5(5308):24.

66. Miller A. Subset Selection in Regression. CRC Press; 2002.

67. Grömping U. Relative importance for linear regression in R: The package relaimpo. Journal of Statistical Software. 2006 Jan 1;17(1):1–27.

68. Westin SN, Ju Z, Broaddus RR, Krakstad C, Li J, Pal N, et al. PTEN loss is a context-dependent outcome determinant in obese and non-obese endometrioid endometrial cancer patients. Molecular Oncology. No longer published by Elsevier; 2015 Oct 1;9(8):1694–703.

69. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. Nature Publishing Group; 2008 Oct 1;455(7216):1061–8.

70. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell. Cell Press; 2015 Nov 5;163(4):1011–25.

71. Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun. Nature Publishing Group; 2018 Aug 13;9(1):1–13.

72. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, et al. The Immune Landscape of Cancer. Immunity. Cell Press; 2018 Apr 17;48(4):812–4.

73. Angelova M, Charoentong P, Hackl H, Fischer ML, Snajder R, Krogsdam AM, et al. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome Biol. BioMed Central; 2015 Dec 1;16(1):1–17.

74. Chen L, Guo D. The functions of tumor suppressor PTEN in innate and adaptive immunity. Cell Mol Immunol. Nature Publishing Group; 2017 Jul 1;14(7):581–9.

75. Jamaspishvili T, Berman DM, Ross AE, Scher HI, De Marzo AM, Squire JA, et al. Clinical implications of PTEN loss in prostate cancer. Nat Rev Urol. Nature Publishing Group; 2018 Apr;15(4):222–34.

76. Zhao J, Chen AX, Gartrell RD, Silverman AM, Aparicio L, Chu T, et al. Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. Nat Med. Nature Publishing Group; 2019 Mar 1;25(3):462–9.

77. Wolf DM, Lenburg ME, Yau C, Boudreau A, van t Veer LJ. Gene Co-Expression Modules as Clinically Relevant Hallmarks of Breast Cancer Diversity. Haibe-Kains B,

editor. PLOS ONE. Public Library of Science; 2014 Feb 7;9(2):e88309.

78. Lausen B, Schumacher M. Maximally Selected Rank Statistics. Biometrics. 1992 Mar;48(1):73.

79. Toso A, Revandkar A, Di Mitri D, Guccini I, Proietti M, Sarti M, et al. Enhancing Chemotherapy Efficacy in Pten-Deficient Prostate Tumors by Activating the Senescence-Associated Antitumor Immunity. Cell Reports. Cell Press; 2014 Oct 9;9(1):75–89.

80. Armstrong CWD, Maxwell PJ, Ong CW, Redmond KM, McCann C, Neisen J, et al. PTEN deficiency promotes macrophage infiltration and hypersensitivity of prostate cancer to IAP antagonist/radiation combination therapy. Oncotarget. Impact Journals, LLC; 2016 Feb 16;7(7):7885–98.

81. Peng W, Chen JQ, Liu C, Malu S, Creasy C, Tetzlaff MT, et al. Loss of PTEN Promotes Resistance to T Cell–Mediated Immunotherapy. Cancer Discov. American Association for Cancer Research; 2016 Feb 1;6(2):202–16.

82. Liu H, Hussain F, Tan CL, Dash M. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery. Kluwer Academic Publishers; 2002 Oct 1;6(4):393–423.

83. Gallego SR, García S, Talín HM, Rego DM, Canedo VB, Betanzos AA, et al. Data discretization: taxonomy and big data challenge. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. John Wiley & Sons, Ltd; 2016 Jan 1;6(1):5–21.

84. Camara PG. Methods and challenges in the analysis of single-cell RNA-sequencing data. Current Opinion in Systems Biology. Elsevier; 2018 Feb 1;7:47–53.

85. Smith AG, Heath JK, Donaldson DD, Wong GG, Moreau J, Stahl M, et al. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature. Nature Publishing Group; 1988 Dec 15;336(6200):688–90.

86. Williams RL, Hilton DJ, Pease S, Willson TA, Stewart CL, Gearing DP, et al. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. Nature. 1988 Dec 15;336(6200):684–7.

87. Tam PPL, Behringer RR. Mouse gastrulation: the formation of a mammalian body plan. Mechanisms of Development. Elsevier; 1997 Nov 1;68(1-2):3–25.

88. Berge ten D, Kurek D, Blauwkamp T, Koole W, Maas A, Eroglu E, et al. Embryonic stem cells require Wnt proteins to prevent differentiation to epiblast stem cells. Nat Cell Biol. Nature Publishing Group; 2011 Aug 14;13(9):1070–5.

89. Koch F, Scholze M, Wittler L, Schifferl D, Sudheer S, Grote P, et al. Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors. Dev Cell. 2017 Sep 11;42(5):514–7.

90. Abranches E, Silva M, Pradier L, Schulz H, Hummel O, Henrique D, et al. Neural differentiation of embryonic stem cells in vitro: a road map to neurogenesis in the embryo. Parise G, editor. PLOS ONE. Public Library of Science; 2009 Jul 21;4(7):e6286.

91. Waddington CH. … strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. Strateg genes a discuss some…; 1957. p. ix+ ….

92. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. Cell. 2007 Feb 23;128(4):635–8.

93. Atlasi Y, Stunnenberg HG. The interplay of epigenetic marks during stem cell differentiation and development. Nat Rev Genet. Nature Publishing Group; 2017 Nov;18(11):643–58.

94. Thomson M, Liu SJ, Zou L-N, Smith Z, Meissner A, Ramanathan S. Pluripotency Factors in Embryonic Stem Cells Regulate Differentiation into Germ Layers. Cell. Cell Press; 2011 Jun 10;145(6):875–89.

95. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006 Aug 25;126(4):663–76.

96. Hynes RO. US policies on human embryonic stem cells. Nat Rev Mol Cell Biol. Nature Publishing Group; 2008 Dec 1;9(12):993–7.

97. Murry CE, Keller G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell. 2008 Feb 22;132(4):661–80.

98. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. Nature Publishing Group; 2013 Mar;14(3):204–20.

99. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. Nature Publishing Group; 2019 Oct;20(10):590–607.

100. Li E, Zhang Y. DNA methylation in mammals. Cold Spring Harb Perspect Biol. 2014 May 1;6(5):a019133–3.

101. Neri F, Krepelova A, Incarnato D, Maldotti M, Parlato C, Galvagni F, et al. Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. Cell. 2013 Sep 26;155(1):121–34.

102. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. Cold Spring Harbor Lab; 2002 Jan 1;16(1):6–21.

103. Myant K, Termanis A, Sundaram AYM, Boe T, Li C, Merusi C, et al. LSH and G9a/GLP complex are required for developmentally programmed DNA methylation. Genome Res. Cold Spring Harbor Lab; 2011 Jan;21(1):83–94.

104. G9a/GLP Complex Maintains Imprinted DNA Methylation in Embryonic Stem Cells. Cell Reports. Cell Press; 2016 Apr 5;15(1):77–85.

105. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. Nature Publishing Group; 2015 Apr 9;520(7546):243–7.

106. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, et al. Intragenic DNA methylation prevents spurious transcription initiation. Nature. Nature Publishing Group; 2017 Mar 1;543(7643):72–7.

107. Gelfman S, Cohen N, Yearim A, Ast G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. Genome Res. Cold Spring Harbor Lab; 2013 May;23(5):789–99.

108. Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, et al. Single-cell DNA methylome

sequencing of human preimplantation embryos. Nat Genet. Nature Publishing Group; 2018 Jan 1;50(1):12–9.

109. Auclair G, Guibert S, Bender A, Weber M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. Genome Biol. BioMed Central; 2014;15(12):545–16.

110. Liao J, Karnik R, Gu H, Ziller MJ, Clement K, Tsankov AM, et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat Genet. Nature Publishing Group; 2015 May;47(5):469–78.

111. Ziller MJ, Ortega JA, Quinlan KA, Santos DP, Gu H, Martin EJ, et al. Dissecting the Functional Consequences of De Novo DNA Methylation Dynamics in Human Motor Neuron Differentiation and Physiology. Cell Stem Cell. Cell Press; 2018 Apr 5;22(4):559–9.

112. Okano M, Bell DW, Haber DA, Li E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell. Cell Press; 1999 Oct 29;99(3):247–57.

113. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992 Jun 12;69(6):915–26.

114. Gao L, Emperle M, Guo Y, Grimm SA, Ren W, Adam S, et al. Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. Nat Commun. Nature Publishing Group; 2020 Jul 3;11(1):3355–14.

115. Manzo M, Wirz J, Ambrosi C, Villaseñor R, Roschitzki B, Baubec T. Isoform-specific localization of DNMT3A regulates DNA methylation fidelity at bivalent CpG islands. EMBO J. John Wiley & Sons, Ltd; 2017 Dec 1;36(23):3421–34.

116. Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, et al. Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat Genet. Nature Publishing Group; 2014 Jan;46(1):17–23.

117. Doetschman TC, Eistetter H, Katz M, Schmidt W, Kemler R. The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. J Embryol Exp Morphol. The Company of Biologists Ltd; 1985 Jun;87(1):27–45.

118. Griffiths JA, Scialdone A, Marioni JC. Using single-cell genomics to understand developmental processes and cell fate decisions. Molecular Systems Biology. John Wiley & Sons, Ltd; 2018 Apr 1;14(4).

119. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. Nature Publishing Group; 2017 Oct 1;14(10):979–82.

120. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. Nature Publishing Group; 2017 Mar 1;14(3):309–15.

121. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature. Nature Publishing Group; 2007 Jul 12;448(7150):196–9.

122. Veillard A-C, Marks H, Bernardo AS, Jouneau L, Laloë D, Boulanger L, et al. Stable methylation at promoters distinguishes epiblast stem cells from embryonic stem cells and the in vivo epiblasts. Stem Cells Dev. 2014 Sep 1;23(17):2014–29.

123. Heger A, Webber C, Goodson M, Ponting CP, Lunter G. GAT: A simulation framework for testing the association of genomic intervals. Bioinformatics. 2013 Aug 15;29(16):2046–8.

124. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28;38(4):576–89.

125. Kuhn M. Building predictive models in R using the caret package. Journal of Statistical Software. 2008 Jan 1;28(5):1–26.

126. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. National Academy of Sciences; 2010 Dec 14;107(50):21931–6.

127. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018 Jan 1;46(D1):D380–6.

128. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 2019 May 22;47(W1):W212–24.

129. Zhou HY, Katsman Y, Dhaliwal NK, Davidson S, Macpherson NN, Sakthidevi M, et al. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. Genes Dev. Cold Spring Harbor Lab; 2014 Dec 15;28(24):2699–711.

130. Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani C-A, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature. Nature Publishing Group; 2019 Dec;576(7787):487–91.

131. Bell E, Curry EW, Megchelenbrink W, Jouneau L, Brochard V, Tomaz RA, et al. Dynamic CpG methylation delineates subregions within super-enhancers selectively decommissioned at the exit from naive pluripotency. Nat Commun. Nature Publishing Group; 2020 Feb 28;11(1):1112–6.

132. Betto RM, Diamante L, Perrera V, Audano M, Rapelli S, Lauria A, et al. Metabolic control of DNA methylation in naive pluripotent cells. Nat Genet. Nature Publishing Group; 2021 Feb;53(2):215–29.

133. Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W170–5.

134. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. Nature Publishing Group; 2014 Jan;9(1):171–81.

135. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. Nature Publishing Group; 2019 Aug;37(8):907–15.

136. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for

assigning sequence reads to genomic features. Bioinformatics 30, 923e930. Vol. 30. 2014. 8 p.

137. Ben L, methods LSSN, 2012. Fast gapped-read alignment with Bowtie 2.

138. Bioinformatics 1GPDPS, 2009. The sequence alignment/map format and SAMtools. Oxford.

139. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W187–91.

140. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. BioMed Central; 2008;9(9):R137–9.

141. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013 Apr 11;153(2):307–19.

142. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011 Jun 1;27(11):1571–2.

143. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016 May 15;32(10):1446–53.

144. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current Protocols in Bioinformatics. John Wiley & Sons, Ltd; 2014 Sep 1;47(1):11.12.1–11.12.34.

Publications

* first author/equal contribution

Published:

3.1 Riccardo M. Betto, Linda Diamante, Valentina Perrera, Matteo Audano, Stefania Rapelli, **Andrea Lauria**, Danny Incarnato, Mattia Arboit, Silvia Pedretti, Giovanni Rigoni, Vincent Guerineau, David Touboul, Giuliano Giuseppe Stirparo, Tim Lohoff, Thorsten Boroviak, Paolo Grumati, Maria E. Soriano, Jennifer Nichols, Nico Mitro, Salvatore Oliviero, Graziano Martello, Metabolic control of DNA methylation in naive pluripotent cells. Nature Genetics 53, 215–229 (2021). https://doi.org/10.1038/s41588-020-00770-2

3.2 **Andrea Lauria***, Serena Peirone*, Marco Del Giudice*, Francesca Priante, Prabhakar Rajan, Michele Caselle, Salvatore Oliviero, Matteo Cereda, Identification of altered biological processes in heterogeneous RNA-sequencing data by discretization of expression profiles, Nucleic Acids Research, Volume 48, Issue 4, 28 February 2020, Pages 1730–1747, https://doi.org/10.1093/nar/gkz1208

3.3 Edoardo Morandi, Matteo Cereda, Danny Incarnato, Caterina Parlato, Giulia Basile, Francesca Anselmi, **Andrea Lauria**, Lisa Marie Simon, Isabelle Laurence Polignano, Francesca Arruga, Silvia Deaglio, Elisa Tirtei, Franca Fagioli, Salvatore Oliviero. (2019) HaTSPiL: A modular pipeline for high-throughput sequencing data analysis. PLOS ONE 14(10): e0222512. https://doi.org/10.1371/journal.pone.0222512

Submitted:

3.4 Stefania Rapelli*, **Andrea Lauria***, Valentina Proserpio*, Guohua Meng, Mara Maldotti, Isabelle Laurence Polignano, Francesca Anselmi, Danny Incarnato, Anna Krepelova, Francesco Neri, Ivan Molineris, and Salvatore Oliviero, Lack of DNMT3B in mouse epiblasts impairs meso-endoderm differentiation.

3.5 Yu Fujimura, Mika Watanabe, Kota Ohno, Yasuaki Kobayashi, Shota Takashima, Hideki Nakamura, Hideyuki Kosumi, Yunan Wang, Yosuke Mai, **Andrea Lauria**, Valentina Proserpio, Hideyuki Ujiie, Hiroaki Iwata, Wataru Nishie, Masaharu Nagayama, Salvatore Oliviero, Giacomo Donati, Hiroshi Shimizu, Ken Natsuga, Hair follicle stem cell progeny heal blisters while pausing skin development.