GENOME-WIDE PREDICTION OF REGULATORS SHAPING CHROMATIN STATE AND GENE EXPRESSION

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Anne Ilse Krämer
aus Deutschland

Basel, 2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von:

FAKULTÄTSVERANTWORTLICHE:
Erik van Nimwegen, Christoph Handschin

KORREFERENT:
Dirk Schuebeler

Basel, den 26. Mai 2020
Prof. Dr. Martin Spiess, DEKAN

CONTENTS

INTRODUCTION

The process of gene regulation is a fascinating process. Investigated for years, still the scientific world lacks appropriate knowledge to explain gene expression of observed phenotypes with an underlying regulatory network. The control of gene expression is a multi-step process that takes place at the epigenetic, transcriptional and translational level and can adapt dynamically to external or internal stimuli. Gene expression is not only regulated through transcription factors binding at promoter regions, close to the transcriptional start sites: the genome wide chromatin state plays a crucial role in the administration of gene expression, by dynamically changing the accessibility of regulatory regions on the DNA genome wide. Distal regulatory regions (enhancers) can activate transcription [1] even if they locate thousands of bases away from the promoter. One gene can be regulated by multiple enhancers with different spatiotemporal activities, which adds yet another level of complexity to the repertoire of expression levels of a given set of genes [2].
Consequently, the gene expression of an observed phenotype is to some extent a result of its genome wide chromatin state defining which regions are accessible, ready to be bound by factors that will interact with a set of other proteins to initiate target gene expression.

Nowadays, researchers make use of established experimental techniques followed by genome wide sequencing such as ATAC or ChIP-seq to capture exactly these informations on DNA level. That is a very promising approach as it is possible to capture the whole genomic architecture at once. Another approach is to measure directly the mRNA expression to find important or novel genes or infer which transcriptional activators could have enabled the transcription.
However, making use of the data requires thorough computational processing and analysis: Several preprocessing steps have to be followed to clean up the data and get it into the right shape for the actual inference of important regulatory regions. Even after preprocessing, the quantification of chromatin state usually requires a sequence of sophisticated and coordinated statistical tools. In the present work, I will outline our approaches in treating genomic and transcriptomic data.

This thesis is structured as follows:
The first two parts form the introduction and will give insights about the current knowledge and hypotheses in genetic regulation (Chapter 1) and high throughput sequencing (Chapter 2). Chapter 3 will then go deeper into the computational approaches that are used in the following manuscripts, but its main focus are the methods and concepts used in Chapter 4.

Chapter 4 describes the main work of my PhD.
Here, I will describe how our newly developed techniques can and will help our understanding of genome wide gene regulation by the analysis of genomic sequencing data. Using high throughput genomic sequencing data, we identified key transcriptional regulators which drive the changes underlying chromatin state across samples

genome wide.
Our results confirmed transcriptional regulators that have already been implicated in embryonic development, stem cell differentiation and circadian regulation. Further, using our extensive library of transcription factor motifs, we were able to predict new regulators and regulatory pathways. To make our tool accessible to everyone dealing with high throughput data, we implemented it as an automated pipeline on our webserver (Chapter 4).

Two smaller side projects of my PhD deal with methods to analyze gene expression data and are described in the following two chapters:
In Chapter 5 we use RNA-seq data in combination with principal component analysis to infer regulatory mechanisms in exercise response in mice. The last Chapter 6 outlines the usage and integration of self-written and already published tools to predict the response of human cancer patients to immune checkpoint therapy from gene expression data.

BIBLIOGRAPHY

[1] Heintzman, Nathaniel D., and Bing Ren. "Finding distal regulatory elements in the human genome." Current opinion in genetics & development 19.6 (2009): 541-549.

[2] Schoenfelder, Stefan, and Peter Fraser. "Long-range enhancer–promoter contacts in gene expression control." Nature Reviews Genetics (2019): 1.

1 GENE REGULATION

The first step of gene regulation is transcription. In 1958, the central dogma of biology was proposed by Francis Crick: "DNA makes RNA and RNA makes protein". Transcription as the first step is itself divided into three sub steps: initiation, elongation and termination. Ever since then, transcription was seen as the key in understanding genetic regulation, with RNA polymerase II being the core protein necessary for the whole process. With the advent of Next-Generation-Sequencing techniques, the view on transcriptional initiation was dramatically extended and a whole machinery of proteins was found to be necessary: general transcription factors, co-activators, cohesions, insulators, enhancers or silencers and epigenetic mechanisms. The dynamic usage of the genome defines cell types, drives development of tissues, keeps the cell homeostasis intact and initiates response to external stimuli. Misregulation at this level of regulation can lead to severe malfunction and diseases. Regulators can act in a cis- or trans manner, depending on where the regulatory sequences occur. Trans-regulatory elements can act independently on the allele (e.g. general transcription factors), cis-regulatory elements are classically defined as enhancers or promoters, depending on their location. Among those, distal regulatory elements and epigenetic regulation have been found to play a vital role in the complex process of gene regulation.

1.1 REGULATION THROUGH TRANSCRIPTION FACTORS

1.1.1 *The different roles of transcription factors*

The first level of regulation is exerted by transcription factors (TF). The number of TFs seems to depend on the genome size of the organism [49], from around 300 for e.coli up to 3000 for humans, where each of them has one or more distinct function. However, this system is highly redundant, a loss of function of one factor can be replaced by another one in many cases. TFs perform the first step in decoding the DNA, which makes them indispensable for all kinds of genetic pathways in living organisms. The number of transcription factors – compared to the 20000 expressed genes – seems rather small. Of course, one transcription factor not only targets one gene and their action is highly dynamic: their regulatory programs differ depending on condition, binding partners and tissues.

Some can also function as repressor by competing for the binding with activating factors and thus block transcription, this mechanism is frequently found in bacteria [23]. The mechanisms that TFs use to interact with the DNA are highly variable: Whereas some can directly recruit RNA polymerases, some can unwrap the DNA and others need a to form complex with other proteins and factors to initiate the transcription.

1.1.2 *Co-regulators*

The tightly coordinated program of gene transcription makes use of co-regulators to fine-tune the transcriptional program. Co-regulators either activate (co-activator) or inactivate (co-repressor) regulatory loci by binding to another transcription factor or altering the chromatin state, largely employed in multiple physiological and pathogenic contexts.

Besides acting as silencer or activator, co-regulators can be classified into three groups: Those which covalently modify histones by acetylation/deacetylation (HATs), examples would be histone acetyltransferase p300 (p300) and CREB-binding Protein (Cbp). Factors of this group can act as activators or repressors, by acetylating or deacetylating the side chains of histone lysines, thus increasing or decreasing the accessibility of the DNA. Members of the second group are part of the TRAP/DRIP/Mediator complex, can recruit that bind to transcription factors, recruit RNA polymerase II (PolII), and interact with the whole transcriptional machinery. The last group consists of factors which make use of ATP to unwind the DNA (SWI/SWF) complex (see Figure 1.1) [1, 20].

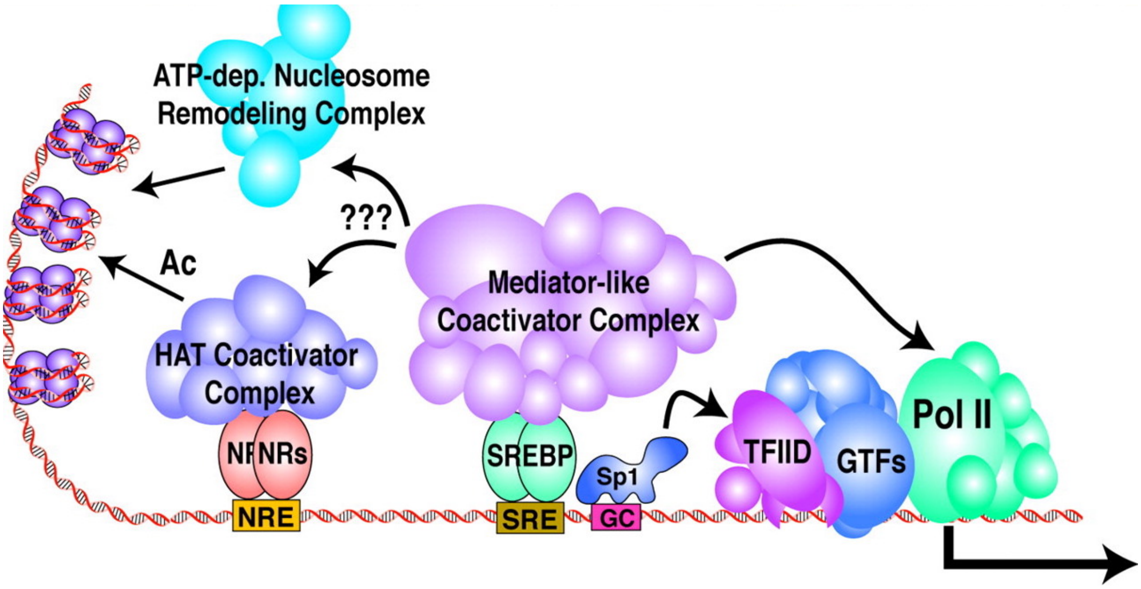


Figure 1.1: **Schematic representation of the three modes of co-regulators.** Chromatin remodeling factors help unwind the DNA and provide access to the DNA (light blue). Nuclear Receptors (NR, orange) have been shown to recruit co-activators with histone acetyltransferase activity (HATs) (purple) to help enabling the transcription. Factors that recruit or are part of the mediator complex (light purple) are needed to interact with the whole transcriptional machinery, for instance the sterol regulatory element-binding protein (SREBP, blue-green). RNA Polymerase II (PolII, green) can initiate the transcription together with general transcription factors (GTFs,blue) and Transcription factor II D (TFIID, pink). Activators like Sp1 (dark blue) help forming the this pre-initiation complex. Taken from [21].

1.1.3 *The transcriptional co-activator Pgc1α*

An example for a highly variable transcriptional co-activator is Peroxisome proliferator activated receptor gamma co-activator 1-alpha (Pgc1α), which uses both histone modification and interaction with other transcription factors. As Pgc1α is a key nodal regulator of metabolism and energy management, its expression is necessary for the regulation of metabolic pathways in many tissues: It guides the adaptation to endurance exercise in skeletal muscle, the thermogenesis in brown adipose tissue (BAT) and has important other functions in liver and brain. Several isoforms of the Pgc1α transcript have been found, their function is still indefinite and remains to be elucidated [32].

Its binding to different partners in a complex has been studied extensively in the past. The most investigated transcription factor that it binds to is the nuclear receptor or estrogen-related-receptor (Errα). The action of the Pgc1α-Errα complex has been found to be important for mitochondrial health and thus metabolism in different tissues.

More specific findings show the implication of PGC1α, ERRα and GA-binding protein (Gabpa) in the oxidative phosphorylation [30], PGC1α coactivation of activator protein 1 (AP1) in the hypoxic gene program [50] and heat shock factor 1 (Hsf1) in the heat shock response [51]. Another study showed forkhead box protein 1 (Foxo1) and hepatic nuclear factor 1 α (Hnf1α) being coactivated by PGC1α in hepatic gluconeogenesis [33, 34].

Chapter 5 examines the implication of Pgc1α in acute response to exercise in mice, for deeper insight into the actions of Pgc1α especially in epigenetic regulation, please refer to Appendix 1.

1.2 EPIGENETIC REGULATION

1.2.1 *Histone marks*

The DNA in each cell is tightly packed into chromatin. But this structure is far from being inert, but rather used to respond quickly by regulating gene expression in response to external stimuli. This is organized by wrapping the DNA around an octameric protein complex, the nucleosome, which itself consists of 8 histone molecules. Intrinsically, this packaging is repressive, supposably because of the bending and physical obstruction. However, histones are not just statically placed on the DNA, they can be also be marked postranslationally which alters their state of accessibility or induce reposition entire nucleosomes [26, 2].

The modifications on histones range from phosphorylation, ubiquitination acetylation and methylation, which have been studied in the past years, to the more recently discovered GlcNAcylation, citrullination, krotonilation and isomerization. This modifications can change dynamically with time. The first histone modification to be discovered was acetylation, found by Allfrey et al in 1964 [27]. Subsequently two groups of proteins that regulate histones were found: histone acetylases (HATs) and deacetylases (HDACs). HATs make use of acetyl CoA to neutralize the initially positively charged lysine by adding an acetyl group to it, which weakens the bond between histones and DNA and thus makes the DNA more accessible. HDACs, contrariwise, restore the positive charge, potentially acting as repressors. Histone methylation is one of the

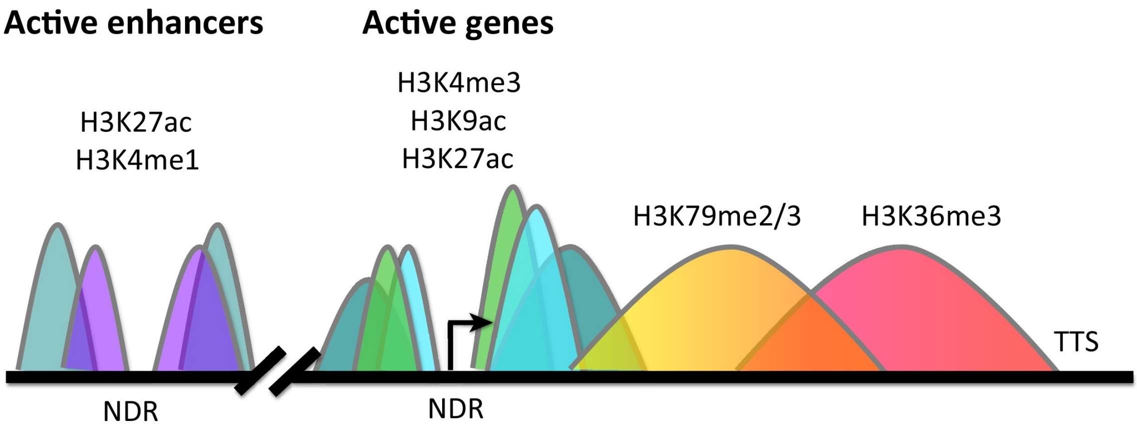


Figure 1.2: **Schematic visualization of the genomic location of different histone modifications.** The arrow denotes the transcription start site (TSS), NDR is the nucleosome depleted regionand TTS is the transcription termination site. Taken from [29].

most studied histone marks, regulated by histone methylases (HKMT) which deposit methyl groups to the histone lysine or arginine side chains (lysines can be mono-, di-, or trimethylated and arginines mono-, symmetrically or asymmetrically methylated). This reaction does not change the charge of the histone and was thought to be static for a long period of time. In 2002, a number proteins with histone-demethylating function were found, which changed the view completely and makes the dynamic usage of histone methylation part of the regulatory network.

Whereas histone acetylation mainly acts as a positive regulator for gene transcription, methylation is more diverse: It does not perturb the initial chromatin structure but rather recruits transcription factors to their loci, which can activate transcription. They can even inhibit the binding of factors which would act as a repressor otherwise. Additionally, histone modifications affect each other: by competition for the same lysine or arginine sites, by disrupting the function of another modification, or by depending on other already deposited modifications [28].
Depending on the function and location of the histone, different lysine tails are targeted (see Figure 1.2). Active promoters have been associated with enriched lysine 4 trimethylation and acetylation of histone 3 and 4 (H3K4me3 and H3ac/H4ac). Histones located in genes in the process of being transcribed show K36me3 and K79me3. Enhancer elements are marked by H3K4me1, if they are activated they show enrichment for H3K27ac. Repressed regulatory elements show H3K9 methylation (H3K9me1) H3K27 trimethylation (H3K27me3) or H3K20 trimethylation (H3K20me3) [29].

1.2.2 *Enhancers*

Enhancers were first discovered in 1981 as a viral repeat sequence (simian virus 40) **??** which could increase the expression of a reporter gene by 200 fold. Two years later, the first enhancer was discovered in mammals. Consecutively, more and more enhancer sequences have been identified or predicted. Nowadays enhancers are thought to be highly implicated into the tissue and cell specific gene programs, especially during embryogenesis [10].

A lot of effort has been put into identifying distal regulatory regions by genetic and biochemical analyses [8, 9]. Although these studies greatly improved the understanding of enhancer function, genome-wide identification of tissue-specific enhancers was precluded by the inability to annotate them systematically based on DNA sequence.
Enhancers are distal elements which are uncoupled from promoters and are able to start the transcription independent of their orientation, which makes them highly variable and enables the activation of distinct pathways during development and a fast response to external stimuli [10]. Genome wide studies have shown that enhancers occur at distal regulatory regions hundreds or thousands of kilobases away, but nevertheless they are targeted in a higher frequency as promoters in lineage-specific gene programs. An example is given by the embryonic stem cells octamer-binding transcription factor 4 (Oct4), SRY-Box Transcription Factor 2 (Sox2) and Nanog Homeobox (Nanog) which share most of their targets, where only 10 % of the binding events take place at the promoter [11, 12]. Most enhancers are not amenable to binding of a TF directly. Specific proteins facilitate the binding: co-activators: histone modifiers, (e.g. acetyltransferases p300/CBP, Gcn5-containing ATAC complex), ATP-dependent chromatin remodelers catalyzing nucleosome movement (e.g. chro- modomain helicase DNA binding protein 7 (CHD7), Brg1 complex (BAF), or mediators of crosstalk with basal transcriptional machinery at promoters (e.g. Mediator complex) [15, 13, 14].
Although it is almost impossible to predict enhancers dirctly from DNA sequence, they leave traces which can be identified: A study conducted in 2007 linked transcriptional regulation with histone modifications (Figure 1.3). Specifically, H3K4me1 was found be deposited at enhancers and H3K4me3 at promoters (1.3A) [17, 16]. Notably, not all enhancers have to be marked, they still can be dynamically activated: Special factors, *pioneer factors*, can bind to wrapped DNA and recruit the necessary machinery to activate the enhancer by unwrapping the DNA or substitution of additional factors which would have been needed otherwise (Figure 1.3B) [18].
Following the hypothesis, histone modifications not only report the location of an regulatory element, but also its mode: Active enhancers are thought to be marked simultaneously with H3K4me1 and H3K27me1, whereas a sole modification of H3K4me1 is only a 'primer' to prepare the enhancer for being activated or poised. Poised enhancers are trimethylated at K27 and still marked by H3K4me1 (Figure 1.3C). The incorporation of the hypermobile histone variant H3.3/H2A.Z makes the enhancer easier accessible and has even been associated with epigenetic memory [52]. Although the annotation of enhancers to their targeted genes remains very challenging, genomic analyses are in concordance about the cell-type specificity of enhancers. Sites which are marked by enhancer associated histone marks vary much more across cell types than promoter marking or CTCF binding sites (see also Chapter 4 of this thesis). Overwhelming is also the fact that the currently predicted number of enhancers which are cell type specific ranges up to hundreds of thousands. Even in yet undifferentiated embryonic stem cells a specific enhancer signature exists whereas promoters are largely invariant [19].

1.2.3 *DNA methylation*

The cytosine nucleotide in the DNA can be methylated (5-methylcytosine (5mC)) and has numerous functions: Earlier, hypermethylation was suggested to be repressive [3, 39, 5], but recent findings associate it with actively transcribed gene bodies and

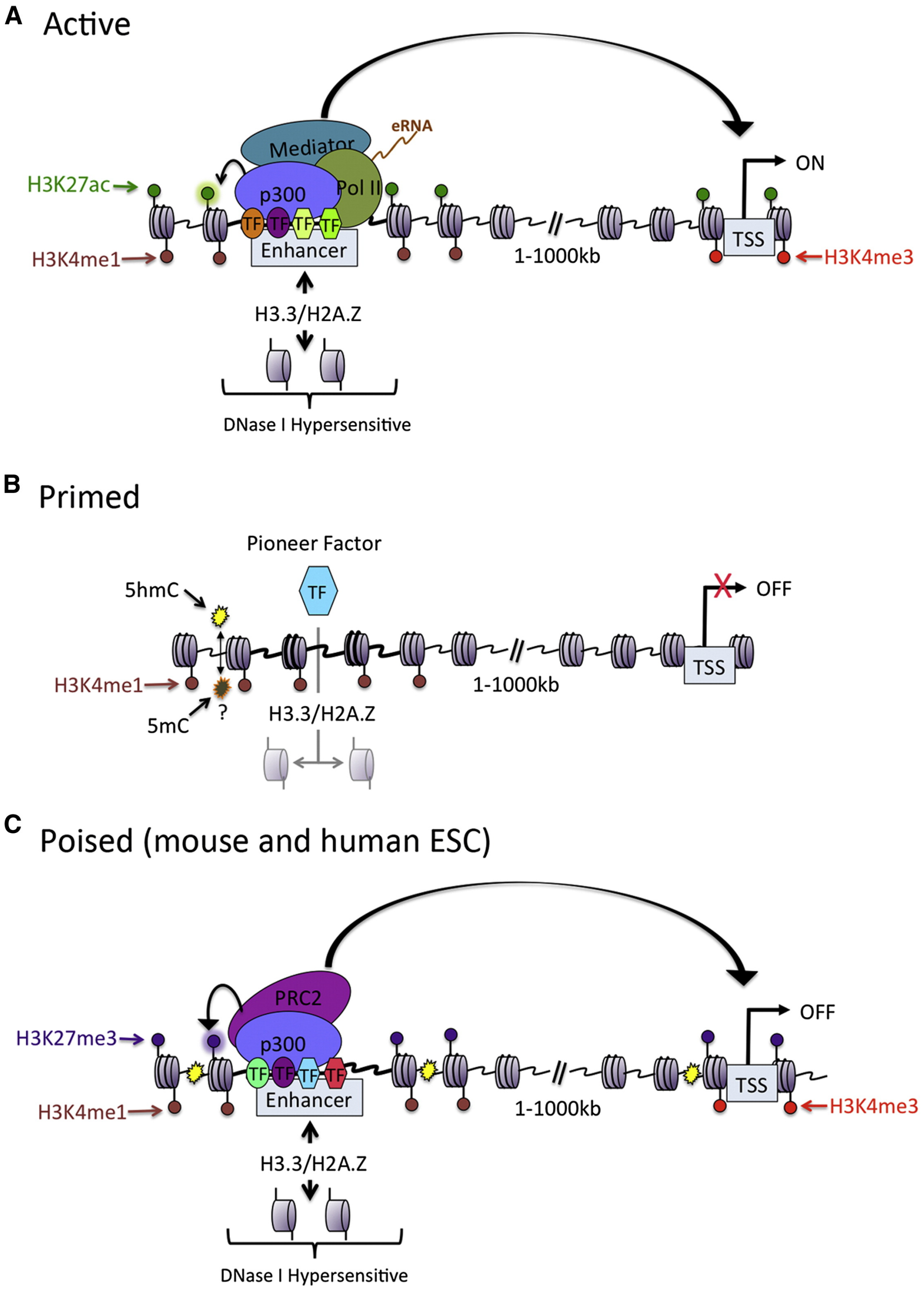


Figure 1.3: **The three suggested principle states of enhancers and the resulting chromatin landscape:** A) Active enhancer: H3K27ac/H3K4me1 is deposited at the enhancer flanking regions and H3K4me3 at the promoter. Hypermobile H3.3/H3A.Z nucleosomes are often incorporated at enhancer locations. A transcription-initiation complex (P300/PolII/Mediator) can bind to the enhancer to enable transcription of the target gene. B) Primed enhancer: Only H3K4me1 is present at the enhancer and no H3K4me3 at the promoter, C) Poised enhancer: H3K27me3 and H3K4me1 can locate both at the enhancer while the promoter may be marked with H3K4me3. PRC binds to the enhancer and communicates with the promoter, but transcription is repressed. Taken from [24].

gene activation per se.
Not all organisms possess 5mC, because methylated regions can easily undergo a point mutation C → T. This leads to fewer CpG content in the genome, explaining the observed lower amount of CpG nucleotides than expected [6, 7].

1.2.4 *miRNA*

The recent advances in high throughput sequencing have revealed several types of RNA, (siRNAs), microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), endogenous small interfering RNAs (endo-siRNAs or esiRNAs), promoter associate RNAs (pRNAs), small nucleolar RNAs (snoRNAs) and sno-derived RNAs. Among those, miRNAs were found to play an important role in post-transcriptional regulation of their messenger RNA (mRNA) targets via mRNA degradation and/or translational repression.

1.3 gene regulation in:

This thesis is about computational analysis and how to get insight into how the complex regulatory structure is governed and how we can make predictions from the data that is being measured. Biological datasets come in many forms – in terms of data structure – but they vary also in terms of hypothesis, experiment, organism, treatment. We analyzed three different types of datasets with various intentions and approaches. This section covers the background knowledge to understand the biological questions in our chosen datasets (Please read Chapters 4 5 and 6 for profound analysis of these datasets).

1.3.1 *Exercise*

Exercise ameliorates skeletal muscle mitochondrial expression and counteracts several severe illnesses such as obesity, cardiovascular diseases, hypertension and type 2 diabetes [32]. It has been shown – among others – to improve lung function, reduce adipose tissue mass and liver fat, increase muscle mass and even has effects on the human psyche [44]. Regular exercise thus affects the whole body. We focus on skeletal muscle, as it is a highly versatile organ capable of adapting to a myriad of external stimuli. Those range from mechanical loading, e.g. the intensity or type of exercise, to the availability of nutrients, hormone signaling, fiber type distribution, temperature and other metabolic programs [43]. Immediately upon initiation of exercise the local demand for ATP, oxygen, glucose and fatty acids increases dramatically. To cope with this, the muscle rapidly starts allosteric regulation and phosphorylation of key enzymes and transporters of glucose and fatty acid oxidation. Simultaneously, the transcription of relevant genes related to energy metabolism is initiated [45]. However, the knowledge of which pathways serve the muscle to adapt to exercise are still relatively unknown.

1.3.2 *Embryonic development and stem cells*

Embryonic development is a highly regulated process, complex and coordinated. A large portion of the regulation is thought to come from alterations in chromatin accessibility and histone modifications. With today's possibilities to examine chromatin state as well as gene expression, a growing number of mechanistic insights and hypotheses have been evolved and proofed in multiple organisms [46, 47]. Most previous studies were aimed at investigating the mRNA expression pattern, although it's becoming more clear that a big part of the dynamic regulation happens in distal regions. The dynamics of enhancer–promoter communication in different cellular contexts and its influence on transcription is not well understood, as the gene-enhancer association is not straightforward and can differ across cell types. We will analyze a dataset in murine embryonic development genome wide and a dataset on human stem cell lineages to decipher parts of the underlying regulatory network.

1.3.3 *Circadian rhythm*

Naturally all living organism show a time-dependent behavior governed by certain external stimuli. The most important stimuli is here the day-night cycle, that has set the oscillation period of time-dependent gene programs in most organisms to a 24h cycle. The supracharismatic nucleus (SCN) receives signals from the ocular photoreceptors, to synchronize and induce tissue-specific independent circadian clocks. On the molecular level, these circadian clock programs are governed by transcription-translation feedback loops: Clock Circadian Regulator (Clock) and Brain And Muscle ARNT-Like 1 (Bmal1) drive the expression of Period 2 (Per2) and cryptochrome 1/2 (Cry1/2) which in turn represses further transcription of Clock and Bmal1. Thousands of genes are downstream of those two master regulators and regulate for example metabolic, humoral signals and body temperature [48].

BIBLIOGRAPHY

[1] Carlberg, Carsten. Target genes of vitamin D: spatiotemporal interaction of chromatin, VDR, and response elements. Vitamin D. Academic Press, 2011. 211-226.

[2] Lawrence, Moyra, Sylvain Daujat, and Robert Schneider. "Lateral thinking: how histone modifications regulate gene expression." Trends in Genetics 32.1 (2016): 42-56.

[3] Ben-Hattar, Jean, and Josef Jiricny. "Methylation of single CpG dinucleotides within a promoter element of the Herpes simplex virus tk gene reduces its transcription in vivo." Gene 65.2 (1988): 219-227.

[4] Watt, Fujiko, and Peter L. Molloy. "Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter." Genes & development 2.9 (1988): 1136-1143.

[5] Iguchi-Ariga, S. M., and Walter Schaffner. "CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation." Genes & development 3.5 (1989): 612-619.

[6] Bird, Adrian P., and Mary H. Taggart. "Variable patterns of total DNA and rDNA methylation in animals." Nucleic Acids Research 8.7 (1980): 1485-1497.

[7] Cooper, David N., and Michael Krawczak. "Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes." Human genetics 83.2 (1989): 181-188.

[8] Granier, Céline, et al. "Nodal cis-regulatory elements reveal epiblast and primitive endoderm heterogeneity in the peri-implantation mouse embryo." Developmental biology 349.2 (2011): 350-362.

[9] Yeom, Young Il, et al. "Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells." Development 122.3 (1996): 881-894.

[10] Buecker, Christa, and Joanna Wysocka. "Enhancers as information integration hubs in development: lessons from genomics." Trends in Genetics 28.6 (2012): 276-284.

[11] Young, Richard A. "Control of the embryonic stem cell state." Cell 144.6 (2011): 940-954.

[12] Chen, Xi, et al. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." Cell 133.6 (2008): 1106-1117.

[13] Roeder, Robert G. "Transcriptional regulation and the role of diverse coactivators in animal cells." FEBS letters 579.4 (2005): 909-915.

[14] D'Alessio, Joseph A., Kevin J. Wright, and Robert Tjian. "Shifting players and paradigms in cell-specific transcription." Molecular cell 36.6 (2009): 924-931.

[15] Weake, Vikki M., and Jerry L. Workman. "Inducible gene expression: diverse regulatory mechanisms." Nature Reviews Genetics 11.6 (2010): 426.

[16] Heintzman, Nathaniel D., et al. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nature genetics 39.3 (2007): 311.

[17] ENCODE Project Consortium. "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature 447.7146 (2007): 799.

[18] Zaret, Kenneth S., and Jason S. Carroll. "Pioneer transcription factors: establishing competence for gene expression." Genes & development 25.21 (2011): 2227-2241.

[19] Hawkins, R. David, et al. "Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency." Cell research 21.10 (2011): 1393.

[20] Spiegelman, Bruce M., and Reinhart Heinrich. "Biological control through regulated transcriptional coactivators." Cell 119.2 (2004): 157-167.

[21] Näär, Anders M., Bryan D. Lemon, and Robert Tjian. "Transcriptional coactivator complexes." Annual review of biochemistry 70.1 (2001): 475-501.

[22] Lambert, Samuel A., et al. "The human transcription factors." Cell 172.4 (2018): 650-665.

[23] Akerblom, Ingrid E., et al. "Negative regulation by glucocorticoids through interference with a cAMP responsive enhancer." Science 241.4863 (1988): 350-353.

[24] Calo, Eliezer, and Joanna Wysocka. "Modification of enhancer chromatin: what, how, and why?." Molecular cell 49.5 (2013): 825-837.

[25] Hernandez-Garcia, Carlos M., and John J. Finer. "Identification and validation of promoters and cis-acting regulatory elements." Plant Science 217 (2014): 109-119.

[26] Bannister, Andrew J., and Tony Kouzarides. "Regulation of chromatin by histone modifications." Cell research 21.3 (2011): 381-395.

[27] Allfrey, V. G., R. Faulkner, and A. E. Mirsky. "Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis." Proceedings of the National Academy of Sciences 51.5 (1964): 786-794.

[28] Zegerman, Philip, et al. "Histone H3 lysine 4 methylation disrupts binding of nucleosome remodeling and deacetylase (NuRD) repressor complex." Journal of Biological Chemistry 277.14 (2002): 11621-11624.

[29] Gates, Leah A., Charles E. Foulds, and Bert W. O'Malley. "Histone marks in the 'driver's seat': functional roles in steering the transcription cycle." Trends in biochemical sciences 42.12 (2017): 977-989.

[30] Mootha, Vamsi K., et al. "Erra and Gabpab specify PGC-1a-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle." Proceedings of the National Academy of Sciences 101.17 (2004): 6570-6575.

[31] Tateishi, Keisuke, et al. "Role of Jhdm2a in regulating metabolic gene expression and obesity resistance." Nature 458.7239 (2009): 757.

[32] Tadaishi, Miki, et al. "Skeletal muscle-specific expression of PGC-1α-b, an exercise-responsive isoform, increases exercise capacity and peak oxygen uptake." PloS one 6.12 (2011).

[33] Puigserver, Pere, et al. "Insulin-regulated hepatic gluconeogenesis through FOXO1–PGC-1α interaction." Nature 423.6939 (2003): 550.

[34] Yoon, J. Cliff, et al. "Control of hepatic gluconeogenesis through the transcriptional coactivator PGC-1." Nature 413.6852 (2001): 131.

[35] Puigserver, Pere, et al. "A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis." Cell 92.6 (1998): 829-839.

[36] Puigserver, Pere, et al. "Activation of PPARγcoactivator-1 through transcription factor docking." Science 286.5443 (1999): 1368-1371.

[37] Wallberg, Annika E., et al. "Coordination of p300-mediated chromatin remodeling and TRAP/mediator function through coactivator PGC-1α." Molecular cell 12.5 (2003): 1137-1149.

[38] Lin, Jiandie D. "Minireview: the PGC-1 coactivator networks: chromatin-remodeling and mitochondrial energy metabolism." Molecular endocrinology 23.1 (2009): 2-10.

[39] Monsalve, María, et al. "Direct coupling of transcription and mRNA processing through the thermogenic coactivator PGC-1." Molecular cell 6.2 (2000): 307-316.

[40] Li, Siming, et al. "Genome-wide coactivation analysis of PGC-1α identifies BAF60a as a regulator of hepatic lipid metabolism." Cell metabolism 8.2 (2008): 105-117.

[41] Lin, Jiandie, Christoph Handschin, and Bruce M. Spiegelman. "Metabolic control through the PGC-1 family of transcription coactivators." Cell metabolism 1.6 (2005): 361-370.

[42] Handschin, Christoph, and Bruce M. Spiegelman. "The role of exercise and PGC1α in inflammation and chronic disease." Nature 454.7203 (2008): 463.

[43] Spiegelman, Bruce, ed. Hormones, Metabolism and the Benefits of Exercise: Furrer, Regula and Handschin, Christoph: Optimized Engagement of Macrophages and Satellite Cells in the Repair and Regeneration of the Exercised Muscle, Springer, 2017.

[44] Voss, Michelle W., et al. "Bridging animal and human models of exercise-induced brain plasticity." Trends in cognitive sciences 17.10 (2013): 525-544.

[45] Catoire, Milene, et al. "Pronounced effects of acute endurance exercise on gene expression in resting and exercising human skeletal muscle." PloS one 7.11 (2012).

[46] Arbeitman, Michelle N., et al. "Gene expression during the life cycle of Drosophila melanogaster." Science 297.5590 (2002): 2270-2275.

[47] Zhang, Wen, et al. "The functional landscape of mouse gene expression." Journal of biology 3.5 (2004): 21.

[48] Ko, Caroline H., and Joseph S. Takahashi. "Molecular components of the mammalian circadian clock." Human molecular genetics 15.suppl_2 (2006): R271-R277.

[49] van Nimwegen, Erik. "Scaling laws in the functional content of genomes." Power Laws, Scale-Free Networks and Genome Biology. Springer, Boston, MA, 2006. 236-253.

[50] Baresic, Mario, et al. "Transcriptional network analysis in muscle reveals AP-1 as a partner of PGC-1α in the regulation of the hypoxic gene program." Molecular and cellular biology 34.16 (2014): 2996-3012.

[51] Xu, L., et al. "The transcriptional coactivator PGC1α protects against hyperthermic stress via cooperation with the heat shock factor HSF1." Cell death & disease 7.2 (2016): e2102-e2102.

[52] Bano, Daniele, et al. "The histone variant H3. 3 claims its place in the crowded scene of epigenetics." Aging (Albany NY) 9.3 (2017): 602.

[53] Banerji, Julian, Sandro Rusconi, and Walter Schaffner. "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences." Cell 27.2 (1981): 299-308.

2 NEXT GENERATION SEQUENCING TECHNIQUES: PROCEDURES, BENEFITS AND CHALLENGES

Orchestrated through a complex interplay of transcription factors, co-activators, chromatin and histones, the gene expression pattern is a main determinant of a cells phenotype. The DNA sequence is thought to encode the bases of this regulation and is used widely to predict potentially functional pattern genome wide using conversation across species. However, less than 10% of our genome falls into conserved regions, and if so, they still cannot be genuinely depicted as being active. For gene transcription to be initiated, it must be accessible and not packed in histones, its promoter has to be bound by an activator and eventually other factors like co-activators have to be present. Recently, it has been found that a large portion of these binding events also happen at distal regions, enhancers. Rigorous methods have been developed to investigate the chromatin state (ATAC-seq and DNase-seq) or the binding of transcription factors (ChIP-seq) genome-wide and will lift our understanding of genomic regulation to the next level.

2.1 GENOMIC METHODS IN ANALYSING DNA ACCESSIBILITY AND TRANSCRIPTION FACTOR LOCALIZATION

As outlined in the previous chapter, the chromatin state of a cell is highly regulated. By making the DNA accessible, active open regions can interact with other sequences and proteins to initiate genetic programs. Thus, understanding the chromatin state by scanning for open regions in the DNA of cells is the first step in understanding the cell-specific complex regulatory network. Importantly, the binding of specific proteins cannot be assessed by analysis of the chromatin state alone. Therefore researchers make use of DNA binding profiles to detect functional binding sites of their proteins of interest.

This chapter will give an overview over the most common techniques to assess accessibility and functional binding sites.

2.2 METHODS TO DETECT ACCESSIBLE REGIONS OR TRANSCRIPTION FACTOR BINDING SITES

2.2.1 *ChIP-seq*

The eternal process in gene transcription is the binding of a transcription factor to the DNA. Of interest are transcription factors and histone marks which can be targeted by an antibody. By combining chromatin immunoprecipitation with deep sequencing it gets possible to study genome wide regulatory elements targeted by transcription factors, cofactors and histone modifications. A typical ChIP-seq experiment starts with crosslinking the DNA-associated proteins to the DNA, followed by a sonication of the

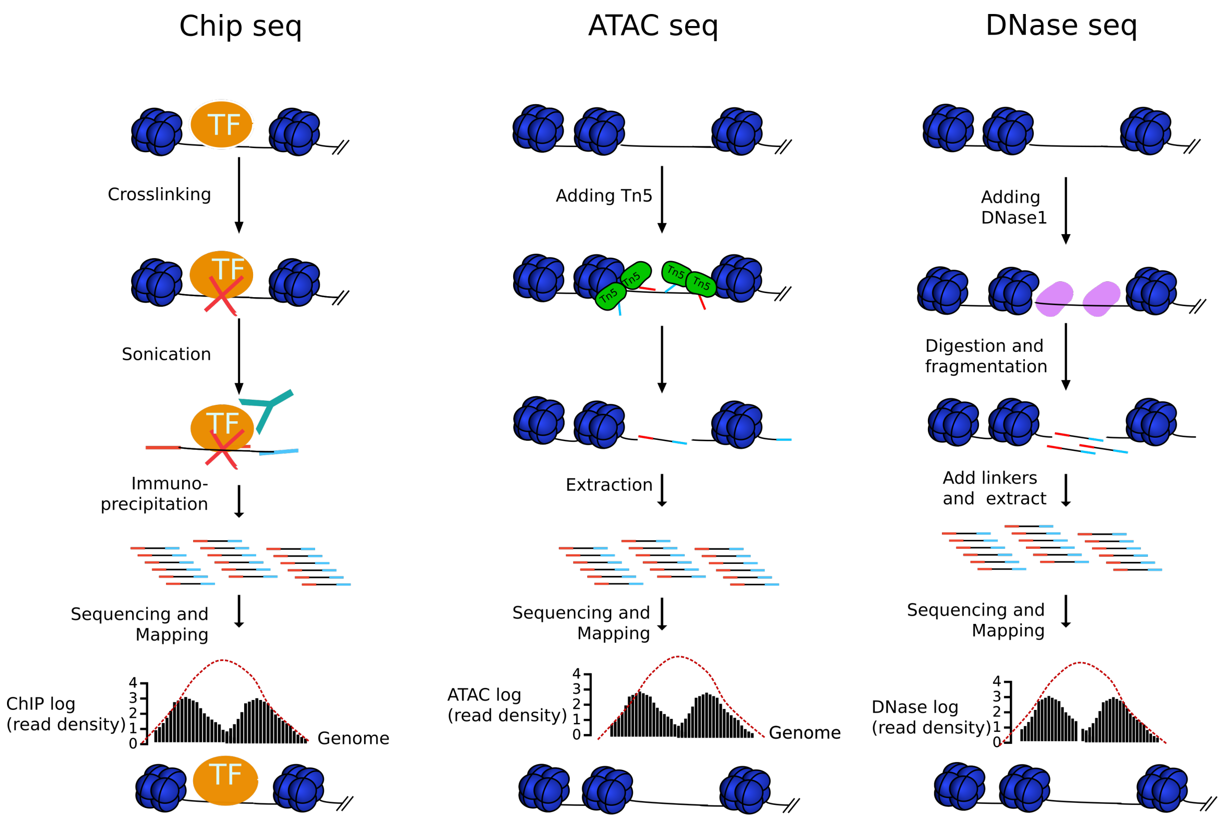


Figure 2.1: **Schematic explanation of the workflow for ChIP seq, DNase seq and Histone ChIP seq.** left: ChIP seq, middle: Histone modification ChIP seq, right: DNase seq/ATAC seq, basic concept taken from [18]. Blue bullets are the nucleosome complexes, TF is an arbitrary transcription factor, the Y-shaped structure is the antibody against the TF. Tn5 is the transposase enzyme.

samples to obtain equally long fragments (see Figure 2.1,*left*). Immunoprecipitation is done using protein specific antibodies (=foreground sample). To find the baseline of fragment counts, a control sample contains a random selection of fragments from the same sample. Fragments are then amplified and oligonucleotide adapters are attached to them to allow for deep sequencing. After sequencing, the location of the transcription factor can be inferred by searching for regions that are enriched for reads in the foreground sample compared to the control sample (see also Chapter 3). ChIP-seq for histone modifications uses a slightly adapted protocol: The fragmentation of DNA is done by using MNase to digest DNA. Unfortunately, ChIP-seq has certain limitations. A large amount of data is necessary to achieve the necessary sequencing depth, in the range of 10^7 cells which is not always feasible – for example in small organisms or very specific tissues. However, recent protocols even managed to perform ChIP-seq in single cells using a microfluidics [23]. Another bottleneck of the ChIP seq protocol is to find the right antibody. If a lot of non-specific binding events occur, the analysis will be confined by false positive peaks. Therefore, ChIP-seq experiments always include a control sample. This can be either an "input" DNA, which is the DNA before immunoprecipitation or substitution of an unspecific antibody such as anti-IgG [24]. One of the most important problems is the required prior knowledge of important transcription factors which makes it hard get an unbiased overview of the genetic state of a cell or organism in an experiment. Recently, new techniques have emerged, the most important being DNase and ATAC-seq.

2.2.2 *DNase-seq*

DNase-seq makes use of enzymatic digestion of accessible regions with the non specific endonuclease DNase1. Initially this technique included running twice the experiment on a gel with and without the protein of interest and checking which basepairs change in enrichment [7]. The first DNase followed by deep sequencing was established in 2006 [13].
To perform DNase-seq, nuclei are isolated from cells and exposure to DNase1, then degrading RNA and proteins and purifying the DNA (see Figure 2.1,*middle*). The fragments of desired size (usually around 130-160bp in length, according to the length of DNA wrapped around 1 nucleosome [8]). The fragments least abundant are those targeted by DNase1 and recognized as most accessible (DNase 1 hypersensitive sites). After sequencing, enriched regions show almost the same pattern as in ChIP-seq and can be inferred by traditional ChIP-seq analysis (Chapter 3)). DNase-seq has influenced the genomic research by providing information about promoters and enhancer. Despite DNase has a slight sequence bias towards minor grooves in the DNA [16], lots of highly influential papers have been published using this method, for example an atlas of all known cis-regulatory regions [9], and the atlas of tissue- and cell-specific differences, provided by the ENCODE project and the Roadmap Epigenomic Consortium [10, 11, 12].

2.2.3 *ATAC-seq*

This method is a fast and sensitive alternative to DNase-Seq for assaying chromatin accessibility genome-wide, first applied in 2013 [14]. As with DNase-seq, the sequenced reads give information of regions with increased accessibility, active sites of transcription factor binding and nucleosome position. Here, a hyperactive Tn5 transposase binds to open regions not covered by histones, yet accessible regions (see Figure 2.1,*right*). The transposase directly inserts sequencing adapters to the bound fragment [1]. Also the ATAC-seq derived regions which denote the accessibility can be inferred via methods for ChIP seq analysis (Chapter 3). The successful technology has been used in a multitude of different contexts, for example in yeast, plants, nematodes, flies, mammals, and even frozen tissues [15]. Because ATAC-seq relies on the insertion of the transposase into open chromatin, rather than digesting it, mitochondrial (mt)DNA is often overrepresented and has to be corrected for [17]. In comparison, ATAC-seq and DNase-seq agree principally on the identified sites, while both methods have distinct sequence biases. DNase showed better results in footprinting TFs, where both approaches yielded different results [3]. While DNase is rather time - consuming (a standard protocol requires approximately 3days to finish), an ATAC-seq library can be prepared in about 2-3h [4]. The simplified procedure also diminished experimental errors and thus increases reproducibility.

2.3 METHODS TO ASSESS TRANSCRIPTION

2.3.1 *RNA-seq*

RNA sequencing was developed more than 10 years ago. Mostly used for finding differential expressed genes (DE), it is used for finding isophorms, new transcripts like long non-coding (lnc)RNAs and distal regulatory regions (enhancers). The experimental procedure to prepare the library for RNA-seq consists of RNA extraction and enrichment (or depletion or ribosomal (r)RNA), then synthesis of complementary (c)DNA and ligation of adapters for the sequencing. Normal sequencing depth is about 10-30 million reads per sample.

2.3.2 *Microarray*

Before RNA sequencing was established, the microarray technique was the one most used to determine the expression level of a gene across different samples. Here, certain probes (cDNA sequences of known identity) are immobilized on a plate, and samples from the different conditions are marked with different fluophores. When the samples are washed over the plate, they hybridize with the probes and the fluorescence emission is measured to quantify from which sample the fragments for each gene came [25].

2.4 SEQUENCING TECHNIQUES

Sanger sequencing was for almost 30 years the acronym for sequencing, named by Frederic Sanger who invented the successful "chain termination method". In summary, this technique uses single strand DNA and chemically-altered deoxynucleotide (dATP, dCTP, dGTP, dTTP), which are incorporated randomly by the DNA polymerase and terminate the chain replication of single-stranded DNA. The resulting fragments differ in size and can be divided by gel-electrophoresis. By doing 4 such runs with different modified nucleotides, it is possible to reconstruct which nucleotide belongs to which position.

Most of the data available nowadays was produced by the widely used NGS platform Illumina, former Solexa. This technique is called cyclic reversible terminator technology and takes place on flow cells, containing billions of fragments. Specific adaptors are ligated to the fragments 3′ and 5′ ends, and complementary adapters are placed as anchors on the flow cell. The fragments coming from the experiment then attach to the hybridize with the anchor and are bound to the flow cell (Figure 2.2, *upper panel*). After this, a few cycles of amplification take place: The single strands bend over to other adjacent anchors and are copied starting from the anchor region *bridge amplification* (Figure 2.2, *middle*). This process is repeated until the desired amount of fragments is present. Then the original template strands (the reverse strands) are washed away and the 3′ ends are blocked to prevent further amplification. Then modified nucleotides – the same as in Sanger sequencing – are washed over the plate, each carrying a fluophore, which can be identified by laser. Each washing one nucleotide is added to the each the sequences and the chain reaction is stopped by the modified nucleotide. After detection of the fluophore, it is removed and the reaction continues (Figure 2.2,

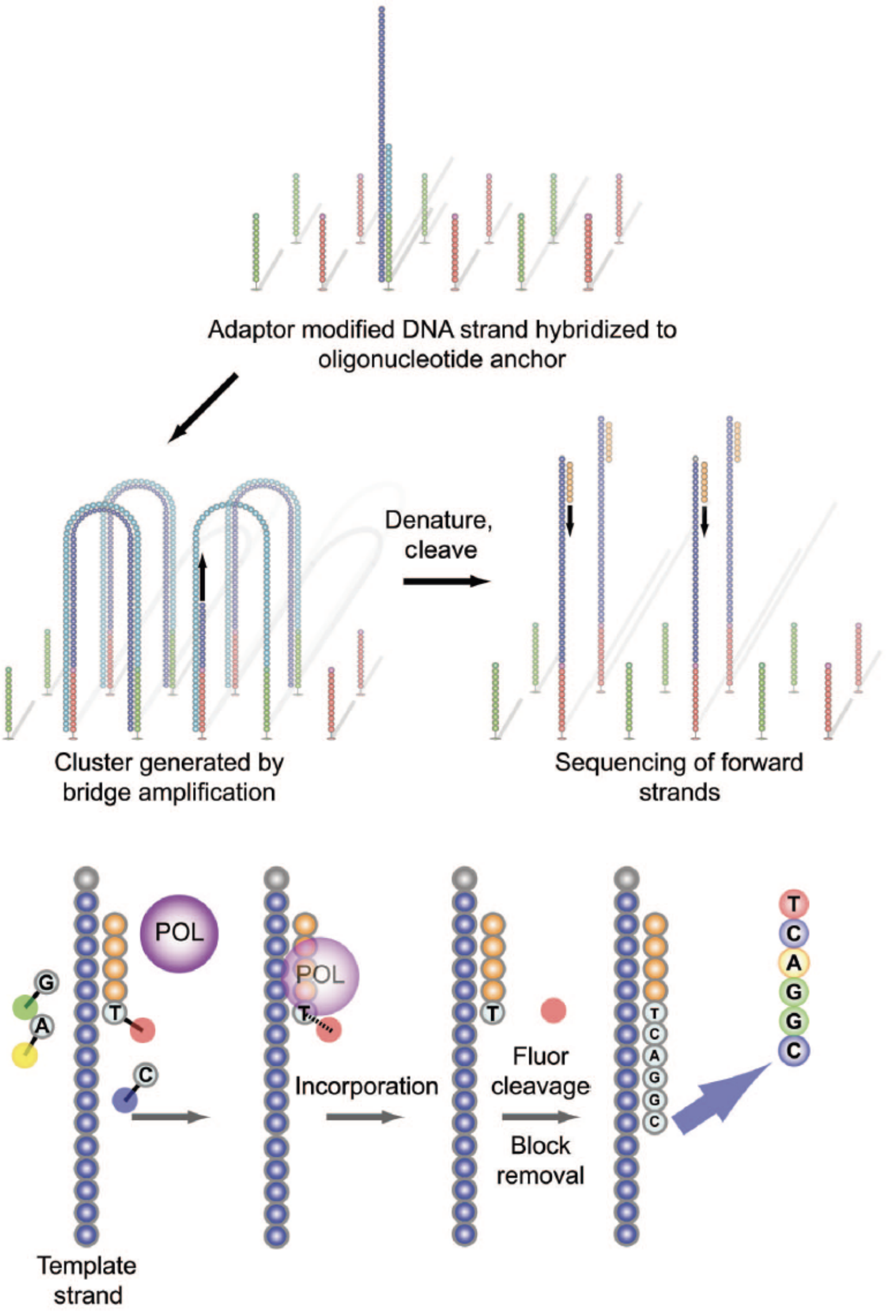

Adaptor modified DNA strand hybridized to oligonucleotide anchor
Denature, cleave
Cluster generated by bridge amplification
Sequencing of forward strands
POL
G
A
T
C
Incorporation
Fluor cleavage
Block removal
T
C
A
G
G
C
Template strand

Figure 2.2: **Representation of the NGS technique for sequencing**. *Upper panel*: Ligation to the anchors. *Middle* Bridge amplification, *Lower panel*: Single strands bend over and are copied. Detailed sequencing process. Each cycle a new nucleotide is incorporated, its color is recorded and the chain reaction stops. Taken from [22].

lower panel). In difference to Sanger, here, each nucleotide terminates the replication, which makes it much more effective. This technique captures the sequence of the coding strand, however it is also possible to obtain both strands by paired end sequencing. In this case, the clusters are regenerated, but this time the the 3′ blocking is removed, another step of bending is performed and this time the forward strand is washed away [22].

2.5 CHALLENGES

The decreasing cost of sequencing led to a tremendous increase of available sequencing data during the last decade. However, the capacities to handle the data are getting more and more pushed to the limits. The analysis of this type of data is highly challenging, due to the extremely heterogeneous structure. Dedicated statistical models are developed to get rid of white noise and model the data to transform it into meaningful results. Given the number of current available tools, it requires certain knowledge to apply the right models to the right data with the right parameters. Using a tool with different parameter settings may result in very different results and interpretations. Even using different version of one program may lead to non-overlapping outcomes. This poses a problem in terms of reproducibility of scientific studies. The increasing amount of data also poses challenges to the available infrastructure: A whole human genome takes up to 140 GB in disk-space. Tools doing calculations on the data can need up to 100GB of RAM to run. Bioinformaticians are thus pressed to find the most economic way of processing and storing the data, without risking the loss of information.

All these reasons lead to eventual poor reproducibility in the analysis of high - throughput experiments. Therefore, adequate documentation of the programming languages, versions, packages and parameters is required. Raw data and the used code has to be made accessible for future researchers. However, as self-built scientific pipelines usually depend on the developer's computing environment, it can be very challenging to make the pipeline publicly available although this should be common practice. To interpret the results of a computational analysis, one has to be aware that everything is a prediction and based on statistical models. Thus, every gene or transcription factor that is reported, is reported with a certain probability. Normally, the false positive rate of the predictions is controlled and set to mostly 5-10 %. Still, a critic view on the data and validation through previously published literature or even experimental approaches is definitely a plus in the process of understanding genetic regulation.

BIBLIOGRAPHY

[1] Buenrostro, Jason D., et al. "ATAC-seq: a method for assaying chromatin accessibility genome-wide." Current protocols in molecular biology 109.1 (2015): 21-29.

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

[3] Calviello, Aslıhan Karabacak, et al. "Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling." Genome biology 20.1 (2019): 42.

[4] Sun, Yuanyuan et al. "Detect accessible chromatin using ATAC-sequencing, from principle to applications." Hereditas vol. 156 29. 15 Aug. 2019, doi:10.1186/s41065-019-0105-9

[5] Klein, David C., and Sarah J. Hainer. "Genomic methods in profiling DNA accessibility and factor localization." Chromosome Research (2019): 1-17.

[6] Crawford, Gregory E., et al. "DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays." Nature methods 3.7 (2006): 503-509.

[7] Galas, David J., and Albert Schmitz. "DNAase footprinting a simple method for the detection of protein-DNA binding specificity." Nucleic acids research 5.9 (1978): 3157-3170.

[8] He, Housheng Hansen, et al. "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification." Nature methods 11.1 (2014): 73.

[9] Thurman, Robert E., et al. "The accessible chromatin landscape of the human genome." Nature 489.7414 (2012): 75-82.

[10] Dunham, Ian, et al. "An integrated encyclopedia of DNA elements in the human genome." (2012).

[11] Maurano, Matthew T., et al. "Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo." Nature genetics 47.12 (2015): 1393.

[12] Kundaje, Anshul, et al. "Integrative analysis of 111 reference human epigenomes." Nature 518.7539 (2015): 317-330.

[13] Crawford, Gregory E., et al. "DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays." Nature methods 3.7 (2006): 503-509.

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

[15] Corces, M. Ryan, et al. "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues." Nature methods 14.10 (2017): 959-962.

[16] Lazarovici, Allan, et al. "Probing DNA shape and methylation state on a genomic scale with DNase I." Proceedings of the National Academy of Sciences 110.16 (2013): 6376-6381.

[17] Rickner, Hannah D., Sheng-Yong Niu, and Christine S. Cheng. "ATAC-seq Assay with Low Mitochondrial DNA Contamination from Primary Human CD4+ T Lymphocytes." JoVE (Journal of Visualized Experiments) 145 (2019): e59120.

[18] Wikipedia Commons image. Downloaded in February 2020 at `https://commons.wikimedia.org/wiki/File:Figure.1_ATAC-Seq_illustration.svg`

[19] Carninci, Piero, et al. "High-efficiency full-length cDNA cloning by biotinylated CAP trapper." Genomics 37.3 (1996): 327-336.

[20] Bray, Nicolas L., et al. "Near-optimal probabilistic RNA-seq quantification." Nature biotechnology 34.5 (2016): 525-527.

[21] Stark, Rory, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years." Nature Reviews Genetics 20.11 (2019): 631-656.

[22] Voelkerding, Karl V., Shale A. Dames, and Jacob D. Durtschi. "Next-generation sequencing: from basic research to diagnostics." Clinical chemistry 55.4 (2009): 641-658.

[23] Grosselin, Kevin, et al. "High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer." Nature genetics 51.6 (2019): 1060-1066.

[24] Flensburg, Christoffer, et al. "A comparison of control samples for ChIP-seq of histone modifications." Frontiers in genetics 5 (2014): 329.

[25] Govindarajan, Rajeshwar, et al. "Microarray and its applications." Journal of pharmacy & bioallied sciences 4.Suppl 2 (2012): S310.

3 NEXT GENERATION SEQUENCING: DATA ANALYSIS

With the increasing amount of high throughput sequencing data, a myriad of new doors have been opened to increase our understanding of the complex nature of genetic regulation. As the data is highly dimensional and noisy, it is crucial to analyze the data in the right way to draw meaningful conclusions. The natural noisiness of the data stems from from experimental and sequencing biases, but also on the innate fluctuations in biological systems. To disentangle real effects from artefacts and to reliably make predictions about the underlying biological system, a multitude of tools using sophisticated statistical models and machine learning technologies have been developed. However, most of these tools require proper understanding of the algorithms in order to applicate them correctly. This prerequisite makes data analysis very challenging for researchers without computational background, but also induces a huge problem of reproducibility of results if not documented correctly. We will outline our way of treating the data starting from raw files. Our approaches have partly been included into a standalone pipeline.

3.1 INFERRING TRANSCRIPTION FACTOR BINDING SITES FROM HIGH THROUGHPUT GENOMIC SEQUENCING DATA

The current format of data coming from the sequencing facilities is fastq. A fastq file stores, for every sequenced read, 4 measures:

- line 1: the identifier, usually with the instrument name and the flowcell lane number and tile number.
- line 2: the sequence of the read based on the IUPAC notation, consisting of "A","T","C","G" and "N".
- line 3: contains just a "+", may followed by optional identifiers or description.
- line 4: contains a qualityscore in ASCII characters for each letter of the sequence in line 2 ("!" denotes lowest quality and ≈ highest quality).

For ChIP-seq the data comes usually with a control sample as outlined in the previous chapter. In case of DNase or ATAC-seq, no control is provided. There are different ways to deal with a lacking control: several tools estimate the background read density by assuming a random uniform background distribution (MACS2 [3], HOMER[4], SICER [5]) or from flanking windows of enriched regions (SPP [6]). [2]. We assume the background in cases like this to be distributed uniformly.

3.1.1 *Preprocessing*

We use and adapt the previously published pipeline CRUNCH [10] for the first steps of our analysis. The preprocessing consists in general of qualityfiltering and adapter trimming.

QUALITY FILTERING There are several ways to check for the quality of the reads. FastQC [23] can be used to assess the files in terms of quality scores, GC content, sequence length distribution, sequence duplication levels, k-mer overrepresentation and contamination of primers and adapters in the sequencing data. This is useful to get a grasp on the overall quality of the data. `cutadapt` [22] for example uses a 5′ and 3′ end quality trimming, which is useful for Illumina reads as their quality degrades towards the 3′ end. As the quality is already encoded in the 4th line of the fastq file, `cutadapt` makes use of the *Phred* score: $Q_{\mathtt{phred}} = -10 \log_{10} p$ with p as the probability that the called base is incorrect. Note that the ascii values encoding the quality in fastq files range from 33 to 126, thus there is an offset of 33. This offset can be different depending on the sequencing pipeline which is used, so it is important to first guess which the encoding of the files. `cutadapt` trims the reads according to the following procedure:

Our threshold is 20, so assuming we have these quality scores:
`52, 50, 36, 37, 18, 17, 21, 14, 12, 13`
then we subtract our threshold:
`32, 30, 16, 17, -2, -3, 1, -6, -8, -7`
Form the partial sums: sum up the numbers starting from the end. The threshold is reached with the first number which is greater than 0 (here in brackets):
`(70), (38), 8, -8, -25, -23, -20, -21, -15, -7`
Then the read is trimmed up to the minimum of this series (in this case -25).

ADAPTER TRIMMING For the adapter trimming it is important to find the most represented adapter in the files. We make use of `cutadapt` as it can quality trim and trim the adapters in the same step. This is especially useful as we are dealing with large datasets and economic usage of the tools saves time and resources without lowering quality standards of the analysis.

```
cutadapt --quality-cutoff 20 --quality-base=$ENCODING --minimum-length 25 --trim-n
    \
--max-n 2 -a $ADAPTER --overlap 10 --cores 12 \
```

We use a pre-defined list of commonly used adapter sequences [34]:

`GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG` (OligonucleotideSequencesforGenomicDNA)
`ACACTCTTTCCCTACACGACGCTCTTCCGATCT` (TrueSeq Universal Adapter)
`TGGAATTCTCGGGTGCCAAGG` (normally used for RNA trimming, we include here, too)
`GATCGGAAGAGCACACGTCTG` (TrueSeq index Adapter)
`TCGTATGCCGTCTTCTGCTTG` (TrueSeq index Adapter)
`CAAGCAGAAGACGGCATACGAGAT` (PCR Primer)
`AATGATACGGCGACCACCGAGATCTACAC` (PCR Primer)
`GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG` (Paired End Adapters)

`GATCGGAAGAGCACACGTCTGAACTCCAGTCAC` (TrueSeq index Adapter)

MAPPING TO THE REFERENCE GENOME After obtaining the cleaned and quality-filtered reads, the next step is mapping to the reference genome. A multitude of tools exists here, for DNA sequencing most frequently used are STAR [19], Bowtie [20], BWA [21]. In our approach, we make use of Bowtie1, a fast, effective and accurate algorithm using the Burrows Wheeler indexing procedure. This only uses 1.3GB of memory and 2GB of RAM to map to the human genome. [40]. Bowtie has also been used especially for short read mapping.
In our case, we make use of the older version, Bowtie1. This version runs faster and and was found to be more sensitive, especially to reads that are <50bp long [10].
Bowtie reports only one mapping position for one read(-a strata -best), choosing the ones with least mismatches , while allowing for 3 mismatches (-v 3). Multi mapping reads are then distributed across their matching positions n , keeping a weight $w = 1/n$. We call this format *bedweight* and it is used later in the pipeline to estimate exact read counts for specific regions on the genome. Typically, more than 70-80% of the input reads are mapped.

FRAGMENT SIZE ESTIMATION Another important step in quantifying DNA sequencing data is to infer the fragment length as the signal obtained by mapping the sequenced reads stems from the regions flanking the binding site of the chipped transcription factor. As most of the datasets we looked at are sequenced single end, they don't contain any information about the fragment length. We estimate the fragment length d with maximizing the correlation c given by:

$$c(d) = \sum_{i \notin r} r_+(i) r_-(i+d) \tag{1}$$

the sum runs over all regions excluding repeat sequences r, those are repeated in the same way across the genome.

$$r_+(i)r_-(i+d) = \begin{cases} 1, & \text{if a read occurs at position } i \text{ and another one at position } i+d \\ 0, & \text{if no read occurs at position } i \text{ together with one at position } i+d. \end{cases}$$

Having determined the fragment length, we shift the reads $\frac{d}{2}$ up or downstream, depending whether they map to the sense or antisense strand respectively. Like this we are able to quantify the strength of the binding directly at the estimated binding position of the transcription factor, or the middle of the accessible region.
We use the exact same procedure here for ATAC and DNase-seq data, although we are aware that determining the middle of the fragments as center of the peaks only works for fragments that do not include a nucleosome (no insert).

3.1.2 *Identification of important regulatory regions*

COUNTING FRAGMENTS IN SLIDING WINDOWS Prior to the actual peak calling step, we generally identify regions which have a high number of reads coming from the ChIP, ATAC or DNase experiment. Assuming that if there hasn't been any binding event or accessible region, those read densities shouldn't differ from the control samples. We thus compare genome-wide the read densities between foreground and background samples.

We slide a window of 500bp across the genome, shifted by 250bp in each step. The length of the window is in accordance with the expected signal: Transcription factor ChIP-seq usually yields 100-150bp long fragments of the DNA which is also the length of a DNA fragment wrapped around histones, and thus the obtained fragment length for ATAC or DNase experiment (see also figure 3.3) [2]. The length of 500bp thus provides good resolution and will capture most of the binding events and accessible regions. In case of overlapping or adjacent enriched regions, we merge them before scanning explicitly for peaks (see 'merging of peaks'). The background is fluctuating much slower and the read density is typically lower. We account for this by taking a larger window for the background samples, 2000bp by default.

When we run both sliding window across the whole genome we simply count all the reads that fall into each region each for foreground n and background m. In the case of replicates, the raw counts for all foreground and background replicates are summed up.

IDENTIFICATION OF TRULY ENRICHED REGIONS We then compute the read density *counts* for each region i by normalizing the fore- and background counts m and n by the total library size N and M respectively:

$$\text{counts}_i = \frac{\log(n_i/N)}{\log(m_i/M)} \tag{2}$$

We observed previously that some regions yield abnormally high read densities in ChIP-control samples. We exclude those regions from our analysis in ChIP-seq analysis, as well as regions mapping to chrM in ATAC-seq and DNase-seq analysis.

Once we have quantified the counts across the genome, the crucial step is now to distinguish bound from unbound regions in ChIP-seq (bound by a transcription factor) and open from closed regions in ATAC-seq (bound by the transposase / digested by DNase).

We then assume the background read density to be a mixture of 1) variations in the biological state of the cells and experimental variations in the library preparation, and 2) a sampling noise from the sampling of fragments itself. That means we can approximate the read density in an unperturbed sample (without any intervention like immunoprecipitation, DNase digestion or addition of a transposase) by a multiplicative model of log normal- and poisson noise. As we have shown previously [10], the fluctuations in next-generation sequencing read-densities across replicate experiments can be well approximated by a combination of multiplicative noise (which may results from uncontrolled variations both in the biological state of the cells and variations in the process of preparing a sequence library from the sample) and poisson sampling noise (from the sequencing itself), which leads to an approximately log-normal distribution of read-counts.

In other words, we want to estimate the probability P to find n out of N fragments in the foreground and m out of M fragments in the background:

$$P(n, m|\sigma, \mu, N, M) = \frac{1}{\sqrt{2\pi(2\sigma^2 + \frac{1}{n} + \frac{1}{m})}} \exp\left(- \frac{\left(\log(\frac{n}{N}) - \log(\frac{m}{M})\right)^2}{2(2\sigma^2 + \frac{1}{n} + \frac{1}{m})} \right) \tag{3}$$

Additional parameters of this model are $2\sigma^2$ which is the variance of the multiplicative noise components, $\frac{1}{n} + \frac{1}{m}$ the noise coming from the poisson sampling in foreground and background.

We do not expect the average read densities to have the same mean value. A large fraction of the reads in the foreground will be located in enriched regions, thus the read density in not-enriched regions will be substantially lower than in the background. To account for this, we fit an additional parameter, μ to the model:

$$P(n, m|\sigma, \mu, N, M) = \frac{1}{\sqrt{2\pi(2\sigma^2 + \frac{1}{n} + \frac{1}{m})}} \exp\left(-\frac{\left(\log(\frac{n}{N}) - \log(\frac{m}{M}) - \mu\right)^2}{2(2\sigma^2 + \frac{1}{n} + \frac{1}{m})}\right) \quad (4)$$

μ is thus the $\log(\frac{FG}{BG})$ shift in unbound regions (see Figure 3.1), which should scale reversely with the amount and size of the peaks. μ allows us to quantify also differences in binding strength across samples.

We further assume that - if a region is significantly enriched - this distribution is just shifted up by a constant factor $(1-\rho)\frac{1}{W}$, where ρ is the estimated fraction of unbound regions and W is the difference between the maximum and the minimum of normalized read densities:

$$W = \max\left(\frac{\log(n/N)}{\log(m/M)}\right) - \min\left(\frac{\log(n/N)}{\log(m/M)}\right) \quad (5)$$

Taken all together, the total log likelihood of the data is then given by:

$$L(D|\sigma, \mu, \rho, N, M) = \prod_i \rho P(n_i, m_i|\sigma, \mu, N, M) + (1-\rho)\frac{1}{W} \quad (6)$$

We fit σ, ρ and μ by maximizing the log likelihood (eq. 6) by an expectation-maximization approach.

To finally quantify which regions differ significantly between fore- and background,

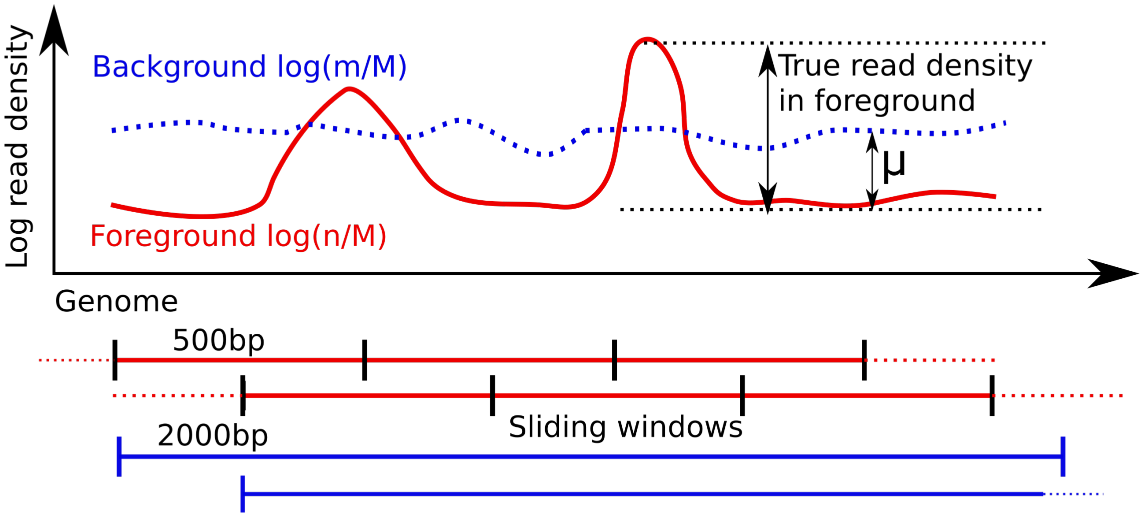


Figure 3.1: **Inference of the log read density.** Sketch of the foreground and background read densities and sliding windows across a piece of the genome. By sliding a window across the genome, we count all fragments in this window. Note that the window in the background is 2000bp. We expect most of the reads in the foreground to map to bound regions, which reduces the amount of reads in unbound regions. We therefore infer the difference of foreground and background levels in unbound regions μ.

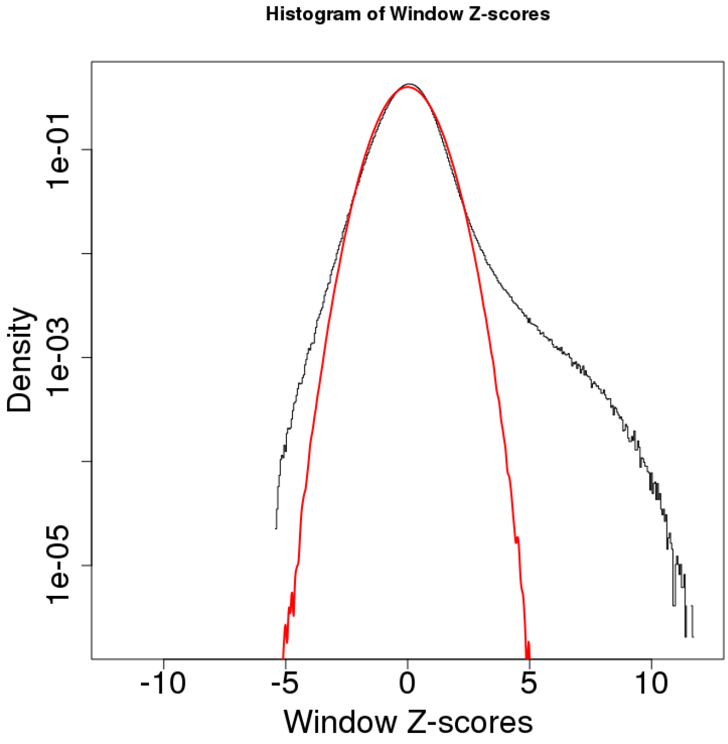


Figure 3.2: **Distribution of zScores for one the dataset of ATAC-seq in murine embryos.** Red line denotes the standard distribution, black is the real data. Note that we plot the histogram in log scale. Clearly visible is the deviation of real data from the standard distribution on the right.

we then compute a zScore for each region.

$$z = \frac{\log(\frac{n}{N}) - \log(\frac{m}{M}) - \mu}{\sqrt{2\sigma^2 + \frac{1}{n} + \frac{1}{m}}} \tag{7}$$

If there was no binding, the zScores should follow a standard distribution. Indeed, most of the zScores follow the log normal distribution, confirming our understanding of the noise distribution (Figure 3.2). Enriched regions deviate from the log normal distribution and locate in the right tail, meaning they are significantly different from the assumed background.

To rigorously take into account only regions which show significant differences from the background distribution, we calculate a false discovery rate, with summing over the top T regions which comply with $FDR \leqslant 0.1$.

$$FDR = \frac{1}{T} \sum_{i=1}^{T} P_{false}(w_i|D) \tag{8}$$

with P_{false} standardly defined as:

$$P_{false}(w_i|D, \sigma, \rho, \mu) = \frac{\rho P(n_i|m_i, \sigma, \mu, N, M)}{\rho P(n_i|m_i, \sigma, \mu, N, M) + (1-\rho)\frac{1}{W}} \tag{9}$$

Regions which fall above this cutoff in at least one sample are taken into account for further analysis. Regions which overlap or are adjacent to each other are merged to bigger regions.

increasing the resolution So far, our approach allows to detect regions of size $> 500\text{bp}$ potentially involved in regulatory actions. The binding sites for most transcription factors range from $7 - 21\text{bp}$, so we aim to inspect each of the regions in more detail to exactly determine the regulatory region and potential regulators.

We therefore construct a coverage profile for each region in basepair-resolution. As the lengths of the fragment are naturally not all similar, they don't locate all at the exact same position and will form gaussian-shaped coverage profiles. For the exact detection of individual active regulatory sites, we fit the coverage profile of each region with a gaussian mixture model:

$$L(\vec{C}|\vec{\mu}, \vec{\sigma}, \vec{\rho}) = \prod_{i=1}^{l} \Big[\sum_j \rho_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(i-\mu_j)^2}{2\sigma_j^2}) + \big(1 - \sum_j \rho_j\big)\frac{1}{l}\Big]^{C(i)} \quad (10)$$

$C(i)$ is the coverage at position i, with i is going over every position inside the region. j runs over all the gaussians in the model with parameters μ_j and σ_j and ρ_j being the fraction of reads in the gaussian. The number of gaussians we include in the model depends on the fragment size and the length of the region, but is at least 2. This likelihood is again fitted with expectation maximization and we obtain all the gaussians which were fitted to the coverage profile. Gaussians are merged, if their means are smaller than the sum of their standard-deviations. The definite peak region is then defined as $\mu - \sigma$ until $\mu + \sigma$.

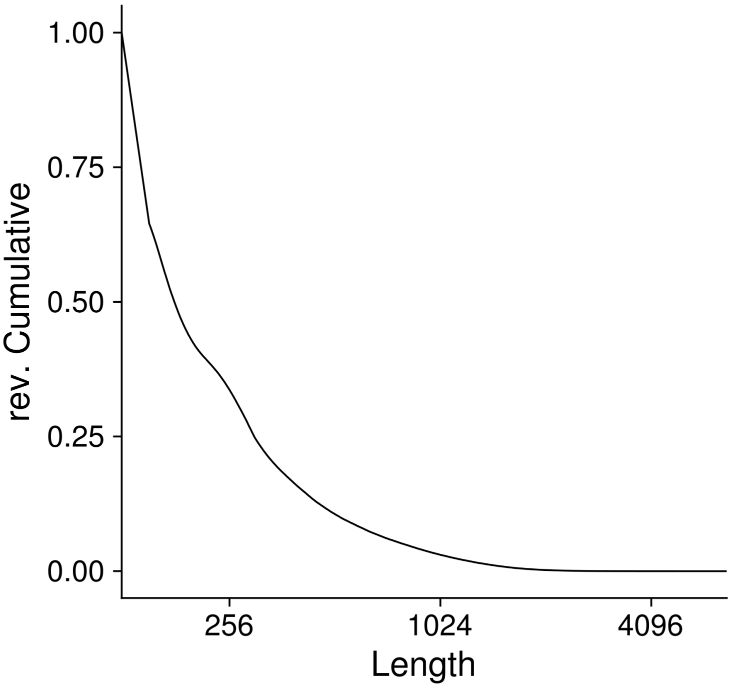


Figure 3.3: **Length of inferred CREs.** Reverse cumulative distribution of the lengths of all CREs. CREs were taken from dataset 2 in the CREMA paper (see chapter (4).

3.2 USING THE PEAKS TO INFER TRANSCRIPTION FACTOR ACTIVITIES

Having found significant peak regions for each sample, we want to go one step further and infer which key regulators may drive the observed system and changes across samples. For an application of this method, please also read chapter 4.

3.2.1 *Finding common cis-regulatory elements*

After calling the peaks for each sample, we need to create a common set of regulatory regions which captures all the peaks in each sample.

We retain all the peaks which were significantly enriched in at least one of the samples. Then we overlap all the peaks from the different samples and merge them if their centers are closer than 75bp by single-linkage clustering. In the case of merging, the resulting cis-regulatory element (CRE) thus spans all the centers of peaks belonging to it +- 75 bp up and downstream. We chose 75bp as cutoff as it has been shown that one nucleosome approximately spans 150bp on the DNA, which fits with the general width of a transcription factor binding peak as well. Indeed, looking at the length of all CREs, we find that most of them are in the range of 150-750bp. Only about 3% of the CREs are larger than 1000 (Figure 3.3)

The analysis of genomic data imposes challenges in the analysis and comparison across different samples: For ChIP-seq we aim to estimate the strength of the binding of the TF to the DNA from the height of the peak, for DNase and ATAC seq it is the fraction of cells that have increased accessibility at the same region. The inferrence of these parameters implies an accurate estimation of the read density and its fluctuations along the genome. We compute a signal S_{cs} for each CRE c in each sample s, which is normalized for background and inter-sample library size differences:

$$S_{cs} = \log\left(\frac{f_{cs}}{F_s} \cdot \tilde{F} + 1\right) - \log\left(\frac{b_{cs}}{B_s} \cdot \tilde{F} + 1\right) - \log\left(l_c / L_c\right) \tag{11}$$

where f_{cs} and b_{cs} is the read-count across CRE c in sample s. F_s and B_s are the library size of background and foreground samples s. $\tilde{F}$ denotes the median of library sizes in the foreground. l_c and L_c are the length of the CRE c in the foreground and the corresponding background CRE, these are similar across samples s. If a foreground region is shorter than 750bp, we choose the background region to be fixed to 750bp. As in the step before, we account for slow fluctuations and smooth the ideally uniform background distribution. Is the CRE bigger than 750bp we use the same size for foreground and background CRE.

To know which regulators play a role across samples, we perform an extensive motif search for transcription factor binding sites (TFBS) to construct our sitecount matrix.

3.2.2 *Predicting binding sites genome-wide*

To rigorously find TFBS for our library of ≈ 600 regulatory motifs in our set of CREs, we make use of the algorithm MotEvo [1], which has been developed earlier in our group. MotEvo is based on a hidden markov model and uses our extensive set of motif sequences w (weight matrices) and a pre-calculated prior probability π (which is equal to the probability of w to occur in a randomly chosen position on the input alignments) to model the input sequences [67]. Finally, MotEvo assigns, at each position of the

sequence and for each w, a posterior probability that a site for the corresponding w occurs at this position.
Note that the binding sites are non-overlapping as it is assumed that redundant motifs compete for the binding and don't increase the overall binding free energy. In our case, the prior probabilities p were chosen to optimize the fraction of explained variance in the MARA model on a large set of input samples.
To account for redundancy in our set of weight matrices, we fuse them given that

- one TF has multiple WMs which are nearly identical and/or
- WMs are not statistically distinguishable.

To find binding sites in all the significant CREs, we download all position weight matrices (PWM) for mouse (680) and human (684) genomes from the SwissRegulon database [24]. PWMs with similar binding pattern are clustered, so at the end we have a library of 503 (mouse) and 501 (human) PWMs [6]. We run the algorithm MotEvo in transcription factor binding site (TFBS) mode [1], across all the CREs. MotEvo reports a posterior probability N_{cm} for a motif m to have a binding site in CRE c. To fix the priors for each motif to occur, we rely on previously obtained values which have been optimized for the ISMARA run on an extensive set of samples taken from the fantom 5 project [16].

3.2.3 *Modeling the observed data in terms of regulators*

As stated previously, a multitude of CREs are changing their chromatin state across time or condition. Making conclusions about which CREs drive the difference is extremely challenging. What we really want to know is: Why does this happen and what are the key regulators driving these differences? Therefore we adapt an approach which was previously published by our group: ISMARA [49]. ISMARA models RNA-seq data in terms of binding sites in the promoters of expressed genes and an unknown activity. The activity for a motif m here denotes the amount the expression of a gene changes if one were to remove the binding site for one motif in this promoter. We want to make use of this model and introduce the concept of 'activity of a regulator' to our approach.

We now have information on the signal S_{cs} for each CRE c and each sample s, plus the probability that CRE c has a binding site for motif m, stored in matrix N_{cm} Further we assume, that the chromatin state across our samples and therefore the signal S_{cs} depends on the underlying motifs in the sequence of CRE c and an unknown activity $A_{ms}$of the factor binding to it. This approach is adapted from the previously in our group developed MARA model [49], which models RNA-seq data in terms of binding sites in the promoters of expressed genes. We can now fit the linear model:

$$S_{cs} = \sum_{m} N_{cm} \cdot A_{ms} + \tilde{c}_c + c_s + \text{noise} \tag{12}$$

to estimate the activity A_{ms} of each motif in each sample. The CRE- and sample-dependent constants $\tilde{c}_c$ and c_s are CRE and sample dependent constants which are estimated in the next step. The noise term accounts for measurement errors in S_{cs} plus biological fluctuations throughout the samples and the error in the model. In detail, MARA makes use of a bayesian procedure assuming the noise is gaussian distributed with variance σ^2 and equal for all CREs and samples. The likelihood of obtaining the signal table S_{cs} is then given by:

$$P(S|A) \propto \prod_{c,s} \frac{1}{\sigma} \exp\left[-\frac{\left(S_{cs} - \tilde{c}_c - c_s - \sum_m N_{cm}A_{ms}\right)^2}{2\sigma^2}\right] \tag{13}$$

We maximize this likelihood in terms of the CRE- and sample-dependent constants and replace them with the maximum likelihood estimations. Which gives:

$$P(S|A) \propto \sigma^{-CS} \exp\left[-\frac{\sum_{c,s}\left(S'_{cs} - \sum_m N'_{cm}A'_{ms}\right)^2}{2\sigma^2}\right] \tag{14}$$

with C the total number of CREs. Note that the table S'_{cs} is now centralized, such that mean of all the rows and the mean of all columns is zero. The sitecount values in N_{cm} is normalized in such way that the average count across all CREs is zero ($\sum_i N'_{cm} = 0$). This way we obtain activity values A' that the average across all sample for each motif is zero. To avoid the overfitting, each activity gets a gaussian distributed prior:

$$P(A'|\lambda, \sigma) \propto \prod_s \exp\left[-\frac{\lambda^2}{2\sigma^2}A'^2_{ms}\right] \tag{15}$$

With this, the posterior distribution becomes:

$$P(A|EN) \propto \exp\left[-\frac{\sum_{i,m}\left(\left(S'_{cs} - \sum_m N'_{cm}A'_{ms}\right)^2 + \lambda^2 \sum_m A'^2_{ms}\right)}{2\sigma^2}\right] \tag{16}$$

the parameter λ is fitted through a cross validation approach, using 80% of the CREs as train; and the remaining 20% as testset. The λ that minimizes the average square deviation of the expression levels in the test set versus those predicted by the fit of the train set is chosen as optimal λ. This posterior probability can be calculated via a ridge regression procedure, in this case SVD is used. The resulting activities are then sorted by their zScore:

$$z_m = \sqrt{\frac{1}{S}\sum_s^S \left(\frac{A'_{ms}}{\delta A'_{ms}}\right)^2} \tag{17}$$

The absolute signal $Y_{cs} = e^{S_{cs}}$,which is the raw normalized counts can be expressed as:

$$Y_{cs} \propto \prod_m e^{N'_{cm}A'_{ms}} \tag{18}$$

meaning that every reduction of a binding site for motif m decreases the signal for a CRE by $e^{A'_{ms}}$. In other words, the activity A'_{ms} corresponds to the amount by which the signal S'_{cs} would be reduced if a binding site for motif m in CRE c were to be removed.

Using the techniques described in (3.1-3.2), we are able to infer important key regulators shaping chromatin state or transcription factor binding to promoters and distal regulatory regions (see also Chapter 4).

3.3 ANALYSIS OF RNA-SEQ DATA TO PREDICT TRANSCRIPTION FACTOR ACTIVITY

For examples of analysis of RNA-seq data, please refer to Chapters 4 and 5.

3.3.1 *Preprocessing and mapping*

The treatment of RNA-seq data differs from what is described for ChIP, ATAC and DNase-seq data. RNA can be mapped in several ways: Once to the genome (e.g. STAR [19], Bowtie [20], BWA [21]) and once to the transcriptome (kallisto [15]). Tools mapping to the transcriptome use the novel technique pseudoalignment, which differs from the conventional practice, where the reads are matched to the genomic sequence. Using a de-Brujin graph, the sequences are split into k-mers and matched to the transcripts. This is a very fast approach as no real mapping procedure takes place. Another advantage of using kallisto is that no adapter-trimming or quality filtering is needed, as contaminated or adapter sequences would never match a transcript's sequence.

3.3.2 *ISMARA on the RNA-seq data*

ISMARA models genome-wide gene expression pattern in terms of predicted functional TFBSs in the respective gene's promoters. Promoter regions are either annotated or taken to be -500 +500 of the TSS defined by CAGE analysis (for more details see [6, 5]). In the model, the expression of promoter p in sample s, E_{ps} is assumed to follow a linear function of the binding sites N_{pm} in promoter p and motif m times an unknown activity A_{ms} of a motif binding site m in sample s.
When summing across all motifs this gives the core ISMARA equation:

$$E_{ps} = \sum_{m} N_{pm} \cdot A_{ms} + c_p + c_s \tag{19}$$

whereas c_p and c_s account for the promoter related basal expression and for the sample-dependent normalization constant, respectively.
The matrix N_{pm} contains information on the binding sites in each promoter and has been inferred using the algorithm *Motevo*. Equivalent to the method in previously described for ATAC, DNase and ChIP-seq, ISMARA also calculates, for each motif, a zScore which gives information about the significance of the motif activity change across the samples.

$$z_m = \sqrt{\frac{1}{S} \sum_{s}^{S} \frac{A'_{ms}}{\delta A'_{ms}}} \tag{20}$$

The term activity can be understood as: a change in activity represents the change in the expression E_{ps} of promoter p in sample s, when motif m in exactly in this promoter p would be removed. This means that the higher the activity, the higher the expression of genes having this motif in the promoter.

The target promoters p_m of motif m are promoters of genes that are expressed in the dataset and have a binding site for motif m in their promoter sequence. To estimate the importance of each promoter, ISMARA calculates a target score which denotes how worse the fit would be if the site for motif m in promoter p would be missing.

3.4 OTHER COMPUTATIONAL TOOLS

Especially in chapters 5 and 6, we make use of different computational strategies, which are explained in the following.

3.4.1 *SVD/PCA*

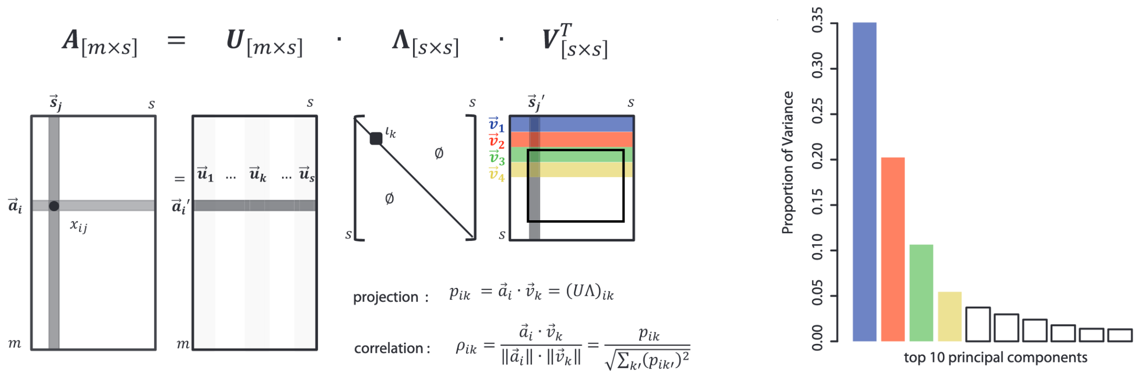

$A_{[m\times s]} = U_{[m\times s]} \cdot \Lambda_{[s\times s]} \cdot V^T_{[s\times s]}$
$\vec{s}_j$
s
$\vec{a}_i$
x_{ij}
m
$= \vec{u}_1 \ldots \vec{u}_k \ldots \vec{u}_s$
$\vec{a}_i{}'$
ι_k
∅
$\vec{v}_1$
$\vec{v}_2$
$\vec{v}_3$
$\vec{v}_4$
$\vec{s}_j{}'$
projection : $p_{ik} = \vec{a}_i \cdot \vec{v}_k = (U\Lambda)_{ik}$
correlation : $\rho_{ik} = \frac{\vec{a}_i \cdot \vec{v}_k}{\|\vec{a}_i\| \cdot \|\vec{v}_k\|} = \frac{p_{ik}}{\sqrt{\sum_{k'}(p_{ik'})^2}}$
Proportion of Variance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
top 10 principal components

Figure 3.4: **Schematic representation of SVD.** *Left panel* shows how we obtain the singular vectors $\vec{v}_k$ from the initial table A. *Right panel* shows the proportion of variance that the singular vectors explain. Taken from [10].

Single value decomposition (SVD) or principal component analysis (PCA) is a generalization of matrix diagonalization to non-square matrices and thus a powerful approach for dimensionality reduction. It is widely used in data science driven approaches to find common sets of variables in large datasets. This makes it especially useful for the application on gene expression data or - in our case - to motif activities. Generally speaking, every matrix A_{mxs} can be expressed in terms of two sets of singular vectors and singular values:

$$A_{mxs} = U_{mxs}\Lambda_{sxs}V_{sxs} \tag{21}$$

where V_{sxs} and U_{mxs} are the right and left singular vectors, it holds $V^TV = \mathbb{I}$ and $U^TU = \mathbb{I}$, with $\mathbb{I}$ the identity matrix. Λ_{sxs} is a diagonal matrix and contains the singular values (see Figure 3.4, *left panel*). The vectors $\vec{v}$ span a new orthogonal coordinate system, and every direction captures in descending order, the directions of variance thoughout the dataset. Each vector a_m with $\{a_m\}_s = A_{mxs}$ can now be written as a linear combination of the right singular vectors (see Figure 3.4, *left panel*).
Note that SVD, which is a more general approach than PCA, as we don't rely on the construction of the covariance matrix A^TA. PCA on A^TA would in this case yield the same eigenvectors $\vec{v}$:

$$A^TA = V\Lambda U^T \cdot U\Lambda V^T = V\Lambda^2V^T \tag{22}$$

which yields the so-called eigenvalues in Λ^2, which are the square of the singular vectors obtained above. The singular values give thus information on how much variance σ^2 was captured by the respective singular vectors.

$$\sigma_j^2 = \frac{\lambda_j^2}{\sum_{i=1}^m \lambda_i^2} \tag{23}$$

Generally, the follow-up analysis focuses on the singular vectors that capture most of the variance (Figure 3.4, *right panel*). To know which pattern of $\vec{a}_m$ are following which singular vector, we make use of a geometric approach which allows to identify vectors in a_m that follow the direction of the principal components.
As now, the vectors a_m are vectors in the principal component space, their projection on each of the axes (=principal components) indicates how strongly the vector a_m overlaps with the singular vector v_k. We an now calculate the projections according to $q_{mk} = \vec{a}_m \cdot \vec{v}_k$, and as, according to the SVD, $AV = U\Lambda$ it can be written in matrix multiplication $q_{mk} = (U\Lambda)_{mk}$.
Note that the projection captures also how strongly the vector a_m follows a singular vector, e.g. how 'long' it is. Still, short vectors can still correlate pretty well with the singular vectors. The correlation can be obtained by calculating: $p_{mk} = q_{mk}/\sqrt{\sum_k (q_{mk})^2}$. Note that the singular vectors v form a new orthogonal basis and are thus independent of each other, each capturing another pattern in the data. An application of SVD or PCA is described in Chapters 4 and 5.

3.4.2 *Machine learning algorithms*

Machine learning (ML) has evolved as a very powerful technique in all to treat the highly heterogeneous data coming from biological experiments. In Chapter 6, we make use of several machine learning techniques to determine maker genes in highly divergent human data.
ML algorithms in *supervised learning* build a mathematical model when subjected to training data, which then can be applied to new data to make predictions. Classification algorithms, a subset of supervised learning algorithms are trained to assign data-points to classes, for example, 'healthy' or 'sick'. Making use of kernel methods, a linear regression model can be performed in a non-linear problem (Figure 3.5): A kernel function maps the data in to an n dimensional space where the regression can be performed. For example, is the kernel function is $K(x, y) = \langle f(x), f(y) \rangle$, with x and y inputs and f the mapping to another space. The calculation of $\langle f(x), f(y) \rangle$ would normally requires the calculation of the dot product of $f(x)$ and $f(y)$. This is a rather computationally expensive step. Knowing the kernel function makes the computation much easier and saves computational time and resources. Recently, neural networks have been developed and used for classification tasks as well. In general a two-layer neural network could be considered as an SVM: the hidden layer transforms the input, the output is then classified. However, neural networks can contain several hidden lay-

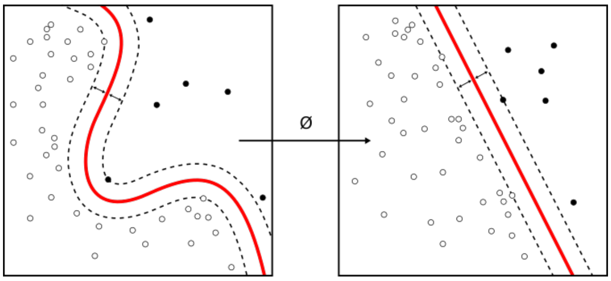


Figure 3.5: **Classification of nonlinear data.** Using kernel methods, non-linear data can be classified: Non-linear data is transformed in such a way that it is possible to separate it linearly.

ers. Each layer can apply different functions to the data and transform it. A function f for example maps the input x to the single hidden layer value $f(x) = h$. If the output is y and is calculated by $g(h) = y$, the output is connected to the input as $g(f(x))$. For an application of ML methods, please refer to Chapter 6.

BIBLIOGRAPHY

[1] Langmead, Ben, et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome biology 10.3 (2009): R25.

[2] Yan, Feng, et al. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis." Genome Biology 21.1 (2020): 22.

[3] Zhang, Yong, et al. "Model-based analysis of ChIP-Seq (MACS)." Genome biology 9.9 (2008): R137.

[4] Heinz, Sven, et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Molecular cell 38.4 (2010): 576-589.

[5] Xu, Shiliyang, et al. "Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells." Stem Cell Transcriptional Networks. Humana Press, New York, NY, 2014. 97-111.

[6] Kharchenko, Peter V., Michael Y. Tolstorukov, and Peter J. Park. "Design and analysis of ChIP-seq experiments for DNA-binding proteins." Nature biotechnology 26.12 (2008): 1351.

[7] Illumina. "Illumina adapter sequences." (2016), taken at 10/02/20 from `https://support.illumina.com/downloads/illumina-adapter-sequences-document-1000000002694.html`

[8] Balwierz, Piotr J., et al. "Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data." Genome biology 10.7 (2009): R79.

[9] Park, Peter J. "ChIP–seq: advantages and challenges of a maturing technology." Nature reviews genetics 10.10 (2009): 669-680.

[10] van Nimwegen, Erik. "Finding regulatory elements and regulatory motifs: a general probabilistic framework." BMC bioinformatics 8.S6 (2007): S4.

[11] Arnold, Phil, et al. "MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences." Bioinformatics 28.4 (2012): 487-494.

[12] Balwierz, Piotr J., et al. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs." Genome research 24.5 (2014): 869-884.

[13] Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics 26.1 (2010): 139-140.

[14] Luisier, Raphaelle, et al. "Computational modeling identifies key gene regulatory interactions underlying phenobarbital-mediated tumor promotion." Nucleic acids research 42.7 (2014): 4180-4195.

[15] Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology 34, 525-527(2016), doi:10.1038/nbt.3519

[16] Lizio, Marina, et al. "Gateways to the FANTOM5 promoter level mammalian expression atlas." Genome biology 16.1 (2015): 22.

[17] Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics 26.1 (2010): 139-140.

[18] Luisier, Raphaelle, et al. "Computational modeling identifies key gene regulatory interactions underlying phenobarbital-mediated tumor promotion." Nucleic acids research 42.7 (2014): 4180-4195.

[19] Dobin, Alexander, et al. "STAR: ultrafast universal RNA-seq aligner." Bioinformatics 29.1 (2013): 15-21.

[20] Langmead B, Salzberg S. "Fast gapped-read alignment with Bowtie 2". Nature Methods. 2012, 9:357-359.

[21] Li H. and Durbin R. (2009) "Fast and accurate short read alignment with Burrows-Wheeler Transform". Bioinformatics, 25:1754-60. 2009

[22] Martin, Marcel. "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet. journal 17.1 (2011): 10-12.

[23] Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`

[24] Pachkov, Mikhail, et al. "SwissRegulon: a database of genome-wide annotations of regulatory sites." Nucleic acids research 35.suppl_1 (2007): D127-D131.

4

CREMA: AUTOMATED MODELING OF GENOME-WIDE CHROMATIN STATE IN TERMS OF LOCAL CONSTELLATIONS OF REGULATORY SITES

ABSTRACT

The variety of cells is partly governed by the chromatin state. It is determined by the DNA accessibility, the presence of a binding site for specific transcription factors (TFs) and the presence of co-activators, substrates and mediators. The ongoing development and usage of high throughput sequencing techniques have opened a wholly new perspective to quantitatively characterize the chromatin state which may extend our knowledge about how the genome-wide regulatory structure across cell types, conditions or tissues is orchestrated. Here we present cis regulatory element motif activity analysis (CREMA). By combining previously published tools CRUNCH and ISMARA, CREMA analyzes high throughput genomic data quantifying TF binding (ChIP-seq) and DNA accessibility (DNase-seq, ATAC-seq) to rigorously model genome-wide chromatin state in terms of local constellations of transcription factor binding sites. Using a sophisticated model, CREMA infers key transcription factors which drive the observed changes in chromatin state and makes reliable predictions about their implication in the regulatory processes.

Here, our tool applied to a selection of published human and mouse data, infers important key regulators in circadian regulation, embryonic development and stem cell differentiation in agreement with the current state of knowledge.

We find that the chromatin state is in general more variable at distal regions and predict novel TFs implicated in circadian oscillation in murine liver and in human stem cell fate. To get an insight into the functional roles of our TFs, we predict target genes and functional terms and pathways that are being regulated by the TF. Our algorithm runs as a fully automated pipeline on a dedicated web-server and is designed to process raw data coming directly from the sequencing facilities.

RUNNING TITLE

Modeling of genome-wide chromatin state

KEYWORDS

Gene regulation, ATAC-seq, DNase-seq, ChIP-seq, transcription factor binding, regulatory motifs

INTRODUCTION

During the last decade, the amount of researchers using high-throughput measurements has increased drastically, as it allows a genome-wide quantification of chromatin state or transcription factor binding. However, making sense of the generated data requires sophisticated computational analysis. A common problem here is the high dimensionality of the data, as the number of DNA loci which are subject to chromatin state changes can be very large. Dealing with data of that size requires not only computational power and storage, but also robust models to analyze the data in a way that provides meaningful insights. As we know from the biological side, compared to the number of regions that are regulated, there are only few transcription factors that drive the observed chromatin state. Our tool combines ChIP-seq analysis [9] with a previously developed model for inferring key regulators for gene expression from RNA-seq data [49] to robustly quantify and explain changes in chromatin state by the activity of regulators. Our tool is designed to be easily accessible for any researcher who deals with high throughput data.

Cell identity is defined and stabilized through a complex regulatory network. Although ultimately this regulatory network is encoded in the constellations of regulatory sites in the whole genome, we still understand quite little about how this regulatory code is translated into regulatory circuitry that defines and stabilizes cell identity. Besides promoters located close to their target genes, a large part of the genetic regulation is additionally guided by distal regulatory modules (enhancers) that, in mammals, can occur tens to hundreds of kilobases away from the transcription start sites of genes that they regulate.

Additionally, the expression of genes is regulated by the chromatin state: DNA is compacted in several layers into chromosomes, on the first level it is wrapped around 8 core histones forming a nucleosome [43]. This packaging is intrinsically repressive as it prevents regulatory factors from binding and thus unwanted transcriptional action [41]. However, histone occupation is not static: By a dynamic unwrapping/wrapping of DNA, it defines and changes the chromatin structure and enables or disables transcription. Histones can be modified postranslationally in multiple ways which affects chromatin state and gene expression, making them powerful predictors of DNA accessibility and active regulatory elements along the whole genome [12, 17].

Hence, the minimum requirement for a gene to be expressed in a certain condition is that the associated promoter: 1) has a binding site for a condition-specific TF and eventual co-regulators, 2) be accessible for its binding and 3) be temporarily bound by the TF which recruits the transcription machinery. In case of a distal regulatory region, TFs additionally have to compete with nucleosomes for the binding and the region has to interact with the promoter. Depending on the gene and condition, supplemental fine-tuning processes such as the interplay of multiple regions or transcriptional regulators be involved in this regulation [55].

Thus, from a combination of both the accessibility and underlying binding sites of transcription factors, it is possible to determine which regulatory regions are potentially active, even if they occur in intergenic regions. In fact, it has been shown previously, that the state of a cell can be precisely predicted by analysis of its chromatin state [53], while the chromatin state itself depends on specific sequence characteristics and binding sites of transcription factors [2]. In this work, we provide a rigorous method to analyze the regulatory key players involved in shaping the chromatin state genome-wide.

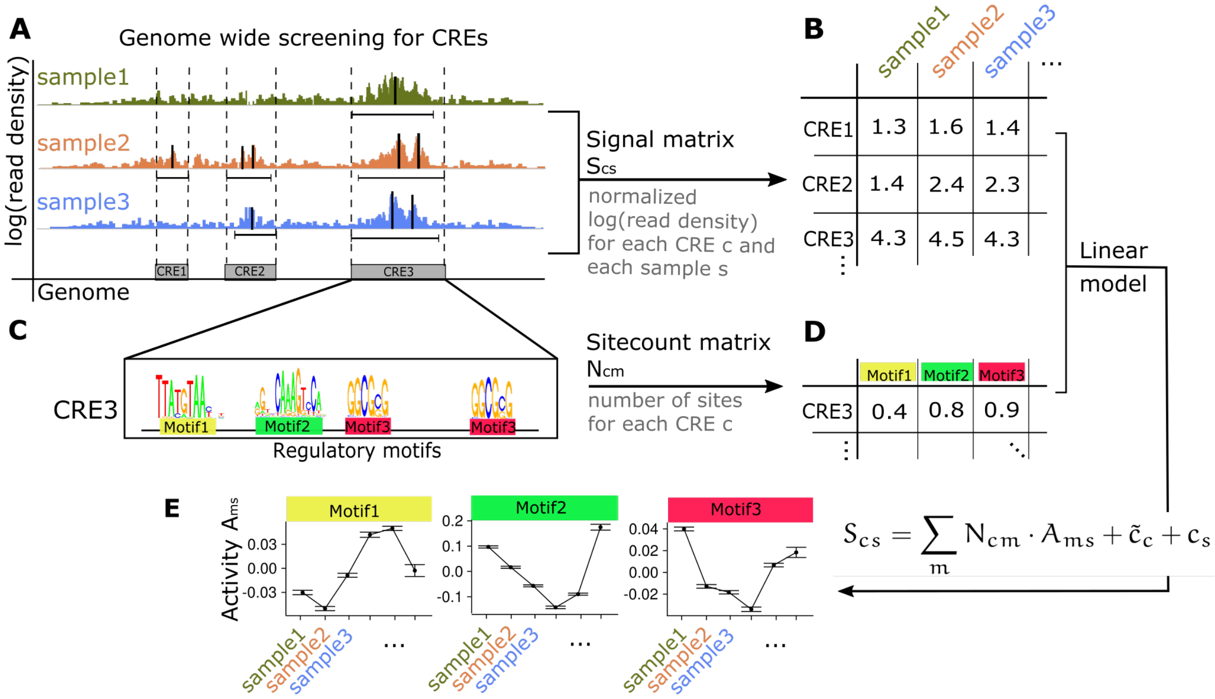

A
Genome wide screening for CREs
log(read density)
sample1
sample2
sample3
Genome
CRE1
CRE2
CRE3
Signal matrix
Scs
normalized log(read density) for each CRE c and each sample s
B
sample1 sample2 sample3 ...
CRE1 1.3 1.6 1.4
CRE2 1.4 2.4 2.3
CRE3 4.3 4.5 4.3
Linear model
C
CRE3
Motif1 Motif2 Motif3 Motif3
Regulatory motifs
Sitecount matrix
Ncm
number of sites for each CRE c
D
Motif1 Motif2 Motif3
CRE3 0.4 0.8 0.9
E
Activity Ams
Motif1 Motif2 Motif3
sample1 sample2 sample3 ...
$S_{cs} = \sum_m N_{cm} \cdot A_{ms} + \tilde{c}_c + c_s$

Figure 4.1: **From raw data to genome-wide identification of cis regulatory elements (CREs) and motif activities.** Analysis and modeling steps: A) Detection of significant peaks genome-wide is performed for each sample independently (colored read density profiles), followed by the construction of a universal set of cis regulatory elements (CREs) (grey boxes) by merging close peaks from all samples. Black vertical lines represent the centers of peaks used to construct the common CRE set. B) Normalized log read density for each CRE c and each sample s is stored in a signal matrix S_{cs}. C) We computationally predict transcription factor binding sites for a large collection of regulatory motifs in all CREs. This generates the sitecount matrix N_{cm} which contains the number of sites for each motif m in each CRE c. E) Using the matrices from B) and D) we fit a linear model to explain the observed signal across samples S_{cs} in terms of regulatory sites N_{cm} and an unknown motif activity $A_{m}s$. $\tilde{c}_c$ and c_s are CRE and sample-dependent constants.

RESULTS

CREMA: cis regulatory element motif activity analysis

We developed CREMA, a combination of two approaches which have been published by our group earlier: CRUNCH [9] and ISMARA (FANTOM Consortium and Riken Omics Science Center 2009, [6]). ISMARA models the genome-wide mRNA expression level changes across samples in terms of transcription factor binding sites in (TFBS) promoters for around 600 regulatory motifs. This is highly successful in inferring which regulators are key regulators within a given system, how these regulators change activity across samples, and what genes and pathways are targeted by each of the regulators. However, much of the regulation is not directly controlled by TFBSs at promoters, but rather by large numbers of cis regulatory elements (CREs) which include, in addition to promoters, distal regulatory elements. Distal CREs are generally more specific than promoters, likely because they can be activated in a condition-dependent manner by particular transcription factors complexes. In addition, CREs are more numerous than promoters, meaning that the number of active CREs in one condition exceeds the number of active promoters. Consequently, the genome-wide signals at CREs contain more information about the regulatory states of cells within a given condition than the mRNA expression levels. To infer regulators that are driving the observed changes in chromatin state at CREs genome-wide, we developed CREMA. CREMA processes ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) and DNAse-seq (DNase I hypersensitive sites sequencing), which assess accessibility of regions on the DNA, as well as ChIP-seq (Chromatin ImmunoPrecipitation DNA-Sequencing) for e.g. a specific TF or histone modifications. CREMA is implemented as an automated pipeline on our web-server at `https://crema.scicore.unibas.ch`. Users can upload raw sequencing data (.fastq) to CREMA. The information on the samples, such as condition and grouping of replicates has to be provided in a simple table.

From raw data to cis regulatory elements

Calling the peaks is done by following the steps used in CRUNCH [9]: After thorough preprocessing, the data is modeled using multiple sophisticated statistical models and reports significant peaks [9]. Having determined the significant peaks (Figure 4.1A, colored tracks), for each sample individually, we define a common set of peaks across all samples: If the centers of individual peaks in a sample and/or across samples are closer than 150bp we merge the peak to one cis regulatory element (CRE) (Figure 4.1A gray boxes). The typical lengths of the CREs after merging range, depending on the dataset, from 150bp up to 5600bp, whereas 90% of the CREs are smaller than 500bp and 97% are smaller than 1000bp in length (Figure S1).
We then calculate the log read density for each CRE and each sample. To make the signal comparable across samples, we construct a common CRE set including peaks of all samples and construct the signal matrix S_{cs} which contains the normalized log read density (peak height) at CRE c in sample s (Methods).

PREDICTION OF TRANSCRIPTION FACTOR ACTIVITIES We then assume that the resulting peak height at any CRE, stored in the signal matrix S_{cs}, depends on the TF binding sites sites for motif m in CRE c and a yet unknown motif activity A_{ms}. To infer TFBS, we make use of the curated library of regulatory motifs which are used also in ISMARA [6]. With MotEvo[1], we computationally predict TFBS for each CRE c (=sitecount matrix N_{cm}) (Fig 4.1,C and D).
Using the matrices S_{cs} and N_{cm}, we model the observed signal matrix using the core linear model of ISMARA to explain the peak height at each CRE c in sample s with a linear combination of each motif N_{cm} and an unknown motif activity A_{ms} (Figure 4.1E).
The inferred motif activities A_{ms} in each sample s for each motif m are stored in the activity matrix A_{ms}. For each motif activity we additionally report the associated uncertainties δA_{ms} (error bars on motif activities). Looking at the data in terms of motif activity reduces the high dimensionality of the data, as the signal in all regions can now be explained by the activities of ≈600 motifs of our library. We then calculate, for each motif, a zScore which summarizes the importance of the motif for explaining the variation in peak height across the samples. While the regulation on DNA level is a complicated, not yet understood interplay of various regulatory variables, we don't

aim to present an accurate snapshot of the chromatin landscape and transcription factor occupancy. Depending on the dataset, our model explains only approximately 10%-20% of the variance in the dataset, yet enough to make reliable predictions of important motifs driving the observed changes in chromatin.
To give further insight into the function of each motif m, we provide an extensive list of "targets". Those are CREs with a TFBS for motif m. To quantify the evidence of a CRE c to be a target of motif m, we compute the target score ζ_{mc} as the log-likelihood ratio of the full CREMA model versus the model with only the sites for motif m removed from CRE c. The target score depicts how much the activity of the motif contributes to the peak height at the target CREs (see Methods). As a large fraction of the CREs locates at distal regions, it is not always trivial to associate a gene definitely. Nevertheless, to obtain meaningful results for functional gene enrichment analysis, we define the gene targeting score $\xi_m(G)$ (see Methods). Here, we first calculate a probability P_{cg} for each CRE c to regulate nearby genes based on the distance of the CRE to the promoters of the nearby genes (= CRE/gene association probability). The total log-likelihood (=gene targeting score) for a motif m to regulate gene g is then the sum over all CREs c of the product P_{cg} and the target score ζ_{mc}. We then run several types of functional analysis considering the gene targeting score, Gene Ontology (GO), Reactome and Kyoto Encyclopedia of Genes and Genomes (KEGG) on the set of genes for each motif. After the analysis finished, we provide a detailed report including information on all motifs (Figure S2-S10).

To test the performance of and versatility of CREMA, we analyzed three different public datasets, including ChIP-, ATAC- and DNase-seq in mouse and human, and validated CREMAs predictions with important established theories and hypotheses on genomic regulation. All datasets analysed in this study are available online at `https://crema.unibas.ch` (see Methods). As outlined in the introduction, the DNA is tightly packed, only a small fraction of the genome is open and thus amenable for the binding of a regulatory protein. Straight-forward methods to get a snapshot of the chromatin state and determine accessible regions are DNase-seq or ATAC-seq. DNase-seq makes use of the enzyme endonuclease to cut accessible, non-nucleosomal DNA, whereas ATAC-seq relies on a transposase enzyme which has intrinsic high affinity to the DNA and cuts out exposed parts of the DNA. Both techniques thus determine accessible regions, which are then selected for sequencing. An accumulation of fragments at a genomic locus thus denotes higher accessibility.

Dataset 1: Preserved circadian cycling of transcription factors in murine liver after > 50h in darkness

Circadian (or near 24-h) oscillations are generated by transcriptional – translational feedback loops involving core transcriptional activators (Bmal1 and Clock) and repressors (Per and Cry) which control the 24-h rhythmicity of gene expression. The liver clock is mainly synchronized by the master circadian clock located in the suprachiasmatic nucleus of the hypothalamus. However, the liver clock can be uncoupled from the light-regulated "master clock", when food access is limited to a temporal window during the day [25]. Studies investigating the liver clock usually make use of this by setting the mice into constant darkness to prevent any possible influence of the liver clock by light. In the chosen study, the authors provide food ad libitum, so we expect

the liver clock to function in a normal circadian manner.
We downloaded a public dataset from the ENCODE consortium [21, 20] (see Table S1 for accessions, datasets are also linked on crema.unibas.ch). DNase-seq has been performed on the liver of adult mice exposed to constant darkness for 70h. Here, mice were first entrained to a standard 24h light/dark cycle (we refer to the timepoints as circadian time CT0-CT24, where CT0 is, according to common notation, equivalent to 6am), the active phase starting at CT12 until CT24 for night-active mice. From 50h after switching the light-dark cycle to constant darkness, the liver was taken every 4h to perform DNase-seq (see Methods). We run CREMA on the whole dataset to predict motifs shaping the chromatin state during one 24h cycle.

Sustained cycling of motif activity in liver

We observe a strong cycling pattern of the motif activity with 24h periodicity (Figure 4.2A). Notably, the most significant motifs correspond to known drivers of circadian regulation (Arntl, Nfil3 and Egr1) or are strongly regulated in a circadian fashion (Hnf1b):
The activity of Arntl, also known as Bmal1 is strongly cycling, with lower activity during the active phase. Almost 15% of the oscillating transcriptome is regulated by the Clock:Bmal1 complex. A study performed in 2018 suggests even a chromatin-shaping role for Clock:Bmal1 which primes tissue and condition dependent enhancers [63].
Nfil3 is a known circadian transcription factor [37] and was found to control lipid metabolism in a circadian way [71]. A predicted target of Nfil3 is Per2, which is an important gene in whole body circadian rhythm.
A study showed that Egr1 is cycling strongly in hepatocytes, regulating the transcription of important clock genes (for instance Per2, Bmal2 and Rev-erbα/β)[59].
Hnf1b was identified as circadianly cycling in a study investigating double knockout (Rev-erbα and Rev-erbβ) in mouse embryonic stem cells. Hnf1β cycling was diminished upon the knockout, suggesting it to be regulated in a circadian manner by the Rev-erb factors [33].
CREMA further predicts several motifs whose activity oscillates in a circadian manner as well. Those novel motifs still yield a high zScore and a strong oscillating pattern but haven't yet been associated circadian regulation or oscillation.
Irf2 shows oscillation with higher activity during the inactive phase. Members of the Irf family have been found to be important for the elicitation of recognition pattern – and thus for the innate immune response. CREMA predicts GO terms related to neuronal cells to be regulated by Irf2 (e.g. "motor neuron axon guidance" (GO:0008045)), but also immune-related pathways, e.g. "negative regulation of interferon-alpha production" (GO:0032687) and "protection from natural killer cell mediated cytotoxicity" (GO:0042270).
The Nfia activity follows strongly a 24h cycle. This TF has been suggested to play an important role in mammalian embryonic development [18], but has not been associated with circadian rhythm. In our analysis, it seems to be involved in signaling and transport, our GO analysis shows "gluthamine transport" (GO:0006868) and "regulation of Rho protein signal transduction" (GO:0051895).
Etv1, is very similar to the Gabpa motif and was associated with organ-size control and tumorigenesis by regulating the transcription of the transcriptional coactivator Yes-associated protein (Yap) [72], which is associated with cell proliferation [68]. We

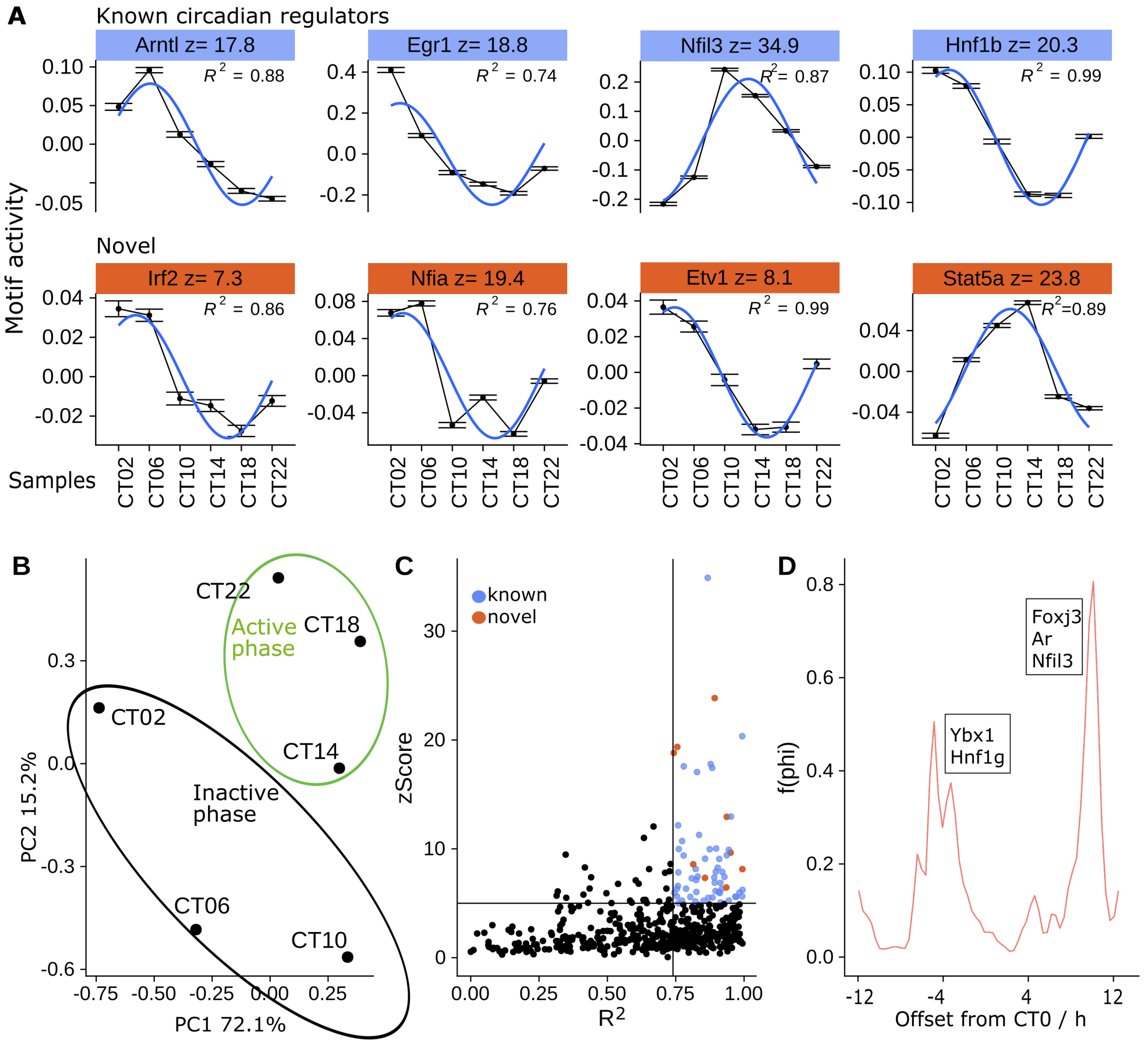


Figure 4.2: **Sustained oscillation of previously known and prediction of novel potential circadian regulators in murine liver after exposure to darkness for multiple days**. A) *upper panel*: Highly active motifs which were previously identified as circadian regulators (Arntl, Egr1 and Nfil3) or show strong circadian behavior (Hnf1b); *lower panel*: Highly active motifs which are predicted to oscillate with 24h period, but have not been associated to act as circadian regulation or being circadianly. regulated before. B) PCA on the activity matrix shows a clear distinction of the inactive and active phases of the mice. Values on the axes represent the fraction of explained variance of the respective component. C) Correlation of R^2 correlation coefficient of the fit of a sinusoidal curve to the activity pattern versus the zScore of our model. Known and novel circadian regulators (including those in A) are denoted as blue (known) and orange (novel) dots. D) Phase distribution $f(\phi)$, showing which phases of motif activity occur most frequently in the dataset. Phase shifts (offsets) are calculated for a sequence of $\phi \in [-\pi, \pi]$ (oscillation period 24h). An offset of 0h corresponds to CT0=6am. Motifs contributing most to the two local maxima of $f(\phi)$ are indicated in the boxes.

observe for Etv1 a significant pattern switching from negative activity to positive activity at around CT12. For Etv1, CREMA indeed infers GO categories related to proliferation and differentiation, for example "megakaryocyte development" (GO:0035855), and "aortic smooth muscle cell differentiation" (GO:0035887). From this we are not

confident enough to define a specific function for Etv1 in liver.
Stat5a was found to be involved in shaping the chromatin landscape in liver, and being essential for liver sexual dimorphism [19]. Stat5a has – to our knowledge – not been investigated with regard to circadian cycling. However, a study included Stat5a into their analysis when looking at the regulation of the liver clock by Bmal/Clock, but without any concrete conclusions related to Stat5a [62]. Our predictions for functional categories include "positive regulation of production of miRNAs involved in gene silencing by miRNA" (GO:1903800) and metabolic terms, such as "negative regulation of glycogen biosynthetic process" (GO:0045719) and "arginine biosynthetic process" (GO:0006526). One of the known circadian regulators, Bhlhe40, shows very little changes in our hands with a zScore of 1.5, however, it was also found to be weakly expressed in liver, which could explain the fact that we don't find any accessible binding sites with this motif [16]. However, we find Bhlhe41 to be significantly cycling (zScore 8.8), which has a similar motif. It is thus likely that we may have a slightly wrong motif-TF association in this case.

Distinction of the circadian timepoints through motif activity

We performed principal component analysis (PCA) on the motif activity matrix. The first two principal components (PC1 and PC2) show exact separation of the timepoints according to the inactive and active phase of the mice - indicating that the light-dark cycle is sustained (Figure 4.2B). Interestingly, the timepoints taken during the active phase seem to lay closer together in the motif activity space than the ones taken from the active phase. This suggests that in general, changes in transcription factor activity are bigger during the inactive phase than during the active phase.

Correlation of oscillation and zScore

To estimate the strength of circadian oscillation, we fitted a harmonic function of the form $y(t) = A_0 \cdot \sin(2\pi/24 \cdot t + \phi)$, assuming a 24h periodicity, starting at CT0, with A_0 the amplitude and ϕ the phase (see Methods). The R^2 is the fraction of variance explained by our fit $y(t)$ compared to the total variance in the underlying activity pattern and quantifies the goodness of the fit. In (Figure 4.2C). we plot the overall zScore for each motif versus the R^2 value. Higher zScores occur thus concomitantly with higher R^2 values. We find 15 known (blue dots) and 7 novel (orange dots) circadian motifs (those may be regulators themselves or regulated in a circadian manner) above a threshold of $R^2 > 0.74$ and $zScore > 5$ (see Table S2), including the motifs mentioned earlier.

Phase shift of motifs and CREs

We found the most common motif activity pattern showing lower activity in the inactive phase. We thus investigated how the phase ϕ of our function $y(t)$ changes across motifs. To take into account the significance of the motif activity pattern too, we we calculate a distribution $f(\phi)$. That is, for a sequence of $\phi \in [-\pi, \pi]$ we calculate a weight w_m for each motif indicating how much its phase deviates from ϕ (phaseshift). Additionally, we weight each w_m by the respective zScore of the motif m. Thus, motifs whose phase ϕ_m is close to ϕ and have a high zScore will have high values and increase $f(\phi)$ (see Methods). We refer to the phaseshift as offset in hours, as we imply a period of 24h. An offset of 0h refers to the cycle starting at 6am=CT0 and the activity is increased during the first three timepoints, whereas an offset of around 10h denotes higher activity in the last 3 timepoints, as the cycle starts at 4pm. What we observe is that our function $f(\phi)$ peaks around an offset $+10h$ and $-4h$. This confirms the previous observation that the majority of the oscillating motifs shows a low activity during the active phase. Examples of motifs having a offset of $-4h$ are indicated in Figure 4.2D (boxes). Ybx1 (Ybx1_Nfya_Nfyb_Nfyc_Cebpz) and Hnf1γ (but also other Hnf factors, Hnf1β, Hnf4α) activities follow this pattern. Further we find motifs with an offset of $+10h$, meaning their activity is increased during the resting phase. This holds for example Nfil3, Foxj3 and Ar. Among the predicted functional GO categories that CREMA predicts for Ar are "negative regulation of peptidase activity" (GO:0010466) and "progesterone biosynthetic process" (GO:0006701), which would suggest that this transcription factor is involved in metabolic processes that are slowed down during the inactive phase. Motif activity plots of Nfya (Ybx), Ar and Hnf4γ are displayed in S11.

Dataset 2: Regulation of the development of organs in mice

To demonstrate the performance of our approach in the analysis of ATAC-seq data, we chose ENCODE data [21, 20] (see Table S1 for accessions), and selected ATAC-seq data of heart, liver, lung, hindbrain, forebrain, intestine, kidney, limb, neural tube and embryonic facial prominence (face) during development in mouse. As an additional dataset we downloaded RNA sequencing data (see Methods). [21, 20] (see Table S1 for accessions) of the same mouse-line at the same timepoints. For this dataset we wanted to answer three questions: 1) Are there motifs which are specific for one or multiple tissues? 2) How do the motif activities change with time? 3) How correlated are the motif activities derived from ATAC-seq data with RNA-seq derived activities?

Motifs responsible for tissue specificity

To answer the previously defined questions we looked first at the pattern across all tissues and timepoints. We manually selected 6 of the top 25 motifs which have already been associated with functions in the respective tissue and in embryonic development and plotted their activity pattern for all tissues (Figure 4.3A). Some motifs show increased activity in a subset of tissues compared to other tissues across all timepoints (e.g. Tal1), while others in- or decrease their activity with time (e.g. Nfia, Cebpb). Specifically, we find the following motifs:
<u>Tal1</u> shows higher activity in liver and heart compared to the other tissues. Additionally, it's activity decreases with time in both tissues. This is consistent with literature: Tal1 has been found in all hematopoietic organs during development, including liver [66]. It is also mandatory for heart development: A study with Tal1 knockout in zebrafish showed severe defects in the heart development [14]. In our plot, the high activity of Tal1 in liver is more than twice as high as the activity changes in heart, but seen independently of liver, the activity of Tal1 in heart is still significantly elevated compared to the remaining tissues. CREMA predicts Tab3, a TGF-β activated kinase, to be regulated by Tal1, which was indeed found to be involved in the TGF-β signaling pathway previously [60].
<u>Nfia</u> increases its activity across all the timepoints in all tissues except for in liver, kidney and intestine. Members of the Nfi transcription factor family were associated to be

important during mouse embryogenesis already 20 years earlier [18]. Especially Nfia has been found to control the transition from neurogenesis to gliogenesis in the central nervous system [47, 22]. This is consistent with our findings, where Nfia increases its activity with time in brain-associated tissues too.

Hnf4a and Hnf1b are known liver-specific factors [27, 48] and show higher activity only in liver, kidney and intestine compared to the other tissues. According to our KEGG pathway analysis, Hnf4a and Hnf1b are involved in lipid metabolism and genes associated to its targets play a role in transcription factor networks, such as the PI3K/AKT and Foxa2/Foxa3 networks.

Rfx3, Rfx1 and Rfx4 have been - amongst other factors in the Rfx family - found to play an important role in axon migration [45, 58]. CREMA infers higher activity in the brain-associated tissues hindbrain, forebrain and neural tube than in other tissues. Notably, the activity increases with time. Among the predicted GO pathways that Rfx3, Rfx1 and Rfx4 are predicted to be regulating are "fasciculation of sensory neuron" (GO:0097155) and "neuronal channel clustering" (GO:0045161).

Mef2b is increasing its activity almost exclusively in heart – with higher overall activity than other tissues. Being considered as one of the core cardiac transcription factors which is involved, among a number of other cellular programs, in direct reprogramming and genome-wide cardiomyocyte gene regulation [23], our findings make sense. Table 1 shows an excerpt of the CREs targeted by Mef2b: genomic location of the CRE, the inferred target score and the regulated genes, as well as the CRE/gene association probability P_{cg} and the distance to the promoter. For Mef2b we find – in concordance with current knowledge – muscle-related genes such as Titin (Ttn) [31] and Prkaa2, which is a catalytic subunit of the energy sensor protein kinase AMPK that plays a key role in regulating cellular energy metabolism especially relevant in muscle [39]. In table 2, we show the top GO categories which were inferred for the set of genes associated to the targets of Mef2b. Here, too, we find muscle related terms which validates our findings.

Cebpb is systematically more active in liver and lung, and it increases activity in both tissues. The family of Cebpb transcription factors is important for pulmonary gene expression and is implicated in several lung-associated diseases such as asthma, pulmonary fibrosis and COPD. Specifically Cebpb was found to support proliferation and to regulate inflammatory and innate immunity gene expression [52]. Indeed, the top predicted target of Cebpb is Interleukin-3, which a cytokine, activated by T-cells, and plays a role in several immunopathologies [13].

Tissue specificity and time dependence of motif activities

A principal component analysis of the activity matrix across all tissues and timepoints shows that the highest variance in the data is due to the difference between tissues. The fraction of explained variances for the first two PCs is 51.7% and 20.9%, respectively (see also Figure S12C). Liver samples have very specific activity patterns which distinguish them from other samples, the tissues which locate closest to liver in the PCA space are intestine and kidney. Brain-associated tissues cluster together, as well as limb and face. We suggest that this may be because these tissues are made up of bone and muscle tissue (Figure 4.3B). We chose top 9 motifs according to CREMAs zScore and plotted the projection of their activity profiles into the PCA space (Rfx comprises two motifs, Rfx3_Rfx1_Rfx4 and Rfx2_7). Interestingly, besides expected TFs such as

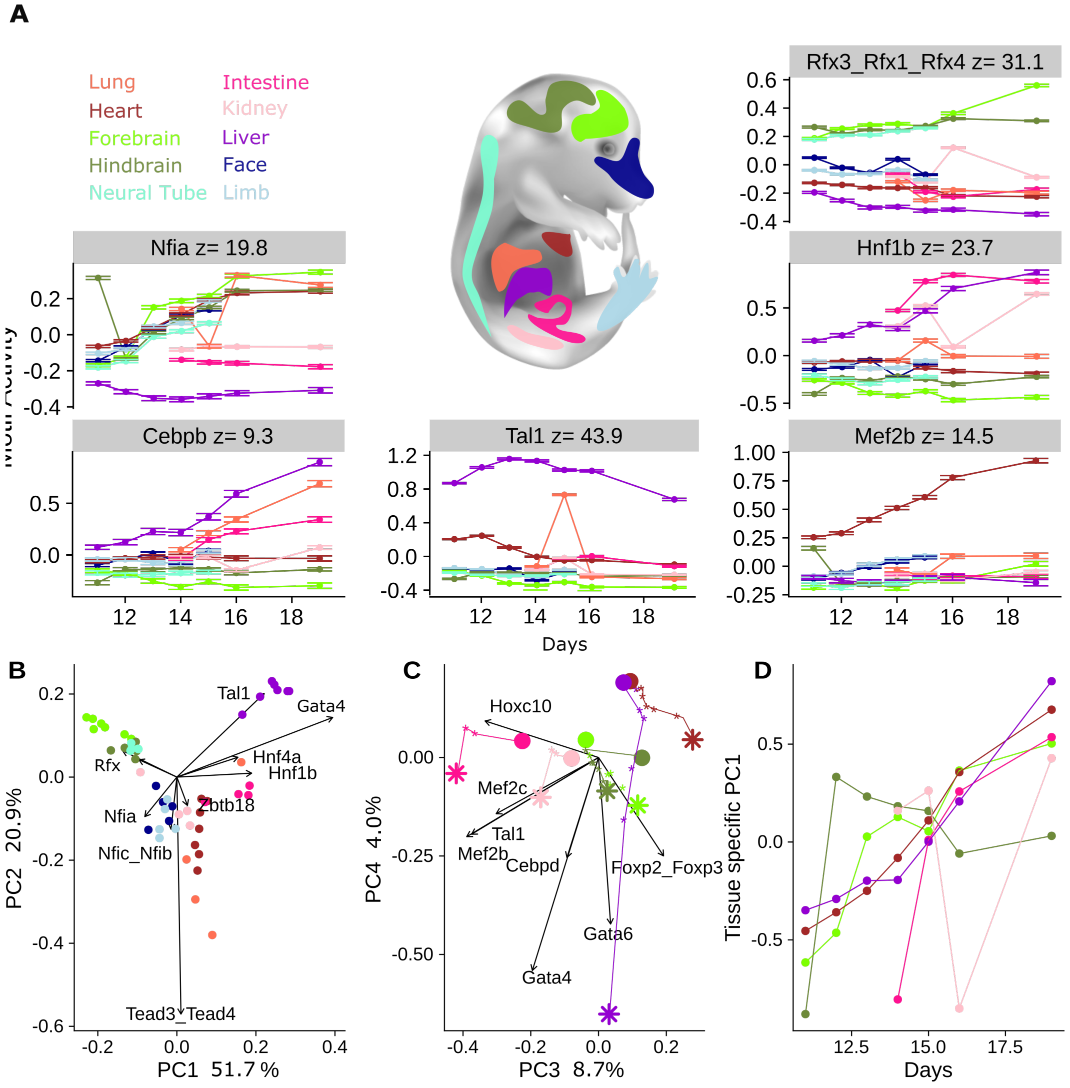


Figure 4.3: **Motif activities are tissue and time dependent and the variability across tissues increases with time**. A) Motifs selected from the top 30 motifs to represent known TFs and to show the variety of activity patterns in a single or multiple tissues. Some motifs are changing their activity in multiple tissues, whereas others are more tissue-specific. B) The first two PCA components show that most of the variation in DNA accessibility patterns are associated with differences between tissues rather than developmental time. Plotted vectors represent the projection of activity profiles of the top 9 significant motifs according to CREMAs zScore into PCA space. Rfx represents Rfx3_Rfx1_Rfx4 and Rfx2_Rfx7. C) Variation in motif activity along the third and fourth PCs is associated with time-dependence. Big dots: activities at 11days (or 14days for intestine, kidney and neural tube). Small dots: activities at intermediate timepoints. Stars: activities at birth. Several tissues which represent the overall behavior have been selected to simplify the plot. Plotted vectors represent the 2 motifs with highest projection score on PC3 and 3 motifs with highest projection on PC4. Several tissues have been preselected to simplify the plot. All samples are shown in S12A. *continued on next page*

the Hnf factors, Gata3 and Gata4 (Figure 4.3B,C) distinguish liver, kidney and intestine from other tissues in the PCA space. Gata factors have been found to be involved in liver development earlier [74], as well as in intestine [11]. Tissues taken from heart,

Figure 4.3: *continued* D) Despite of the specificity for some tissues (groups), around 50% of the variance of motif activities in each tissue can be explained by the first principal component for a PCA across activities in this tissue (tissue specific PC1) which is an increasing or decreasing pattern (Figure S12D). This holds for all the tissues. Several tissues have been selected to simplify the plot, all samples are shown in Figure S12B.

target CREs	target score	associated gene	Gene Info	distance to promoter	CRE/gene ass. prob.
chr8:122451195-22451377	70.65	Gm20735	predicted gene, 20735	1	0.94
chr2:76806390-76806720	67.76	Ttn	titin	19987	0.22
chr4:105100638-105100793	67.56	Prkaa2	protein kinase	9175	0.27
chr15:103355478-103355629	54.81	Itga5	fibronectin receptor	1098	0.34
chr2:91117863-91118209	50.42	Mybpc3	myosin binding protein	108	0.94
ch14:55003891-55004289	47.30	Myh7	myosin, cardiac muscle	9464	0.07

Table 1: **Target CREs and their associated genes for Mef2b.** We show the location on the genome, the target score, the associated gene and more extensive gene info. The distance to the promoter and the CRE/gene association probability (P_{cg}, see Methods) is shown in the last two columns.

kidney or face cluster near the origin of the PC1-PC2-space, but have different values in the third and fourth principal components (Figure4.3C and S12A).

PC3 and PC4 show additionally that with advancing time, tissues separate from each other: Big points show the first timepoint of the measurement (11days or 14days for kidney, intestine and neural tube), small dots show intermediate timepoints and stars the last timepoint at birth. In PC3/PC4, timepoints at the beginning locate close to each other while the motif activities inferred for samples at birth span larger regions. (Figure 4.3C, stars). This reflects that tissues evolve their specific functions subsequently during embryonic development.

Although the tissues are defined by the activity of tissue-specific distinct motifs, we asked whether there were differences in the general time-dependent pattern of motif activity depending on tissue. Therefore we subdivided the dataset into one individual matrix for each tissue and performed PCA on this matrix (see Methods). The obtained first principal component (tissue specific PC1) points in the direction carrying most of the variance in each tissue, here the first component already captures 55.6-79.9% of the variance, depending on the tissue (Figure S12D). The most common behavior of motif activity in individual tissues across developmental time is thus a systematic in- or decrease of the activity (Figure 4.3D and S13). Note that PCA is invariant under point reflection, meaning that the pattern could as well be reversed (decreasing activity with time). We calculated which motifs follow most strongly the PC1-pattern by correlating the motif activity pattern with the PC1 pattern. For a list of motifs with highest correlation to PC1 please refer to Table S3.

Log-likelihood per target	Total log-likelihood	Term	Description
28.1	112.3	GO:0035995	detection of muscle stretch
19.1	57.2	GO:0090292	nuclear matrix organization
15.4	46.1	GO:0031034	myosin filament assembly
13.4	26.8	GO:0014878	response to electrical stimulus involved in regulation of muscle adaptation
12.1	24.1	GO:0002019	regulation of renal output by angiotensin
10.8	32.5	GO:0014873	response to muscle activity involved in regulation of muscle adaptation
10.4	51.8	GO:0098735	positive regulation of the force of heart contraction

Table 2: **Top GO categories inferred for the associated genes to the target CREs of Mef2b, according to their total log likelihood (total enrichment score).** The log likelihood per target shows the fold enrichment compared to random selection. We include the ID for the GO term as well as a more extensive description of the term.

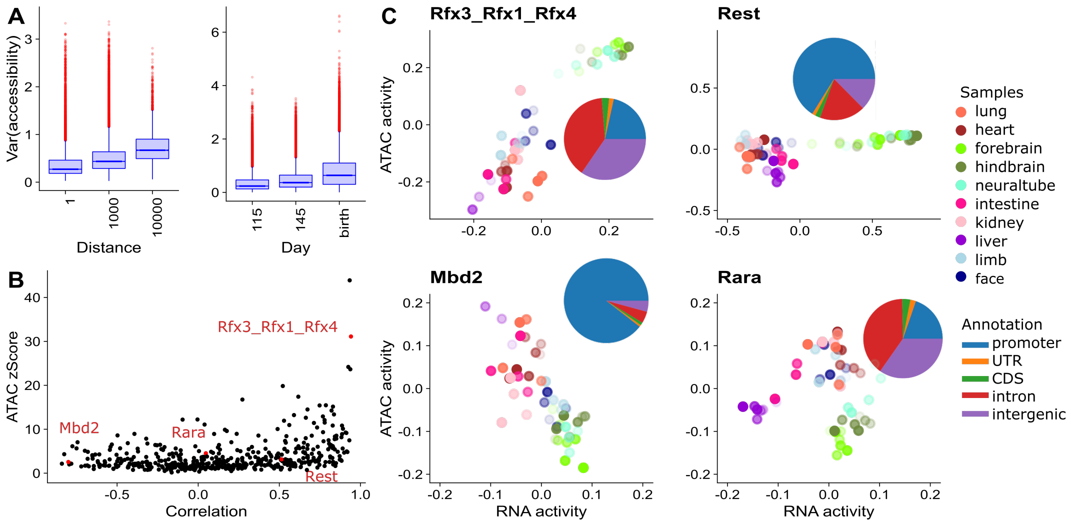


Figure 4.4: **Integration of RNA seq data with the previously analyzed ATAC seq data reveals loci-dependent activity of motifs**. A) Variance of the signal at CREs across samples. The samples taken into account to compute the variance for each CRE are stratified by timepoints B) CREs are grouped depending on the distance to the closest promoter, the box whisker plots show the variance at CREs across all samples (right) C) Scatter plot of the correlation of CREMA and ISMARA-inferred activity pattern and CREMA-inferred zScore. Motifs selected in C) are denoted by orange dots. D) Motifs selected to represent the 4 possible scenarios when comparing ISMARA and CREMA-inferred motif activities: well correlated (Rfx3), only significant changes in the ISMARA analysis (Rest), inverse correlation (Mbd2) and only significant changes in the CREMA analysis (Rara). Opacity of the dots increases with developmental time. Inlays (piecharts) show the fraction of target CREs locating to promoters, untranslated regions (UTR), coding sequences (CDS) and intronic or intergenic regions, weighted by the target score.

Timepoint and tissue-specific marker motifs

To define which motifs are rather time- or tissue dependent, we made use of CREMAs built-in averaging tool which allows users to contrast motif activities between different groups of samples by calculating average motif activities for each group and calculating how significant these averages vary across groups. This can be used to get a broader overview of which regulators drive the chromatin landscape or transcription factor binding either mainly across developmental time independent of tissues or mainly across tissues independent of developmental time. To identify factors that are either specifically active at a particular developmental timepoint or in a particular tissue, we averaged the CREMA run in two ways: 1) across all tissues for each timepoint and (time specific motifs) and 2) across all timepoints for each tissue (tissue specific motifs).

Averaging between different timepoints across all motifs (1) yields generally lower zScores than averaging across all timepoints for each tissue (2), which shows in addition to our PCA analysis that the difference in motif activity between tissues is higher than the difference in time.

Figure S14A shows that the general pattern of time specific motifs (see also Figure 4.3D) is either in or decreasing. Interestingly, also the variance of motif activity increases with time (Figure 4.3C, stars, and increasing error on averaged motif activities in Figure S14A). Motifs like Elf5, Stat4_Stat3_Stat5b and Gmeb2 act as time-specific motifs in all tissues. Using the second way of averaging yields tissue specific motifs (see Figure S14B). We identify Tal1 to have the highest motif activity of all tissues in liver, Klf4_Sp3 in intestine and liver, while Rfx2_Rfx7 and Gfi1_Gfi1b have a high motif activity only in brain associated tissues (forebrain, hindbrain, neural tube).

Comparison with ISMARA

We divided our set of CREs into four subsets, depending on the distance taken from the middle of a CRE to the start of the promoter: 1) closer than 1,000bp 2) between 1,000 and 10,000bp, 3) between 10,000 and 100,000bp and 4) larger than 100,000bp. We refer to a promoter as being the transcription start site of a gene (TSS). Further we subdivided the samples by timepoints. We calculated the variance then for each for the CREs in each group and timepoint and plotted it as a histogram (Figure 4.4A).

Here, independent of the distance, the variance increases from 11days to birth as can be seen by the density of variances across all samples at CREs at birth is much flatter than at 11days, independent of the distance. This confirms our findings from the PCA (Figure 4.3C). But the distance plays a role too: The further away, the flatter and longer the right tail of the distribution, suggesting that indeed, the more distal the CRE, the more variable in terms of accessibility.

To investigate how much the activities of motifs in driving chromatin accessibility genome-wide match or differ from the activities of motifs at promoters in driving gene expression, we used ISMARA [6] on RNA-seq data, matching the tissues and time-points of the ATAC-seq data exactly. ISMARA models the RNA expression data in terms of the same regulatory motifs that are used in CREMA. The difference is that it uses the annotation of promoters to their target genes and infers a motif activity based on the expression of the target gene and the likelihood of this motif having a binding site in its promoter. ISMARA was applied successfully over the past years to a multitude of different datasets to predict novel regulators (e.g. [51, 69, 44]). However, ISMARA relies on the motifs located in promoters to model the regulatory network. Although both CREMA and ISMARA use the same core model, the activity to affect transcription by binding to the promoter is not the same as the activity to affect DNA accessibility. Therefore, we were interested in how the inferred motif activity pattern relate.

Correlation of motif activity pattern and inferred zScores

We compared the RNA-seq derived (ISMARA) with the ATAC-seq derived (CREMA) motif activity pattern. Hence, we calculated, for each motif, the Pearson correlation between the activity pattern obtained by ISMARA and CREMA analysis. For the most significant motifs in the CREMA run also correlate highly with the motif activity patterns inferred by ISMARA (Figure 4.4B). Nevertheless, the correlation values range from -0.85 (Stat1a) to 0.96 (Tal1), so some of the motif activities obtained by CREMA analysis differ from the ones ISMARA infers. We expect to find three different scenarios when comparing motif activities from both analyses:
1) the ISMARA-inferred activity pattern correlates well with the CREMA-inferred pattern. 2) either the motif has a high zScore in the ISMARA run (which means it is located in promoters of regulated genes), but a low one in the CREMA run, or the other way round (which means that it drives the observed changes in accessibility). 3) both activity pattern are anti-correlated.
We selected four motifs to demonstrate the different scenarios (Figure 4.4C).
Motif activities for Rfx3 show a high Pearson correlation value, suggesting that it affects transcription as well as chromatin state. ISMARA and CREMA-inferred activities correlate positively (ISMARA zScore = 8.9; CREMA zScore=31.11). Note that the ATAC zScore is higher because our model runs on CREs in ATAC-seq (number of CREs: 234168, max. zScore=43.9) compared to promoters in RNAseq (number of promoters: 30115,max. zScore=8.9).
Rest is a known repressor of neuronal genes in non-neuronal tissues [4]. As such, the RNA activity pattern shows elevated activity in brain-associated tissues, although it means that Rest is not bound. The activity pattern do not correlate. ISMARA-inferred motif activities (zScore = 8.6) vary strongly across tissues and reach high values in brain-associated tissues, whereas the CREMA-inferred activity (zScore = 3.15) changes are very minor across tissues. Hence, according to our analysis, Rest is predicted to be important for the regulation of gene expression directly at promoters but does not affect change the genome-wide chromatin state. This point is further confirmed by the location of CREs with a Rest motif. Those are mainly located at promoters (inlay in fig 4.4C, Rest).

Mbd2 is one of the examples with a very negative correlation value between ISMARA (zScore = 1.5) and CREMA motif activity (zScore = 2.55), suggesting its binding might increase chromatin accessibility while its binding at promoters may lead to repression of the target genes. Indeed, this protein has been found to bind methylated regions in the DNA and act as transcriptional repressor, for example in cancer cells [57, 8, 30]. Our analysis reveals in addition that Mbd2 motifs are located predominantly at the promoter (inlay in fig 4.4C, Mbd2)
Rara or Rar, was found to be important in embryonic development [36]. With the CREMA-inferred motif activity pattern (zScore = 4.52) showing higher activity compared to the ISMARA-inferred motif activity pattern (zScore = 1.08), our results suggest Rara activity impacts or is strongly impacted the chromatin landscape. Similarly

to Rfx3, the loci of CREs with Rara motifs are located mainly in intronic or intergenic regions (inlay in fig 4.4C, Rfx3 and Rara). Rara's chromatin-modifying behaviour has recently been found leukemia, [70].

Dataset 3: ChIP for histone marks in primary cells

To verify the performance of our approach also in ChIP seq data, we applied CREMA to an extensive selection of samples downloaded from the ENCODE consortium [21, 20] (see Table S1 for accessions). ChIP was performed for different Histone modifications in human primary cells belonging to the hematopoietic (T cell, regulatory T cell (reg T cell), memory T cell (mem T cell), neutrophil, common myeoloid progenitor cell (CMP) and mesenchymal (keratinocyte, osteoblast, fibroblast, astrocyte) lineage (see methods for full names of the primary cells).

Top motifs changing activity at enhancer and promoter marks

We chose H3K4me1 (Histone 3 lysine 4 monomethylation) as one of the most studied histone modifications for enhancers [7, 29, 61]. For a direct comparison, we include H3K4me3 (Histone 3 lysine 4 trimethylation) which is generally associated with promoters [26]. We further find that the zScores for all motifs obtained by the H3K4me1 and H3K4me3 analysis are correlated (Figure S15A). Motifs which show high correlation between their H3K4me3 and H3K4me1-derived activity pattern yield higher zScores in the H3K4me1 analysis S16A). As in the previous dataset, we find that CREs marked by H3K4me1 are generally more distal from CREs marked by H3K4me3 and are more numerous (more than 60% of the H3K4me3 regions are overlapping with H3K4me1, which is reflected in total variance of activity across motifs and samples (Figure S15B)). As our analysis focuses on distal regulatory elements, we continue using the H3K4me1 dataset exclusively. Especially given the large overlap, higher variance and number of significant CREs, the H3K4me1 dataset it is more promising to infer important cell-type specific regulators.

Hematopoietic and mesenchymal cells are clustered in PCA

We performed PCA on the resulting activity matrix (Figure S15C). The first principal component (PC1) clearly separates the two different lineages, mesenchymal and hematopoietic cells with a captured variance of 45.4% (Figure S15C and S16B). The second principal component (PC2), captures the differences of cells in each lineage. Plotted vectors here were chosen to represent the motifs with the top 10 zScores, and highest 5 projection values on PC1. The projection of motif activities to the PCA space gives already first insights into which motifs are important for H3K4 monomethylation in the different cell types, for example the CXXC1 motif activity changes across different cell types in each lineage, whereas HMGA1 activity distinguishes between mesenchymal and hematopoietic cells.

The PCA gives us an overview of which motifs may be important for the difference H3K4 monomethylation at CREs across the cell types. However, this analysis does not detect motifs which are active in specific single cells types. To increase the resolution and to predict which motifs drive the differences in histone modification of specific cell lines and cell types, we performed sample averaging between selected sub-groups of our dataset (see the table in Figure 4.5). As our dataset includes cells at different stages

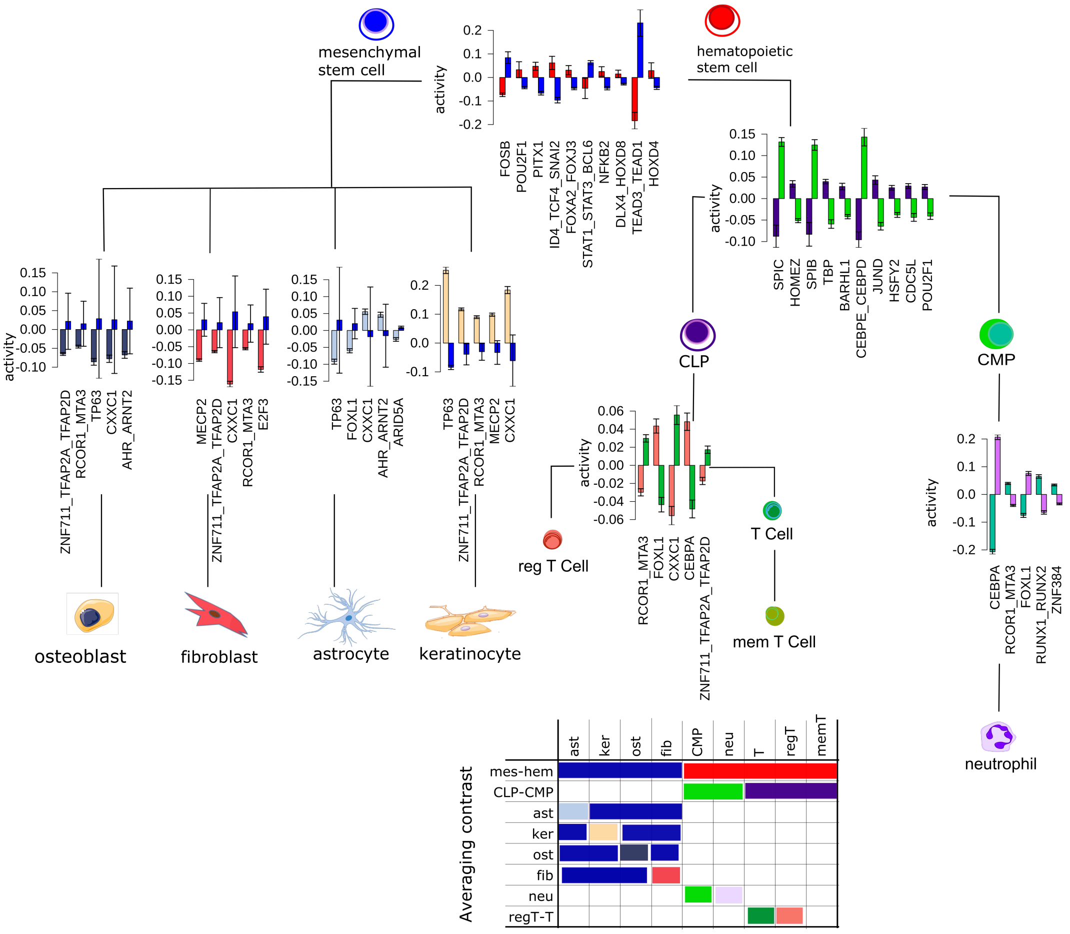


Figure 4.5: **Differentiation of hematopoietic and mesenchymal primary cells.** Inferred motifs driving the changes in H3K4 monomethylation between cell types. Activities were obtained by sample-averaging (Methods). Colors in barplots correspond to the cell types. Table in the left lower corner displays the averaging configurations matched with colors represented in the barplots. Abbreviations in the table are: hem-mes=hematopoietic-mesenchymal. CLP-CMP=common lymphoid progenitor/common myeloid progenitor, ast=astrocyte, ker=keratinocyte, ost=osteoblast, fib=fibroblast of dermis, T=CD4+ T cell, regT= CD4+ regulatory T cell, memT = CD4+ memory T cell. The length of the black lines is not related to similarity of cell types and is only used to visualize cellular differentiation paths.

of the differentiation, our aim was to determine motifs regulating H3K4 monomethylation that are specific for subsets of cells or even single cell types. Overall, we find ZNF711 (zScore 18.5), RCOR (zScore 16.7) and CEBPA (zScore 15.6)) to change their activity strongest across all examined cell types.

CEBPA has been already found to be important for the hematopoietic-mesenchymal transition [42], as has RCOR [65], whereas ZNF711 is a potentially novel motif here and has to our knowledge not been directly associated with hematopoietic or mesenchymal differentiation.

Deciphering transcription factor activity during different stages of differentiation

To go deeper into the transcription factors regulating differentiation, we averaged motif activities for all cells in our dataset which belong to the mesenchymal lineage and compare them to the average activities of cells belonging to the hematopoietic cell lineage (Figure 4.5). Motifs are sorted by their zScore in the overall run, from left to right. We find POU2F1 (zScore, merged: 6.7) for the averaged motif activity) and FOSB (zScore, merged: 7.0) to be important for the difference in H3K4 monomethylation between hematopoietic and mesenchymal cells.

Hematopoietic lineage: In osteoblasts, FOSB was found to play a role in differentiation: stretching induced its transcription and was followed by expression of osteoblast markers in human mesenchymal precursor cells [28]. POU2F1 has been associated previously with the epithelial-mesenchymal transition in cancer [75]. TEAD3 (overall zScore 11.4) is important in the epithelial-mesenchymal transition [73], our analysis predicts it to be a strong indicator for distinguishing mesenchymal from hematopoietic cells based on H3K4 monomethylation.
Going further into the hematopoietic lineage, we averaged transcription factor activities between cells evolving from common myeloid progenitor cells (CMP) and those evolving from common lymphoid cells (CLP). In case of infections, the decision that cells become CMP or CLP can be forced towards the myeloid cells [24, 46]. In our analysis, the activity of CEBPE/CEBPD and SPIC are driving the differences in H3K4me1 deposition between cells deriving from CMP or CLP . A study from 2017 found CEBPB and BACH2 acting in a feed forward loop to regulate myeloid differentiation [35]. Given the similarity of weight matrices of BACH2 and BACH1 (overall zScore 6.6), the motif-TF association may not be completely correct, thus we would rather infer BACH1 than BACH2. BACH2 indeed shows a peak in its activity pattern in CMP in the overall CREMA run (Figure S17).
Going further, we investigated which regulatory key players are involved in the difference in H3K4 monomethylation in neutrophils versus CMP. Strikingly, CEBPA shows high activity in neutrophils compared to CMP (zScore merged: 23.7), which is in accordance with a previous study that defined CEBPA as marker for neutrophils [3]. Next we questioned what are the most important motifs for the difference in H3K4me1 between regulatory T cells (reg T Cell) and T Cells. Here, we find RCOR1 being highly active in T-cells, but it has not been mentioned in relation to T-cell specification previously. Additionally, a very interesting candidate is CXXC1, which was found to play a role in the differentiation of T cells by regulating the H3K4me3 deposition at promoters of key genes important for thymocyte survival [15]. This confirms our previous finding that H3K4me1 and H3K4me3 marks are highly correlated. Consistent with this, predicted GO categories for CXXC1 are "negative regulation of histone methylation" (GO:0031061), but also T cell associated categories such as "T cell tolerance induction"(GO:0002517) and "CD8positive, alpha-beta T cell activation" (GO:0043374). FOXL1, is more active in regulatory T cells compared to T cells. Although FOXL1 has not been implicated directly with T cell development, FOXP3 was found to be an important regulator in the development and function of regulatory T cells [38]. From our results, also FOXP3 (overall zScore 2.8) is highly active comparing all cells, but not to the extend of FOXL1 (overall zScore 13.8). However, the weight matrices of both motifs differ, which suggests that FOXL1 may be a novel regulator implicated in regulatory

T cell development.

Mesenchymal lineage: Regarding the mesenchymal cell lineage, we compared each cell to all others to determine which regulators are driving H3K4 monomethylation here. We get a very clear signal for keratinocyte, with TP63 (zScore, merged: 17.6) being highly active. Indeed, TP63 is a keratinocyte-exclusive master regulator in epidermal development [50, 64]. Our pathway analysis (Reactome) predicts TP63 to be involved in "Genes involved in collagen formation" and "ErbB receptor signaling network", which is both related to epidermal development.
In astrocytes, we find, besides TP63, Ahr (zScore, merged: 4.5) and CXXC1 (zScore, merged: 4.9) to be active. While CXXC1 has not been associated with astrocytes, Ahr was found to modulate astrocyte-related transcription programs [54]. The top motif differentiating H3K4me1 methylation of other mesenchymal cells from fibroblasts is MECP2 (zScore, merged: 16.8). Notably, a connection of histone modification and myofibroblast differentiation has been found earlier [56, 32]. The data for osteoblasts, fibroblasts and astrocytes is very noisy, so we are not confident enough to draw strong conclusions here.

DISCUSSION

Motif activity response analysis, the core of the ISMARA [6] model has already been applied successfully to RNA-sequencing data [51, 69, 44]. However, the ISMARA predictions are based solely on binding sites in promoters. As it is known that distal regulatory regions play a non-neglectable role in gene regulation, a growing number of scientists aim to get a genome-wide view on their system. Here, we proposed CREMA, a novel approach to analyze the chromatin state genome-wide in terms of regulatory motifs. Using CRUNCH [9] to scan for significant regulatory elements across the whole genome, CREMA applies the core model of ISMARA to infer key regulatory motifs which explain changes in chromatin state across a non-limited number of samples.
Using CREMA has several advantages. First, we reduce the high dimensionality of the dataset, often including up to 500,000 cis regulatory regions by explaining observed changes in chromatin state with the activities of around 600 transcription factors. Narrowing down the number of potential key regulators is a crucial step for the design of follow-up experiments and hypotheses. Secondly, CREMA not only analyzes the chromatin state of each sample separately, but reports the differences of motif activity across samples. This is especially useful to define which key regulators play a role in certain conditions, cells types or treatments. Thirdly, CREMA is implemented as a fully automated pipeline freely accessible our webserver. Thus, everyone dealing with high throughput data can apply CREMA, without the prerequisite of having previous knowledge in computational biology.
To display the performance and versatility of our model, we applied it to three datasets. Using a DNase-seq dataset, we investigated which regulatory key elements are following circadian oscillation pattern across a 24h timecourse. The top motifs that CREMA predicts to be involved in shaping the chromatin state across a 24h cycle were indeed known circadian regulators that show a strongly cycling motif activity pattern. Additionally we find regulators, e.g. Stat5a and Irf2 whose activity oscillates strongly with

24h period but have not been found to oscillate in a circadian manner before.
Further, we applied the model to ATAC-seq data on 10 different murine embryonic tissues at 4-7 different developmental stages. After modeling the data in terms of regulators using CREMA, we find motifs shaping the chromatin state of specific tissues while other motifs are more universal and change their activity in multiple tissues at the same time. Although motif activities are mostly tissue specific, the general pattern of motif activity is similar across tissues. PCA on the activity matrix for each tissue separately shows that most of the variance, around 40-60%, can be explained by a time-dependent in- or decreasing pattern, this holds for all tissues. To highlight the novelty of results obtained by our genome-wide approach, we compared it to the results inferred by ISMARA on RNA expression data for the same tissues and same timepoints. While we find generally a high correlation between CREMA and ISMARA-inferred patterns, we indeed find motifs that are found to be significantly active only in the CREMA analysis, such as Rara, whereas other motifs are presumably recruited only to promoters and thus their activity changes significantly in the ISMARA analysis, but not in the CREMA analysis, for example Rest.
Using a large dataset where ChIP was performed against two different histone modification marks in 9 types of human primary cells, we find that CREs marked by H3K4me1 are more distal and numerous than those marked by H3K4me3. Using the built-in averaging tool of CREMA, we analyzed motifs implicated in the decision of cell fate. Here we show for instance that TEAD factors play a role in distinguishing mesenchymal from hematopoietic cells and identify CEBPA to be a strong regulator for neutrophil differentiation.
Of course, there are some drawbacks in our model. Especially in ChIP-seq data, the differences in antibody efficacy for different target proteins are not distinguishable from biological relevant binding signals. Although we use a sophisticated strategy to normalize the data, there can be biases if samples are prepared using antibodies with differing efficacy. Adjusting for this would require previous exact determination of the antibody efficacy for each sample and could be included into the normalization procedure in future implementations. The linear model we use is very simple which has the advantage to be easily solvable, but of course lacks additional parameters, for example differences in function of these factors (chromatin opening, transcriptional activators or repressors) or the interplay of transcription factors. A sole binding site, as our model takes it into account, may not be enough to explain the changes in chromatin state. Most transcription factors don't act alone but in large complexes and not all of them bind to the DNA. Including the formation of higher order complexes of transcription factors or distinct functions of transcription factors into our model would certainly improve our predictions.

One of the advantages of CREMA, the genome-wide view on chromatin state, makes it – in turn – very hard to assign and reliably predict regulated genes to the motifs. Our inferred CREs often locate in intronic or intergenic regions, and could thus be assigned to a large number of surrounding genes. Knowledge on enhancer-promoter associations would improve our CRE-gene associations and functional analysis tremendously.

METHODS

Datasets

We downloaded raw fastq files from the ENCODE database [20, 21]. For a complete list of accessions, see table S1.

- Dataset 1: DNase-seq of murine liver of mice left in darkness for 50-70 (Panda et al, 2002) [?]
- Dataset 2.1: ATAC-seq for 11 different tissues and at 4-7 different timepoints from E11.5 - birth in the embryos in mice
- Dataset 2.2: RNA-seq for 11 different tissues and at 4-8 different timepoints from E10.5 - birth in the embryos in mice
- Dataset 3: ChIP-seq for H3K4me1 and H3K4me3 across 9 primary cell lines in cultured human cells

The ChIP-seq data we use consists of foreground and background samples. For ATAC-seq and DNase-seq we assume the background to be uniformly distributed.
For a clearer visualization of the data and to be in accordance with common notation, we renamed the samples for the first dataset in the following way: darkness50=CT14; darkness54=CT18; darkness58=CT22; darkness62=CT02; darkness66=CT06; darkness70=CT10.
For second datasets we chose: E11.5=11days, E12.5=12days, E13.5=13days, E14.5=14days, E15.5=15days, E16.5=16days, and 0h=birth For the third dataset: common myeloid progenitor, CD34-positive = CMP, CD4-positive, alpha-beta memory T cell= mem T cell, CD4-positive, alpha-beta T cell= T cell, alpha-beta regulatory T cell= reg T cell, astrocyte, keratinocyte, neutrophil, osteoblast. In the second dataset, we refer to the timepoint denoted in as '0' or postnatal as 'birth'.

Quality filtering and Adapter trimming

We use `cutadapt` [?] as it can quality trim
(`-quality-cutoff 20`) and trim the adapters in the same step. For the adapter trimming, we first find the most abundant adapters by running `cutadapt` on the first 1000000 sequences. We use a pre defined list of commonly used adapter sequences from illumina S5 and scan it across our reads. The adapter with the highest number of matches is chosen. [34]:
Then we run `cutadapt` in the following mode, with `$ENCODING` a placeholder for the sequencing machine and `$ADAPTER` the adapter, both are inferred in the previous step.
We then remove reads which are too short (<25bp) and reads with more than two undefined bases.

```
cutadapt ——quality-cutoff 20 ——quality-base=$ENCODING  ——minimum-length
25 ——trim-n
——max-n 2 -a $ADAPTER ——overlap 10 ——cores 12
```

Mapping to the reference genome

For the mapping we use bowtie1 [40]. Bowtie reports only one mapping position for one read(-a strata -best), choosing the ones with least mismatches, while allowing for 3 mismatches (-v 3). Multi mapping reads are then distributed across their matching positions n , we keep a weight for each read as $w = 1/n$. We call this format *bedweight* and it is used later in the pipeline to estimate exact read counts for specific regions on the genome. Typically, more than 70-80% of the input reads are mapped. We don't take into consideration chrM, to avoid contamination of reads stemming from mitochondrial DNA. In case Paired-end data is available, we treat every read as a separate observation (ATAC and DNase data), whereas in ChIP-seq, we use paired-end data to infer the exact middle of each fragment.

Fragment size estimation

As most of the datasets we looked at are sequenced according to a single end protocol, they don't contain any information about the fragment length. For ChIP seq, we follow the approach denoted in [9]. Note that this procedure is different for ATAC and DNase seq: here, the reads are not shifted by the fragment length. Other than for ChIP seq (described above), where we want to estimate the position of a TF, we use the mapped position of each read without shifting this position to the estimated middle of the fragment. This is based on the length of DNA that is typically wrapped around one nucleosome. For ATAC/DNase seq we set the fragment length to 150 (see also paragraph "Counting fragments in sliding windows".

Counting fragments in sliding windows

We slide a window of 500bp across the genome, shifted by 250bp in each step. The length of the window is in accordance with the expected signal: the typical width of a ChIP peak is usually ± 75 around the TSS, which is the length of DNA that is wrapped around the nucleosome, as well as the ATAC/DNase insert size [**?**, **?**, 9]. The length of 500bp thus provides good resolution and will capture most of the binding events and accessible regions. Nevertheless, histones can span longer regions on the DNA, we account for this, as our approach dynamically merges enriched regions before scanning for peaks. The background is fluctuating much slower and the read density is typically lower. We account for this by taking a larger window for the background samples, 2000bp by default.

When we run those two sliding window across the whole genome we simply count all the reads that fall into this window each for foreground n and background m. In the case of replicates, the raw counts for all foreground and background replicates are summed up. If we lack background data, we generated a set of background data out of pooled ChIP background datasets (available for human, [9] and mouse (Control ChIP-seg data from the Bing Ren Lab for the 11 tissues at up to 7 timepoints in dataset 2, see supplement Table S1 for accessions). In case of ATAC or DNase-seq, we use a constant background, meaning 100 reads per background window.

Peak calling

For peak calling, we use CRUNCH for each sample separately. For details please refer to the methods and supplemental methods used for CRUNCH [9]. Next we estimate the probability of a region to be significantly enriched (foreground counts significantly higher than background counts) using a bayesian mixture model fitted to the regions. An individual zScore for each identified region is calculated and regions passing a previously defined threshold make it into final peak calling step, to find the exact location

of the binding event (ChIP-seq) or open region (ATAC-seq/DNase-seq).
To call the peaks, we fit the read coverage at each region in base-pair resolution to a mixture of multiple gaussians and a constant background. After selecting for significant peaks, we are left with an individual peak set for each sample.

Determination of the number of fitted peaks

as in [9], we use the following estimation, for N the number of possible peaks, R_l the length of the region and fraglen the fragment length:

$$N = \frac{R_l}{fraglen * 2} \tag{24}$$

Construction of the common CRE set

After calling the peaks for each sample, a common set of regulatory regions which captures all the peaks in all samples is created. We retain all the peaks which were significantly enriched in at least one of the samples. Then all the peaks from all different samples are overlapped and merged to cis regulatory elements (CREs), if their centers are closer than 75bp. In the case of merging, the resulting CRE thus spans all the centers of peaks belonging to it ±75 bp up and downstream. We chose 75bp as cutoff as one nucleosome approximately spans 150bp on the DNA, which fits with the general width of a transcription factor binding peak as well.
The analysis of genomic data imposes challenges in the analysis and comparison across different samples: For ChIP-seq we aim to estimate the strength of the binding of the TF to the DNA from the height of the peak, for DNase and ATAC seq it is the fraction of cells that have increased accessibility at the same region. The inferrence of these parameters implies an accurate estimation of the read density and its fluctuations along the genome. We compute a signal (peak height) S_{cs} for each CRE c in each sample s, which is normalized for background and inter-sample library size differences.

$$S_{cs} = \log\left(\frac{f_{cs}}{F_s} \cdot \tilde{F} + 1\right) - \log\left(\frac{b_{cs}}{B_s} \cdot \tilde{F} + 1\right) - \log\left(l_c/L_c\right) \tag{25}$$

For ATAC-seq and DNase, we use:

$$S_{cs} = \log\left(\frac{f_{cs}}{F_s} \cdot \tilde{F} + 1\right) \tag{26}$$

where f_{cs} and b_{cs} is the readcount across CRE c in sample s. F_s and B_s are the library size of background and foreground in each sample s. $\tilde{F}$ denotes the median of library sizes in the foreground. l_c and L_c are the length of the CRE c in the foreground and the corresponding background CRE, these are similar across samples s. If a foreground region is shorter than 750bp, we choose the background region to be fixed to 750bp. As in the step before, we account for slow fluctuations and smooth the ideally uniform background distribution. Is the CRE bigger than 750bp we use the same size for foreground and background CRE.

Predicting binding sites genome-wide

To rigorously find binding sites in our set of CREs, we use MotEvo [1], which has been developed earlier in our group. MotEvo calculates the probability for each motif m to occur in sequence of CRE c. Note that the binding sites are non-overlapping. Motevo is run without information on the genomic coordinate (without alignment) and without background model (UFE). To account for the missing background model, we change the background prior bgp to 1-((1-$\mathrm{bgp}_{\mathrm{withUFE}}$)/100). We store the probabilities for each CRE and each motif in our sitecount matrix N_{cm}. If one sequence has multiple binding sites for one motif, the probabilities are summed up.

The ISMARA model

We now assume that the signal S_{cs} for each CRE c and each sample s, can be explained by the TFBS for motif m in CRE c, stored in matrix $N_{c,m}$ and an unknown motif activity A_{ms}. This approach is adapted from the previously in our group developed MARA model [49], which models RNA-seq data in terms of binding sites in the promoters of expressed genes. We can now fit the linear model

$$S_{cs} = \sum_{m} N_{cm} \cdot A_{ms} + \tilde{c}_c + c_s + \text{noise} \tag{27}$$

to estimate the activity A_{ms} of each motif in each sample. The CRE- and sample-dependent constants $\tilde{c}_c$ and c_s are CRE and sample dependent constants which are estimated in the next step. The noise term accounts for measurement errors in S_{cs} plus biological fluctuations throughout the samples and the error in the model. For more details see also supplemental methods.

Target annotation

To estimate which CREs are most important to explain the motif activity, we calculate a target CRE score ζ_{mc} for each CRE c and each motif m. To do so, we remove the binding site of motif m in region CRE c ($N_{cm} = 0$) and estimate the amount by which the relative square deviation between the model without this binding site (mutated version) and the one with all binding sites (full version), decreases. We only consider CREs for further analyses which have a positive target CRE score. To calculate the squared-deviation between the observed peak height in our signal matrix S_{cs} and the predicted value $\sum N_{cm}A_{ms}$, we define:

$$\chi^2_{cs} = (S_{cs} - \sum_{m} N'_{cm}A'_{ms})^2 \tag{28}$$

and calculate this value for both the full (χ^2_{cs} and the mutated version χ^2_{csm}) of the model. Using the fact that we have much more CREs than motifs, we can calculate the average square deviation per sample/CRE pair as:

$$\langle\chi^2\rangle = \frac{1}{CS}\sum_{i,s}\chi^2_{cs} \tag{29}$$

with C as the total number of CREs and S the number of samples. In the output table we show for each motif and targeted CRE the following target score:

$$\zeta_{cm} = \frac{\sum_s \chi^2_{csm} - \chi^2_{cs}}{\langle\chi^2\rangle} \tag{30}$$

Targets are then assigned as the closest gene found on the genome, up or downstream. We indicate the distance (middle of the region - promoter start) in the report page Figure S4,S8.

Assigning genes to CREs: CRE score and gene score

As we find for each dataset around 60000 - 500000 CREs, every gene may be associated with multiple CREs, and thus even more regulatory motifs. Therefore we calculate a score for each gene to be regulated by a certain motif. Every CRE gets a score for a gene depending on their distance d_{CG}, which is the distance from the middle of the CRE to the start of the promoter of a gene. For more information on our library of promoters please refer to [6].

$$w_c(G) = \frac{0.95}{1+(\frac{d_{CG}}{d_p})^2} + \frac{0.05}{1+(\frac{d_{CG}}{d_d})^2} \tag{31}$$

Promoters that are only $d_p = 150bp$ away should get the CRE assigned with 95% probability. If there is no CRE close to the promoter, we assume that the weight decreases on a 50kb scale with $d_d = 50,000pb$. Having assigned the weight $w_{CRE}(G)$ we can now calculate the probability that CRE c regulates gene G. We call this score CRE/gene association score (P_{cg}):

$$P_c(G) = \frac{w_c(G)}{w_0 + \sum_g w_c(g)} \tag{32}$$

Here, we set $w_0 = 1/100$. Now, the gene targeting score ξ for a gene to be regulated by a motif is the sum of the nearby CREs and the corresponding target CRE score ζ_{mc} for the CRE c with motif m.

$$\xi_m(G) = \sum_c \zeta_{cm} P_c(G) \tag{33}$$

Gene Ontology Analysis

We populate all the categories in the hierarchies "biological process", "cellular component" and "molecular function" with genes associated with CREs with motif m. Then we apply iteratively the following procedure: Until each gene has found a category:

- we calculate a total enrichment score as $S_{GO,m} = 1/N \sum_g \xi_m(g)$ for each category in each hierarchy, with N the number of genes mapping to category GO.
- the top scoring category is reported.
- the top scoring category is removed and all the genes mapping to it are removed from other categories.

This procedure is done similarly for the cellular component (CP) and REACTOME categories of the Molecular Signature Database (MSigDB) [**?**, **?**].

Sample Averaging

We grouped the samples in different ways (Table S4): For averaging the samples we use a dedicated procedure outlined in detail in the supplementary methods and [6] and start from the activity table including all samples.

Fit to sinusoidal curve

As the majority of CREs are accessible in a time-dependent manner, we checked for the significance of circadian cycling using a sinusoidal function for time t, with A the amplitude, ω the frequency and ϕ the phase.

$$y(t) = A\sin(\omega t + \phi) \tag{34}$$

This function can be written as (with $\omega = 2\pi/24$):

$$y(t) = \alpha \sin(2\pi/24 \cdot t) + \beta \cos(2\pi/24 \cdot t) \tag{35}$$

with, written in polar coordinates $\alpha = r\cos(\phi) and \beta = r\sin(\phi)$, we can calculate the phase $\phi = \arctan(\beta/\alpha)$ and the amplitude $r = \sqrt{\alpha^2 + \beta^2}$

Calculation of the Phase Distribution

To determine the phase shift of motifs, we calculated the following function $f(\phi)$ for a sequence of $\phi \in [-\pi, \pi]$. w_m is a motif-associated weight which depends on how much the phase of the motif's activity pattern ϕ_m deviates from ϕ and the second weight is the CREMA-inferred zScore for this motif. That is, we show the phase distribution across all motifs weighted by their zScore.

$$w_m = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2}\frac{(\Delta\phi)^2}{\sigma^2}\right) \tag{36}$$

$$f(\phi) = \sum_m \frac{z_m^2 w_m}{\sum_m z_m^2} \tag{37}$$

with $\Delta\phi = \min(\phi_m - \phi, 2\pi + \phi - \phi_m)$.

SVD / PCA on the motif activities

SVD is a generalization of matrix diagonalization to non-square matrices and can be used for lower-dimensional representation of high dimensional data. We applied it to the activities of 503 motifs across all tissues and time-points (samples). For example, for our second dataset, our matrix A_{ms} contains 55 x 503 (samples x motifs) entries. SVD decomposes the matrix as

$$A = U\Lambda V \tag{38}$$

where V and U are the right and left singular vectors, respectively. Λ contains the singular values s_k. The vectors $\vec{v}_k$ create a new orthonormal basis pointing to the directions that capture most of the variance in across all motifs. For more details, please refer to the supplemental methods.

Output

CREMA provides an extensive report which can be accessed online and downloaded to the users hard drive for analysis. We here provide an overview of the plots and analysis provided (see Figure S2 - S10), note that for the calculation of the PCA in this paper we use the R `svd`.

The main page shows all motifs sorted by their zvalue. Additionally given in this table is the associated transcription factor, a small activity profile plot and the sequence logo of the motif. For further investigation, a click on each motif opens a separate page with more detailed information:

- a logo of all associated motifs
- a list of all associated motifs and their corresponding transcription factors (Gene ID)
- a detailed activity plot and PCA across motif activity and CRE signal
- a barplot showing the zScores of the activity A_{ms} of the selected motif m across all samples
- the targeted CREs and the associated genes with information on the distance and the target CRE score ζ_{mc} and the CRE/gene association score P_{cp}.
- information on the location of CREs containing a binding site for this motif across the whole genome: a distance histogram showing the log10(distance) of CREs weighted by the target CRE score ζ_{mc} and a piechart which shows the fraction (without and with weighting by ζ_{mc}) of CREs mapping to UTRs, exons, intron, promoters and intergenic regions.

Consistent with the notion of ISMARA, we use promoters (transcription start sites (TSS)). These were identified by CAGE data analysis for mouse and human and one promoter may include several co-regulated TSS if they locate close to each other. For details please refer to the supplemental material of the ISMARA paper published earlier [6].

DATA ACCESS Accession numbers can be found in Table S1 and on `crema.unibas.ch`

ACKNOWLEDGEMENTS This research was supported by SystemsX.ch with the iPhD grant.

COMPETING INTERESTS The authors declare that they have no competing interests.

Supplementary Methods

Sample Averaging

We assume that we can write the activities A_s as $A_s = A^g + \delta_s$. Thus, the probability of having activity A_s in group g is:

$$P(A_s|\overline{A}^g, \sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{(A_s - \overline{A}^g)^2}{2\sigma^2} \right]$$

As the activity $\overline{A}^g$ is the expected value $A_s^* \pm \delta A_s$ we can write:

$$P(D|A_s) = \frac{1}{\sqrt{2\pi}\delta A_s} \exp\left[-\frac{(A_s - A_s^*)^2}{2(\delta A_s)^2} \right]$$

The mean and errorbar of the averaged activity in each group is then:

$$\langle \overline{A}^g \rangle = \frac{\sum_{s \in G} \frac{A^*_s}{(\sigma_g *)^2 + \sigma_s^2}}{\sum_{s \in G} \frac{1}{(\sigma_g *)^2 + \sigma_s^2}}$$

$$\delta \overline{A}^g = \sqrt{\frac{1}{\sum_{s \in G} \frac{1}{(\sigma^*_g)^2 + \sigma_s^2}}}$$

and a zScore can be calculated in the following:

$$z_m = \sqrt{\frac{1}{|G|} \sum_g \left(\frac{\langle \overline{A}^g \rangle}{\delta \overline{A}^g} \right)}$$

with $|G|$ the number of groups. When there is little variance between samples in each group for a motif, this motif will have a high significance z_m.

ISMARA model

In detail, MARA makes use of a bayesian procedure assuming the noise is gaussian distributed with variance σ^2 and equal for all CREs and samples. The likelihood of obtaining the signal table S_{cs} is then given by:

$$P(S|A) \propto \prod_{c,s} \frac{1}{\sigma} \exp \left[- \frac{\left(S_{cs} - \tilde{c}_c - c_s - \sum_m N_{cm} A_{ms} \right)^2}{2\sigma^2} \right]$$

We maximize this likelihood in terms of the CRE- and sample-dependent constants and replace them with the maximum likelihood estimations. Which gives:

$$P(S|A) \propto \sigma^{-CS} \exp \left[- \frac{\sum_{c,s} \left(S'_{cs} - \sum_m N'_{cm} A'_{ms} \right)^2}{2\sigma^2} \right]$$

with C the total number of CREs. Note that the table S'_{cs} is now centralized, such that mean of all the rows and the mean of all columns is zero. The sitecount values in N_{cm} is normalized in such way that the average count across all CREs is zero ($\sum_i N'_{cm} = 0$). This way we obtain activity values A' that the average across all sample for each motif is zero. To avoid the overfitting, each activity gets a gaussian distributed prior

$$P(A'|\lambda, \sigma) \propto \prod_s \exp \left[- \frac{\lambda^2}{2\sigma^2} A'^2_{ms} \right]$$

With this, the posterior distribution becomes:

$$P(A|EN) \propto \exp \left[- \frac{\sum_{i,m} \left(\left(S'_{cs} - \sum_m N'_{cm} A'_{ms} \right)^2 + \lambda^2 \sum_m A'^2_{ms} \right)}{2\sigma^2} \right]$$

the parameter lambda is fitted through a cross validation approach, using 80% of the CREs as train; and the remaining 20% as testset. The lambda that minimizes the average square deviation of the expression levels int the test set versus those predicted by the fit of the train set is chosen as optimal lambda. This posterior probability can be calculated via a ridge regression procedure, in this case SVD is used. The resulting activities are then sorted by their z-score

$$z_m = \sqrt{\frac{1}{S}\sum_s \left(\frac{A'_{ms}}{\delta A'_{ms}}\right)^2}$$

SVD / PCA on the motif activities

SVD decomposes the matrix as

$$A = U\Lambda V$$

where V and U are the right and left singular vectors, respectively. Λ contains the singular values s_k. The vectors $\vec{v}_k$ create a new orthonormal basis pointing to the directions that capture most of the variance in across all motifs. Any motif activity pattern $(\vec{a}_m)_s$ can now be represented in a linear combination of the right singular vectors $\vec{v}_k$. The coordinates of the motifs in the sample space thus depict the contribution of each sample to the motif activity pattern. As the singular vectors v_k point in the direction of highest variance in the dataset, they capture consecutively the pattern contributing to the variation in the dataset, e.g. the first component captures the highest amount of variance, followed by the second component. The projection of the motif activities $\vec{a}_m s$ can be obtained by calculating $\vec{q}_m k = \vec{a}_m \cdot v_k$ or $q_{mk} = (AV)_m k$. When plotted across samples, the singular vectors show the distinct pattern of motif activities detected by the SVD. Note that this approach is the same as principal component analysis (PCA): The singular vectors v we obtain in SVD are the same as those obtained by PCA, which is performed on the covariance matrix C, with n the number of samples:

$$C = A^T A/(1-n) = V\Lambda U^T U\Lambda V^T/(1-n) = V\Lambda^2 V^T/(n-1)$$

The only difference is now that the eigenvalues e_i of the covariance matrix C relate to the singular values s_i as $e_i = s_i^2/(n-1)$, with the fraction of explained variance $FOV = s_i^2/\sum_i(s_i^2)$.

To know which pattern of $\vec{a}_m$ are following which singular vector, we make use of a geometric approach which allows to identify vectors in a_m that follow the direction of the principal components.

As now, the vectors a_m are vectors in the principal component space, their projection on each of the axes (=principal components) indicates how strongly the vector a_m overlaps with the singular vector v_k. We an now calculate the projections according to $q_{mk} = \vec{a}_m \cdot \vec{v}_k$, and as, according to the SVD, $AV = U\Lambda$ it can be written in matrix multiplication $q_{mk} = (U\Lambda)_{mk}$.

Note that the projection captures also how strongly the vector a_m follows a singular vector, e.g. how 'long' it is. Still, short vectors can still correlate pretty well with the singular vectors. The correlation p can be obtained by calculating: $p_{mk} = q_{mk}/\sqrt{\sum_k (q_{mk})^2}$. Note that the singular vectors v form a new orthogonal basis and are thus independent of each other, each capturing another pattern in the data.

SUPPLEMENTARY MATERIAL

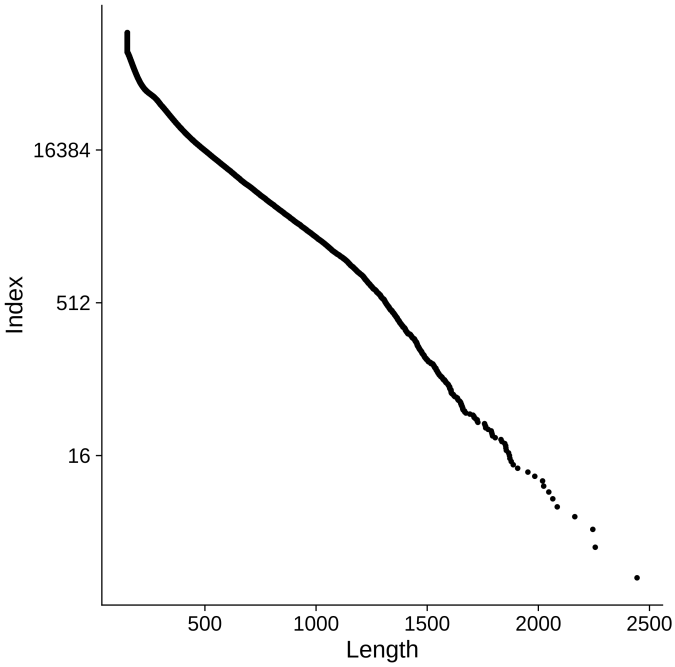
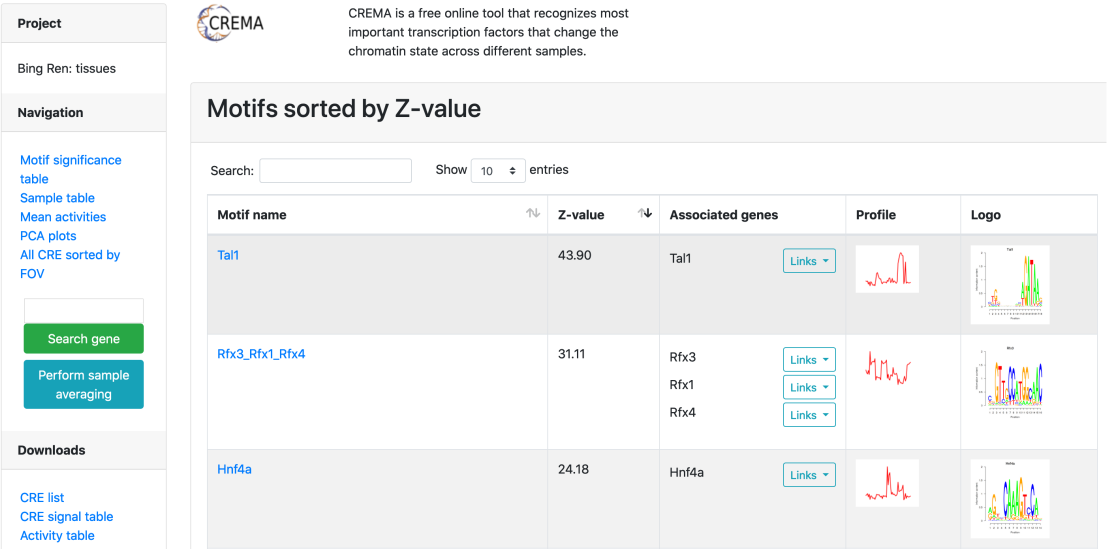


Figure S1: Length of inferred CREs, ordered from short to long. CREs were taken from dataset 2.



CREMA is a free online tool that recognizes most important transcription factors that change the chromatin state across different samples.

Project

Bing Ren: tissues

Navigation

Motif significance table
Sample table
Mean activities
PCA plots
All CRE sorted by FOV

Downloads

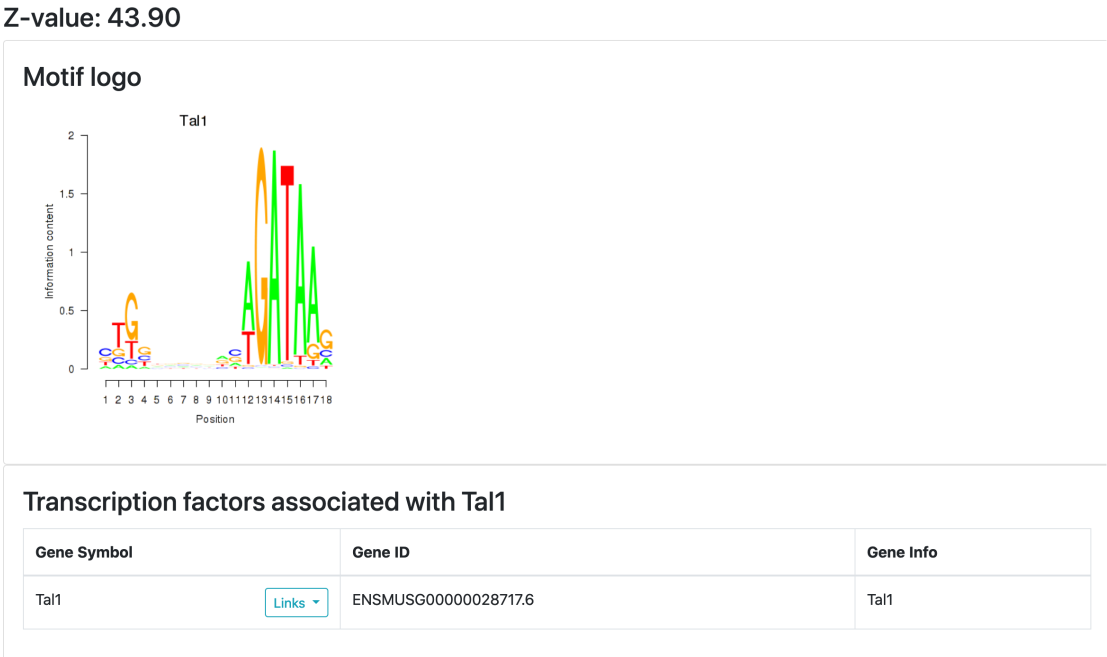
CRE list
CRE signal table
Activity table

Motifs sorted by Z-value

Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links Rfx1 Links Rfx4 Links		
Hnf4a	24.18	Hnf4a Links		

Figure S2: The main page of the report: For each motif, we provide zScore, associated genes, the weight matrix of the motif and a small motif activity profile. The image is a screenshot from `crema.unibas.ch`.



Z-value: 43.90

Motif logo

Transcription factors associated with Tal1

Gene Symbol	Gene ID	Gene Info
Tal1 Links	ENSMUSG00000028717.6	Tal1

Figure S3: A click on one motif opens the motif-specific page. We provide information about the weight matrix (motif binding site) and associated transcription factors. The image is a screenshot from `crema.unibas.ch`.

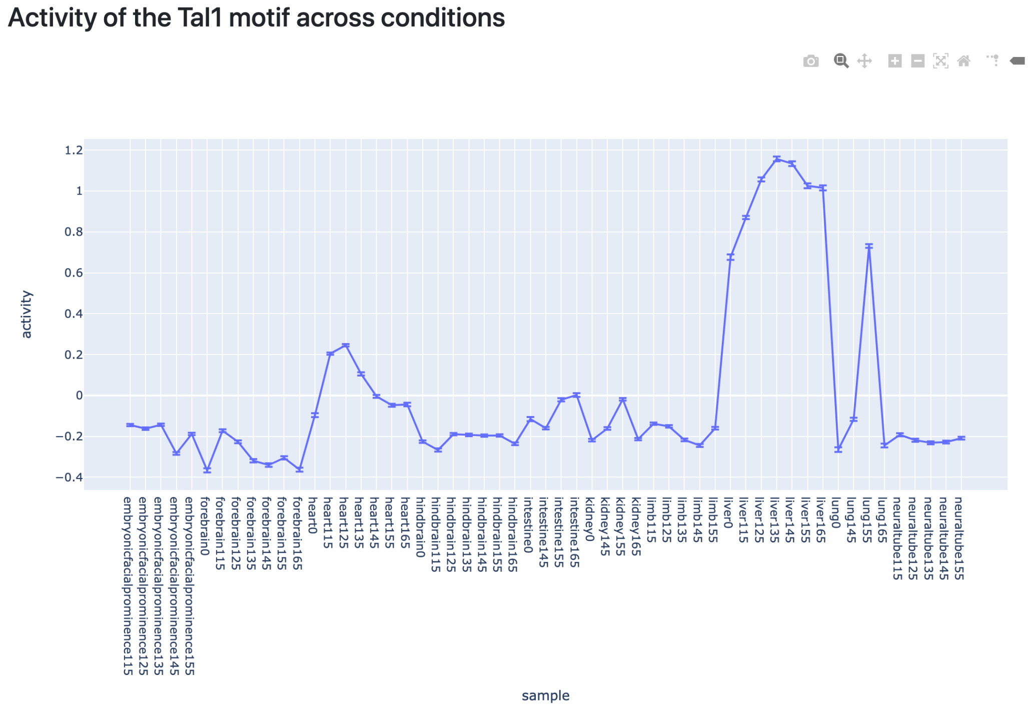
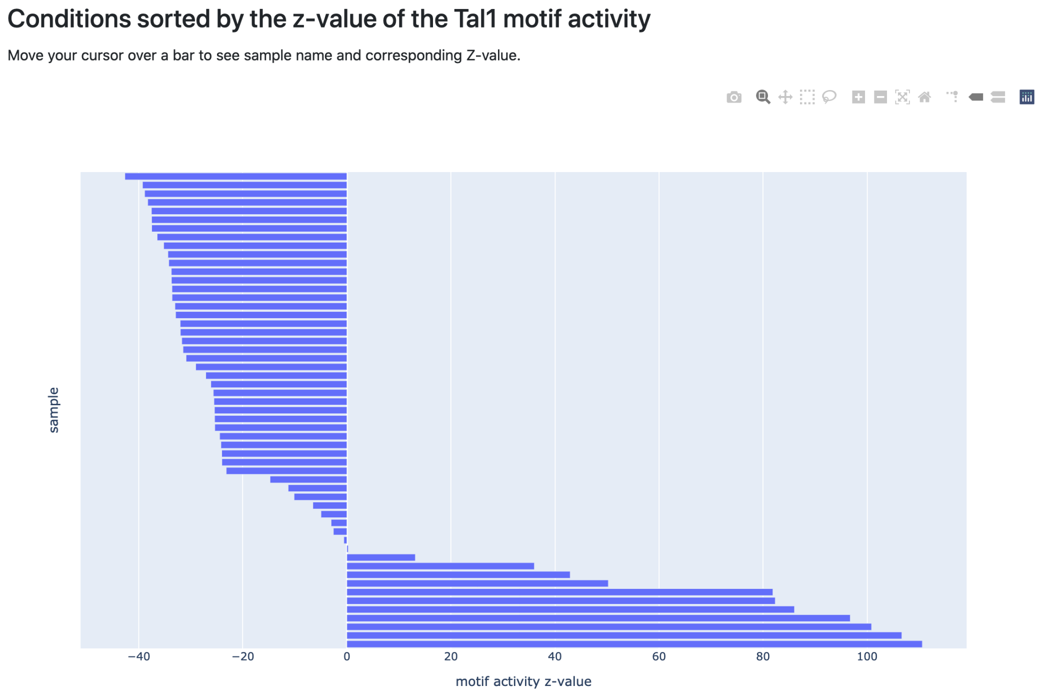


Figure S4: Activity profile. Users are able to zoom into the plot and exact values of activities A_{ms} and the errorbars δA_{ms} are given by sliding the mouse over the dots in the plot. The image is a screenshot from `crema.unibas.ch`

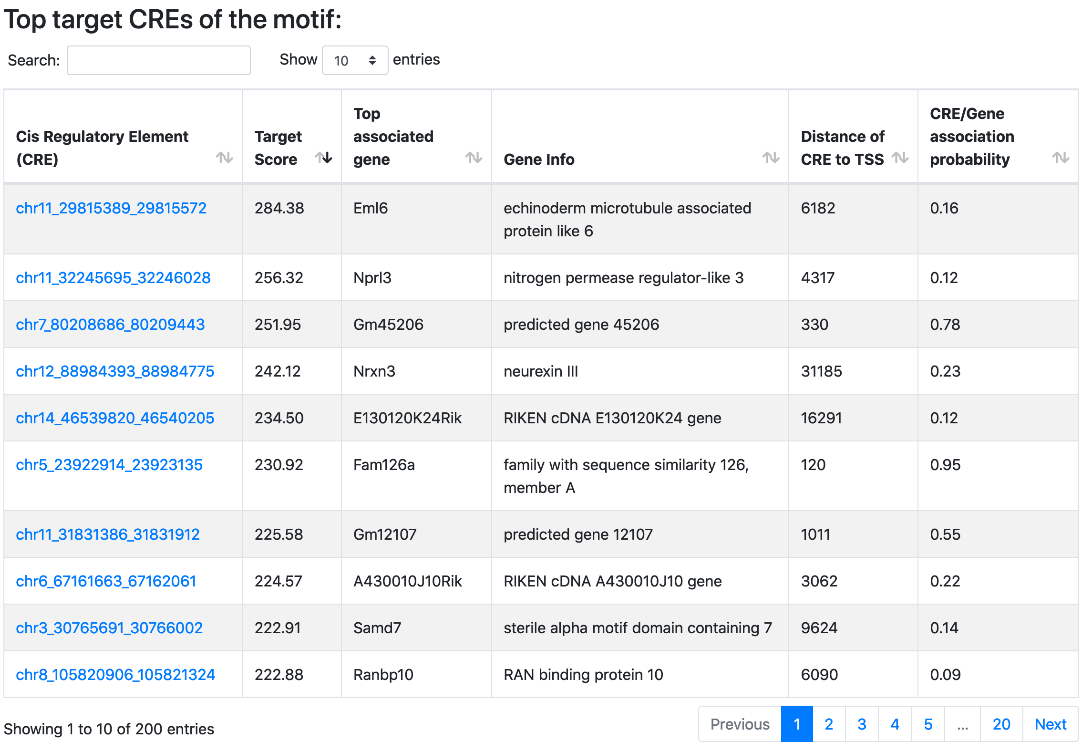
.



Conditions sorted by the z-value of the Tal1 motif activity

Move your cursor over a bar to see sample name and corresponding Z-value.

Figure S5: We compute, for this motif, a zScore for each sample. The barplot shows the zScore and hence the importance of this motif across all samples. The exact value of zScores is given by sliding the mouse over the bars. The image is a screenshot from `crema.unibas.ch`.



Top target CREs of the motif:

Cis Regulatory Element (CRE)	Target Score	Top associated gene	Gene Info	Distance of CRE to TSS	CRE/Gene association probability
chr11_29815389_29815572	284.38	Eml6	echinoderm microtubule associated protein like 6	6182	0.16
chr11_32245695_32246028	256.32	Nprl3	nitrogen permease regulator-like 3	4317	0.12
chr7_80208686_80209443	251.95	Gm45206	predicted gene 45206	330	0.78
chr12_88984393_88984775	242.12	Nrxn3	neurexin III	31185	0.23
chr14_46539820_46540205	234.50	E130120K24Rik	RIKEN cDNA E130120K24 gene	16291	0.12
chr5_23922914_23923135	230.92	Fam126a	family with sequence similarity 126, member A	120	0.95
chr11_31831386_31831912	225.58	Gm12107	predicted gene 12107	1011	0.55
chr6_67161663_67162061	224.57	A430010J10Rik	RIKEN cDNA A430010J10 gene	3062	0.22
chr3_30765691_30766002	222.91	Samd7	sterile alpha motif domain containing 7	9624	0.14
chr8_105820906_105821324	222.88	Ranbp10	RAN binding protein 10	6090	0.09

Showing 1 to 10 of 200 entries

Figure S6: A list of all inferred target CREs. We annotate those CREs with the gene yielding the highest CRE/gene association probability. The target CREs are sorted by the target CRE score. Further information of this associated gene (Gene Info) as well as the distance from the CRE to its promoter is indicated. Users sort the table according to each column by clicking on the arrows on top. The image is a screenshot from `crema.unibas.ch`.

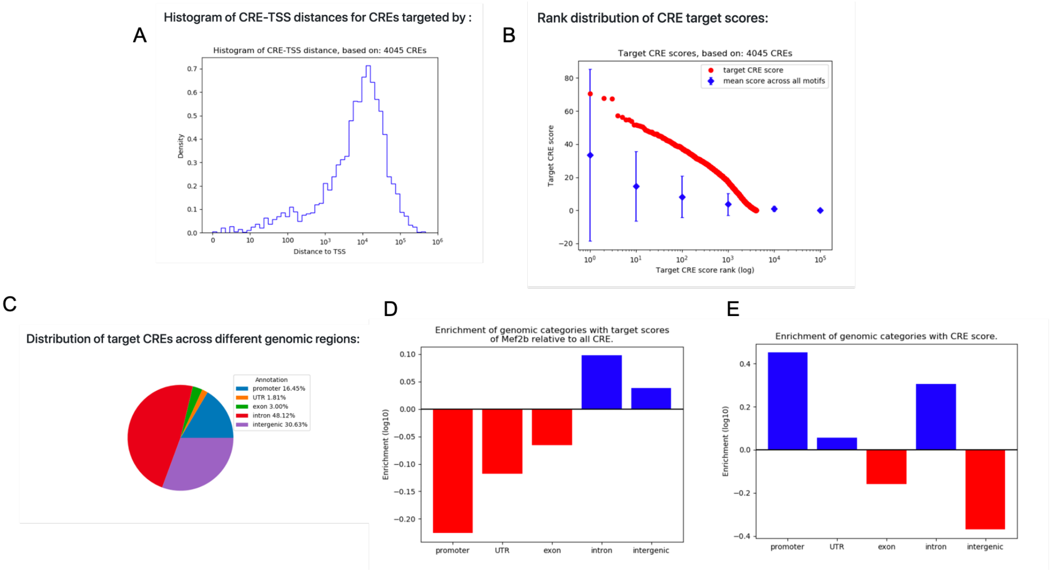
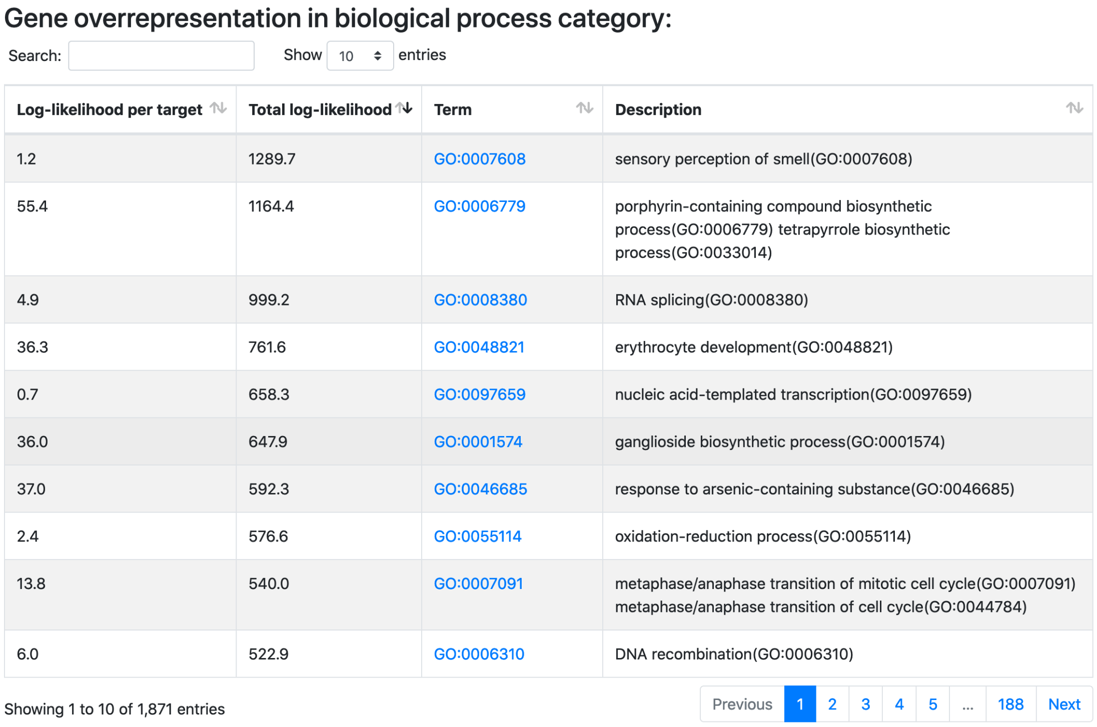


Figure S7: For further insight into the location of CREs targeted by a specific motif (in this case, Mef2b): A) histogram of distances of the CRE to the closest associated gene. The CREs is weighted with the inferred target score. B) Target scores for a specific motif, in comparison to all motifs (blue bars). CREs are weighted by their target score. C) Piechart of locations that the CREs map to: UTR, intron, CDS, promoter and intergenic regions. D) Enrichment of genomic categories with target scores the selected motif relative to all CREs E) Enrichment of genomic categories with CRE score. The image is composed of screenshots from `crema.unibas.ch`.



Gene overrepresentation in biological process category:

Search: Show 10 entries

Log-likelihood per target	Total log-likelihood	Term	Description
1.2	1289.7	GO:0007608	sensory perception of smell(GO:0007608)
55.4	1164.4	GO:0006779	porphyrin-containing compound biosynthetic process(GO:0006779) tetrapyrrole biosynthetic process(GO:0033014)
4.9	999.2	GO:0008380	RNA splicing(GO:0008380)
36.3	761.6	GO:0048821	erythrocyte development(GO:0048821)
0.7	658.3	GO:0097659	nucleic acid-templated transcription(GO:0097659)
36.0	647.9	GO:0001574	ganglioside biosynthetic process(GO:0001574)
37.0	592.3	GO:0046685	response to arsenic-containing substance(GO:0046685)
2.4	576.6	GO:0055114	oxidation-reduction process(GO:0055114)
13.8	540.0	GO:0007091	metaphase/anaphase transition of mitotic cell cycle(GO:0007091) metaphase/anaphase transition of cell cycle(GO:0044784)
6.0	522.9	GO:0006310	DNA recombination(GO:0006310)

Showing 1 to 10 of 1,871 entries

Figure S8: On all associated target genes, we calculate gene ontology categories. Users can choose by themselves how to sort the table, either on the log-likelihood per target or the total log-likelihood. The categories are sorted by their total enrichment score. We also include the log likelihood per target which shows the fold enrichment compared to random selection into the table. The image is a screenshot from `crema.unibas.ch`.

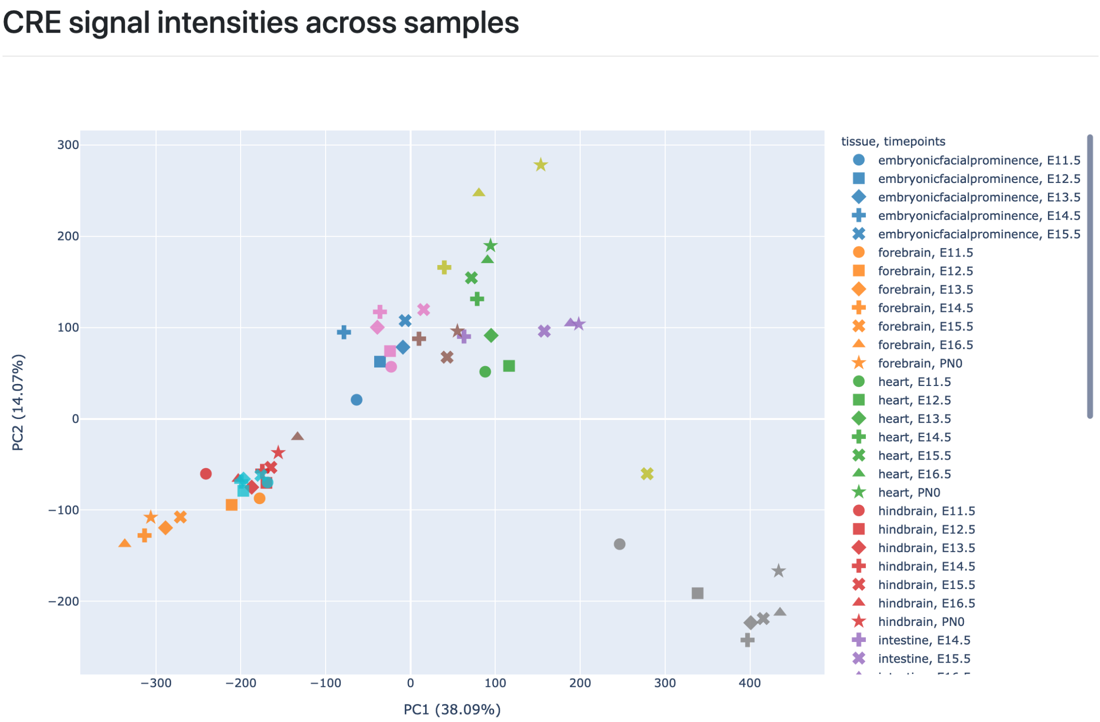

CRE signal intensities across samples

Figure S9: As part of the pipeline we show a PCA plot for principal components 1-4 (shown are PC1 and PC2) of the motif activities. The image is a screenshot from `crema.unibas.ch`.

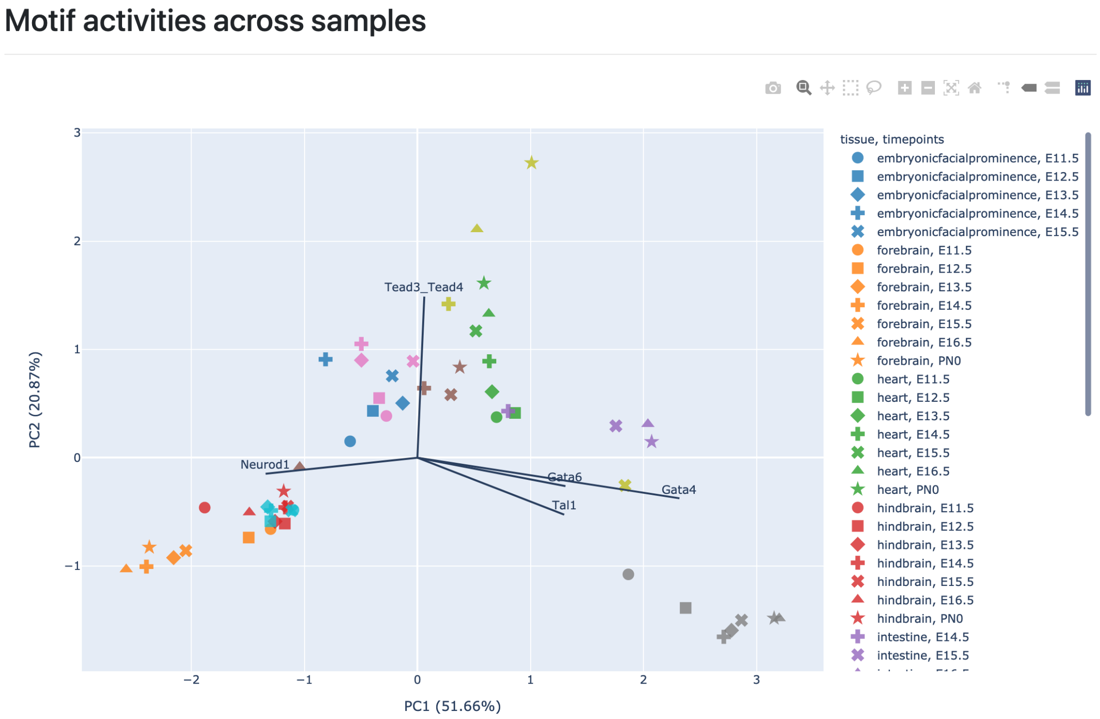

Motif activities across samples

Figure S10: As part of the pipeline we show a PCA plot for principal components 1-4 (shown are PC1 and PC2) of the motif activities. The image is a screenshot from `crema.unibas.ch`.

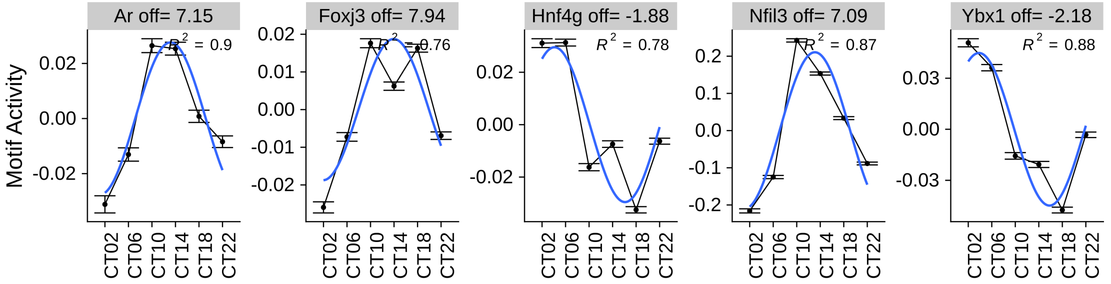


Figure S11: Supplementary Information for dataset 1: Motif activity plots for motifs indicated in 4.2D. Offset (off) is given in hours on top of each plot.

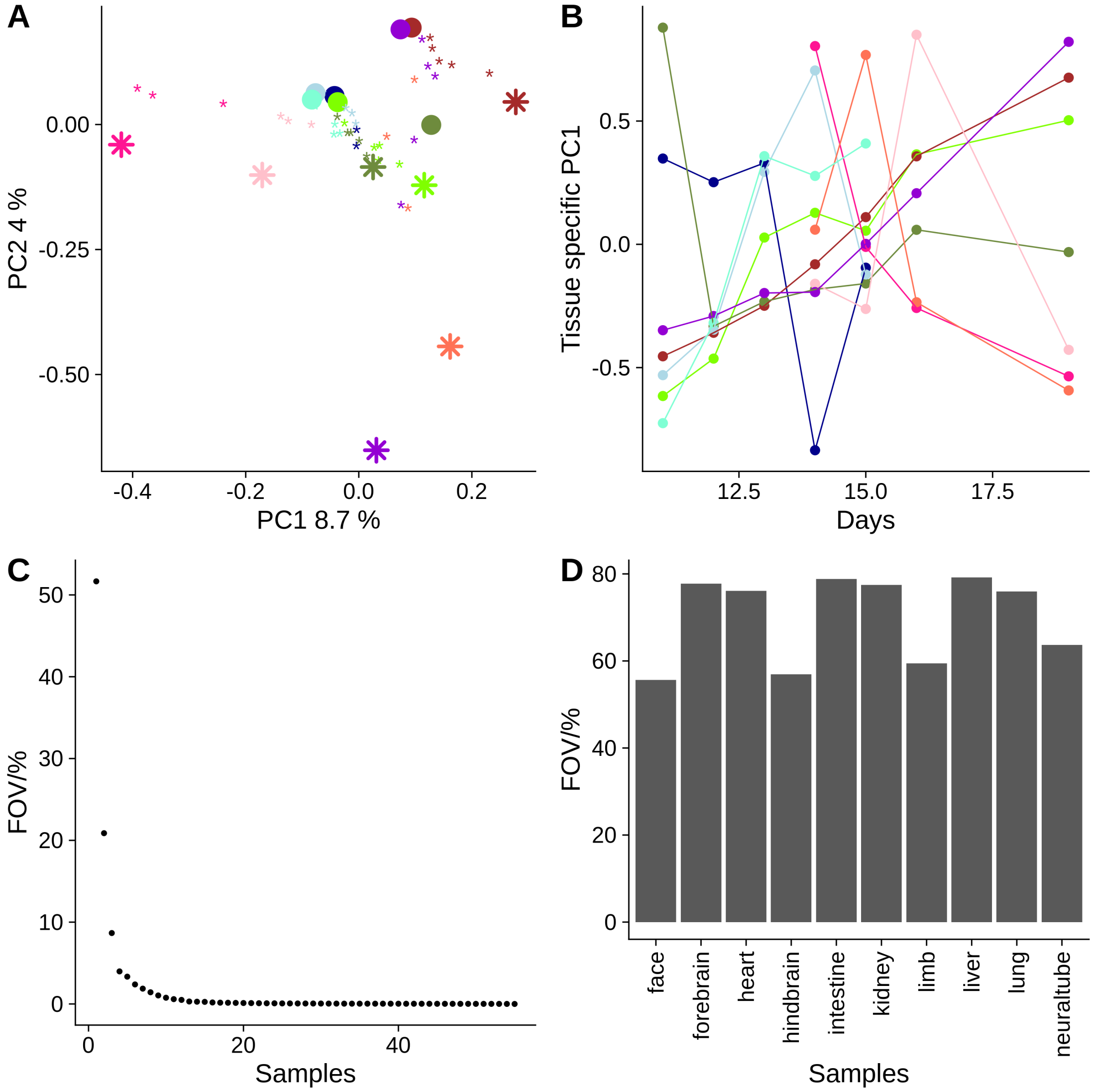


Figure S12: Supplementary information for dataset 2. A) PC3 and PC4 including all samples. B) Tissue - specific PC1 including all samples. C) Fraction of explained variance for the PCA shown A and in 4.3B, D. D) Fraction of explained variance when performing SVD on separate subsets of the activity table for each tissue.

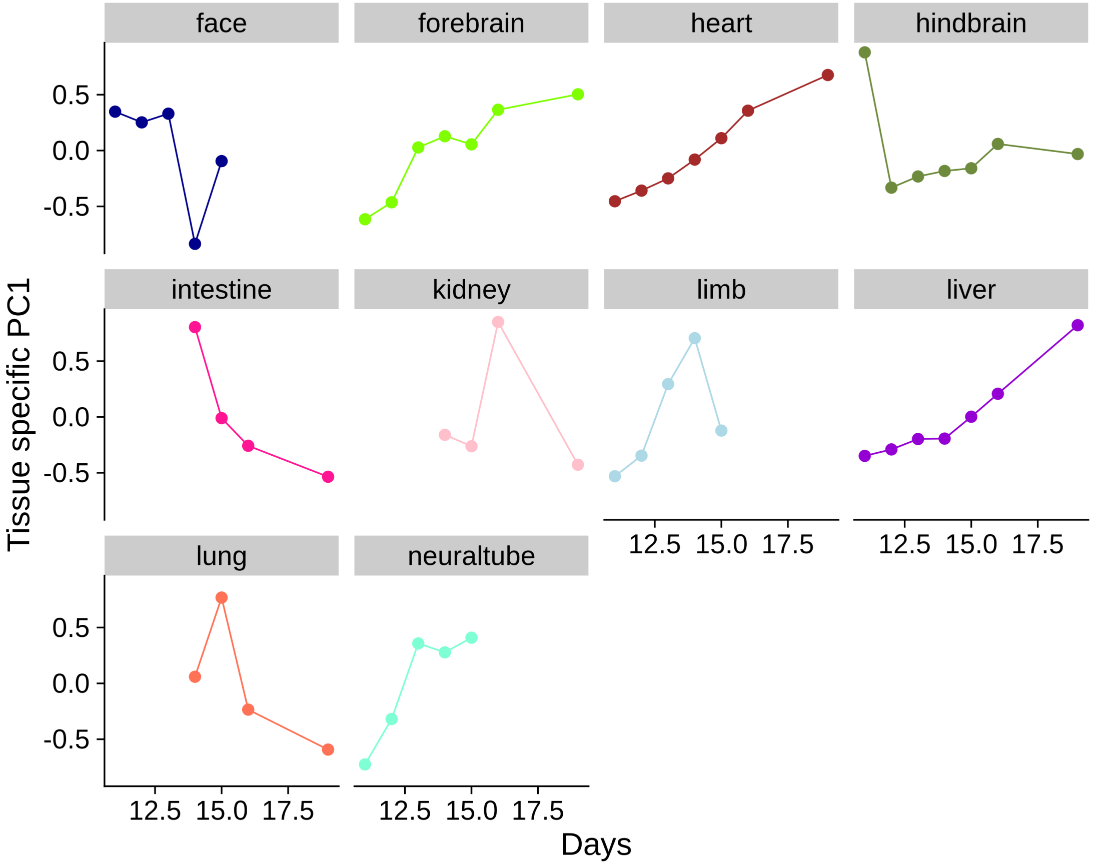


Figure S13: Supplementary information for dataset 2. First Component for all tissues, across measured timepoints

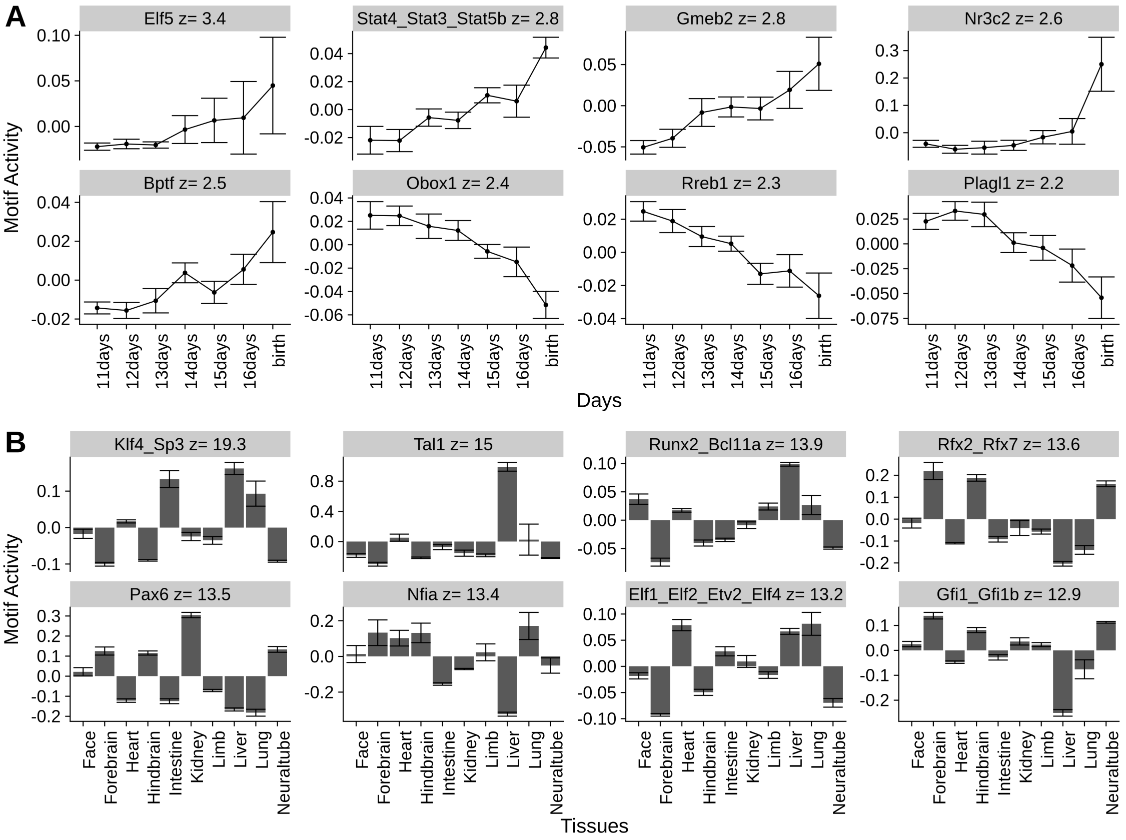


Figure S14: Supplementary information for dataset 2: Averaging over time and tissues. We averaged the MARA run in two ways: 1) average across all tissues for each timepoint 2) average across all timepoints for each tissue A) time-specific motifs: Activity profiles of top 8 motifs averaged across all tissues for each timepoint. B) tissue-specific motifs: Motif activities averaged across all timepoints for each motif.

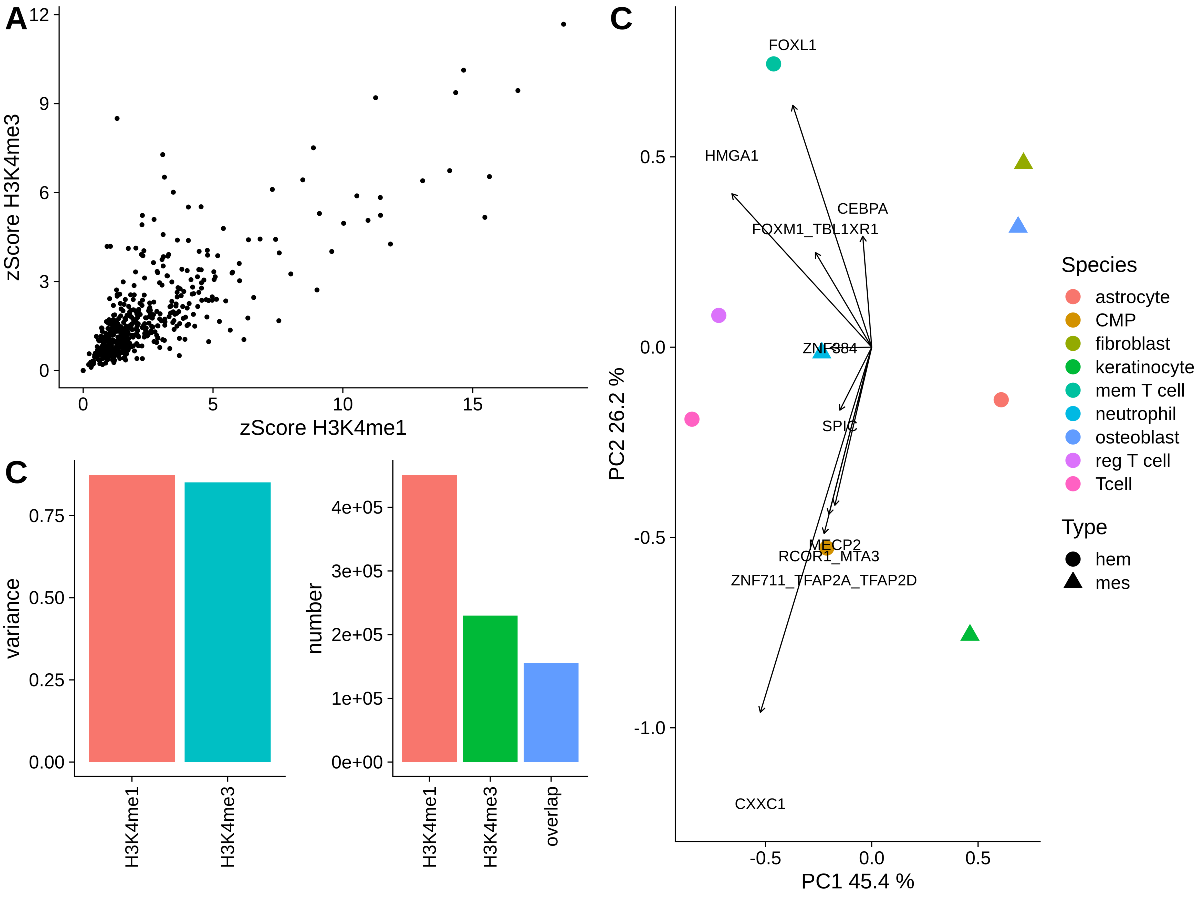


Figure S15: **Analysis of Histone modifications in primary cells shows separation of lineages and higher variability in enhancer regions**. A) Correlation of zScores of motif activity derived by CREMA analysis of ChIP-seq data for H3K4me1 and H3K4me3. B) *left*: Sum of variance in activity across all motifs in the H3K4me3 and H3K4me1 analysis. *right*: Number of significant CREs found uniquely for H3K4me1 and H3K4me3 and overlapping CREs. As long as one basepair was overlapping, the CRE was counted as overlapping. C) PCA on the activity matrix for H3K4me1. Plotted vectors were chosen to have the highest projection on PC1 and/or belong to one of the top significant motifs yielding highest zScores.

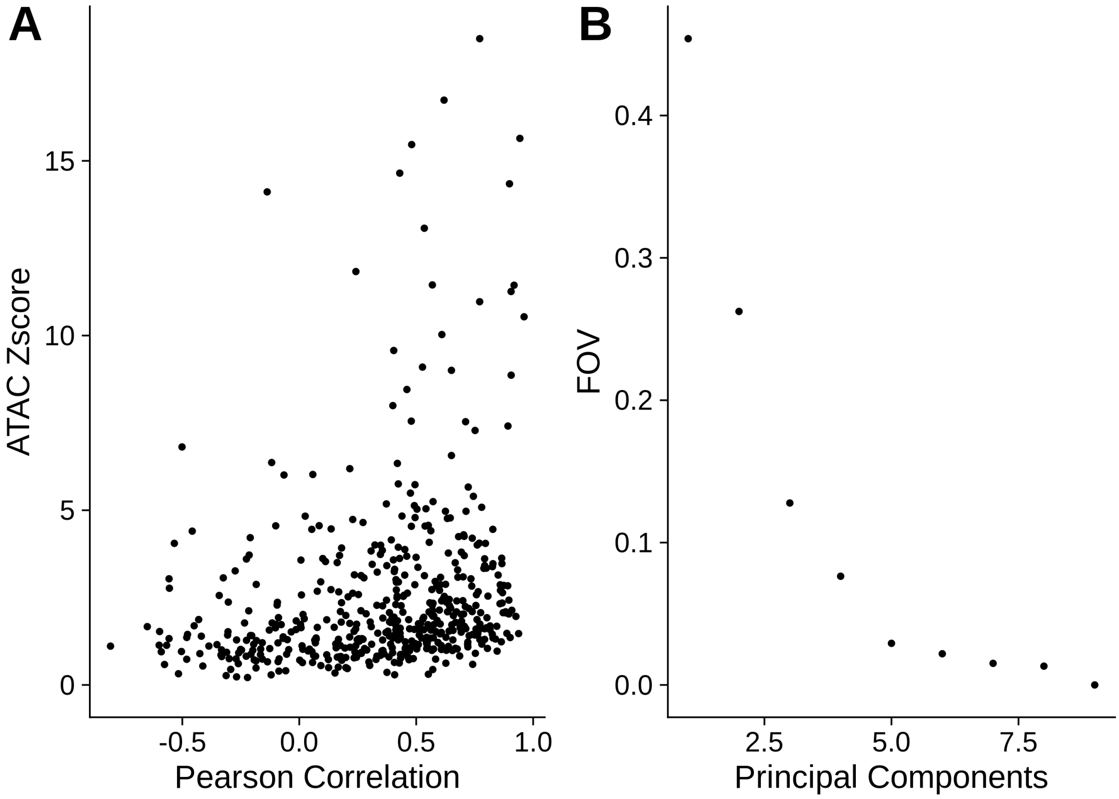


Figure S16: Supplementary information for dataset 3. A) Pearson correlation coefficient vs. ATAC zScore. B) Fraction of explained variance in the SVD shown in Figure S15,C.

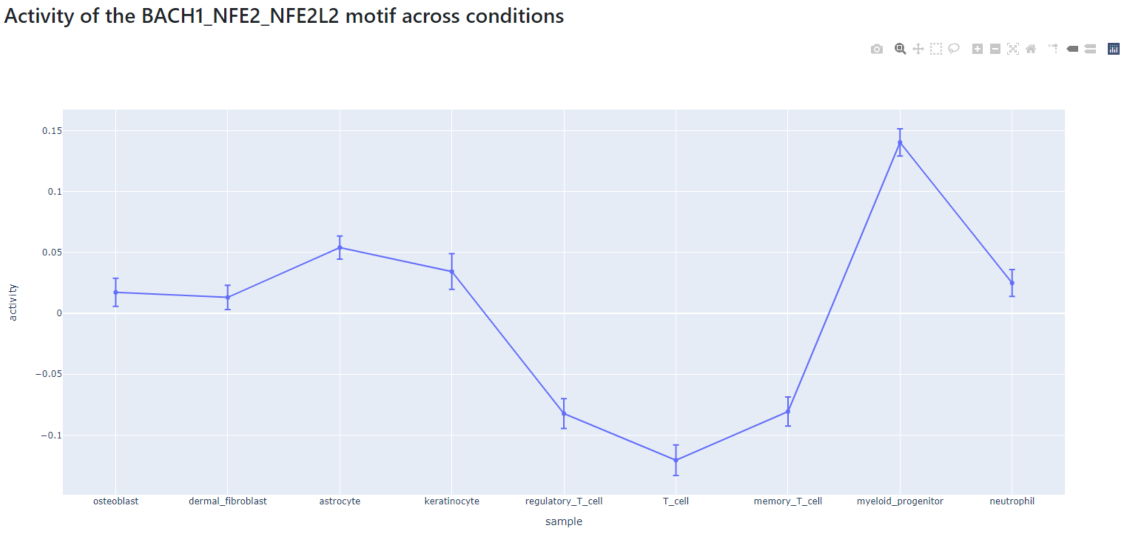


Figure S17: Supplementary information for dataset 3. Activity Pattern for BACH1_NFE2_NFE2L2 across all conditions, taken from the example results at `crema.unibas.ch`. The activity peaks at myeoloid progenitor (CMP).

Dataset	Accession
Dataset 1 accession:	DNase seq of murine liver of mice left in darkness for 50-70h (by John Stamatoyannopoulos, UW) ENCSR904DTN
Dataset 2 accessions:	ATAC seq for 11 different tissues and at 4-7 different timepoints from 11.5 days until birth in the embryos in mice (Bing Ren, UCSD) ENCSR012YAB, ENCSR023QZX, ENCSR031HDN, ENCSR032HKE, ENCSR068YGC, ENCSR079GOY, ENCSR088UYE, ENCSR096JCC, ENCSR102NGD, ENCSR150EOO, ENCSR150RMQ, ENCSR154BXN, ENCSR176BYZ, ENCSR204ZTY, ENCSR211OCS, ENCSR217NOA, ENCSR255XTC, ENCSR261ICG, ENCSR273UFV, ENCSR282YTE, ENCSR302LIV, ENCSR310MLB, ENCSR312LQX, ENCSR335VJW, ENCSR343TXK, ENCSR358MOW, ENCSR371KFW, ENCSR377YDY, ENCSR382RUC, ENCSR384JBF, ENCSR389CLN, ENCSR451NAE, ENCSR460BUL, ENCSR465PYP, ENCSR468GUI, ENCSR486XAS, ENCSR551WBK, ENCSR552ABC, ENCSR559FAJ, ENCSR603MWL, ENCSR609OHJ, ENCSR618HDK, ENCSR623GSD, ENCSR627OCR, ENCSR652CNN, ENCSR662KNY, ENCSR668EIA, ENCSR690VOH, ENCSR700QBR, ENCSR732OTZ, ENCSR758IRM, ENCSR785NEL, ENCSR798FDL, ENCSR810HQR, ENCSR819QOJ, ENCSR820ACB, ENCSR836PUC, ENCSR876SYO, ENCSR896XIN, ENCSR903GMO, ENCSR961SMM, ENCSR966ORC, ENCSR976LWP, ENCSR983JWA
Dataset 2.1 accessions:	RNA seq for 11 different tissues and at 4-8 different timepoints from E10.5 - birth in the embryos in mice (Barbara Wold, Caltech) ENCSR004XCU, ENCSR017JEG, ENCSR020DGG, ENCSR039ADS, ENCSR049UJU, ENCSR062VTB, ENCSR080EVZ, ENCSR096STK, ENCSR115TWD, ENCSR150CUE, ENCSR160IIN, ENCSR173PJN, ENCSR185LWM, ENCSR216NEG, ENCSR284AMY, ENCSR284YKY, ENCSR285WZV, ENCSR304RDL, ENCSR307BCA, ENCSR331XCE, ENCSR337FYI, ENCSR343YLB, ENCSR347SQR, ENCSR362AIZ, ENCSR367ZPZ, ENCSR370SFB, ENCSR401BSG, ENCSR420QTO, ENCSR448MXQ, ENCSR457RRW, ENCSR504GEG, ENCSR508GWZ, ENCSR526SEX, ENCSR537GNQ, ENCSR538WYL, ENCSR541XZK, ENCSR557RMA, ENCSR559TRB, ENCSR579FCW, ENCSR597UZW, ENCSR611PTP, ENCSR636CWO, ENCSR647QBV, ENCSR648YEP, ENCSR667TOX, ENCSR691OPQ, ENCSR719NAJ, ENCSR727FHP, ENCSR750YSX, ENCSR752RGN, ENCSR760TOE, ENCSR764OPZ, ENCSR792RJV, ENCSR809VYL, ENCSR823VEE, ENCSR826HIQ, ENCSR830IVQ, ENCSR848GST, ENCSR848HOX, ENCSR851HEC,

	ENCSR867YNV, ENCSR906YQZ, ENCSR908JWT, ENCSR921PRX, ENCSR928OXI, ENCSR932TRU, ENCSR943LKA, ENCSR946HWC, ENCSR968QHO, ENCSR970EWM, ENCSR982MRY, ENCSR992WBR
Dataset 3	ChIP seq for H3K4me1 and H3K4me3 across 9 primary cell lines in cultured human cells.
	(by Bradley Bernstein, Broad Institute)
accessions:	ENCSR000ALI, ENCSR000AOT, ENCSR000APJ, ENCSR000ARV, ENCSR170NCG, ENCSR324EFP, ENCSR523BMU, ENCSR586POT, ENCSR660WQO, ENCSR777RWW, ENCSR826VJY, ENCSR887ESB, ENCSR911BCA

Table S1: Accession list for the datasets. Datasets are linked in addition on the webpage `crema.unibas.ch`. Accession list for the backgound datasets to calculate the pooled background for mouse, data is given in a supplementary table.

Motif	Reference
Arnt_Tfe3_Mlx_Mitf_Mlxip_Tfec_Egr1	https://doi.org/10.1371/journal.pbio.1000595
Egr1	https://doi.org/10.1038/srep15212
Elf1_Elf2_Etv2_Elf4	
Esr2	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5393318/
Etv1_Etv5_Gabpa	
Foxi1_Foxo1	https://doi.org/10.1073/pnas.0701599104
Foxp1_Foxj2	
Hcfc1_Six5_Smarcc2_Zfp143	
Hnf1b	https://doi.org/10.1172/JCI96138
Hnf4a	https://doi.org/10.1172/JCI96138
Hnf4g	https://doi.org/10.1172/JCI96138
Irf2_Irf1_Irf8_Irf9_Irf7	
Irx6_Irx2_Irx3	https://www.nature.com/articles/s41598-019-52215-4
Mecp2	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4395115/
Mnt	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6910219/
Nfia	
Nfil3_Tef	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5702268/
Nr2f6	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4066320/
Rorc_Nr1d1	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750502/
Stat5a	
Tcf7_Tcf7l2	
Tfeb_Usf1_Srebf1_Usf2_Bhlhe41_Srebf2	https://doi:10.7150/jca.13748
Ybx1_Nfya_Nfyb_Nfyc_Cebpz	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5351623/

Table S2: The kown/novel circadian regulators which are indicated in (Figure 4.2,C). Motifs were assigned as "known" when they could be associated with circadian regulation either being a regulator or being regulated.

face

Decreasing	Increasing
Tal1	Tead1
Nr2f1_Nr4a1	Nrf1
Nr1h4	Rfx5
Vdr	Gmeb2
Pparg_Rxrg	Atf4
Klf1	Elf5
Tbr1	Zbtb33_Chd2
Hoxc9	Myog_Tcf12
Zfp423	Rela_Rel_Nfkb1

forebrain

Decreasing	Increasing
Esr2	Nfatc2
Gata5	Meis1
Tead3_Tead4	Nfia
Nr1h4	Cdc5l
Pitx1	Nkx1.1_Nkx1.2
Hoxb13	Yy1_Yy2
Hnf4a	Tfdp1_Wt1_Egr2
Zbtb18	Hcfc1_Six5_Smar
Erg	Foxp2_Foxp3

heart

Decreasing	Increasing
Gata3	Mef2b
Snai1_Zeb1_Snai2	Srf
Foxk1_Foxj1	Tead3_Tead4
Pou1f1	Mef2c
Hmga2	Stat4_Stat3_Stat5b
Foxa3	Hes1
Nr1i3	Gmeb2
Zfp784	Mnt
Onecut1_Cux2	Hsfy2

hindbrain

Decreasing	Increasing
Mef2b	Hoxa11_Hoxc12
Hsf2	Zbtb18
Maf_Nrl	Nr1h4
Nfatc2	Zbtb12
Hey2	Pax1_Pax9
Pou6f2_Pou4f2	Hmbox1
Rela_Rel_Nfkb1	Rxra
Gata2_Gata1	Onecut1_Cux2
Vsx2_Dlx3	Snai1_Zeb1_Snai

intestine

Decreasing	Increasing
Pou2f1	Thrb
Ppara	Rarg
Hmga1	Vsx1_Uncx_Prrx2_Shox2_Noto
Glis2	T
Prop1	Nr1i2
Sox2	Gmeb1
Max_Mycn	Atoh1_Bhlhe23
Msx2_Hoxd4	Barhl2
Hoxa1	Nr2f1_Nr4a1

kidney

Decreasing	Increasing
Prdm14	Gsx1_Alx1_Mixl
Obox3	Neurod1
Irf5_Irf6	Bsx
Hoxa7_Hoxc8	Sox14
Hoxd11_Cdx1_Hoxc11	Barhl2
Sin3a	Zkscan1
Mef2b	Gli3_Zic1
Hinfp	Hoxb2_Dlx2
Runx2_Bcl11a	Ddit3

limb

Decreasing	Increasing
Nr1h4	Taf1
Rarg	Nrf1
Lhx2_Hoxc5	Tfap4
Tal1	Irf2_Irf1_Irf8_Irf9_Irf7

liver

Decreasing	Increasing
Nkx6.1_Evx1_Hesx1	Spic
Foxp2_Foxp3	Cebpa_Cebpg
Mecom	Cebpe
Zfp691	Hlf

Stat1	Tcf3	Cebpd	Nr1h4
Nr4a2	Tcf21_Msc	Dmc1	Pparg_Rxrg
Nkx2.5	Neurod1	Prox1	Ets1
Esr1	Nfatc2	Etv3_Erf_Fev_Elk4_Elk1_Elk3	Spib
Gata5	Ascl2	Nkx3.2	Onecut1_Cux2
lung		textbfneural tube	
neg. correlation	pos. correlation	Decreasing	Increasing
Onecut1_Cux2	Tead1	Tlx1	Nfe2l2
Klf8	Tead3_Tead4	Mecp2	Stat4_Stat3_Sta
Smad2	Ehf	Sox3_Sox10	Rest
Figla	Tfcp2	Tcf7_Tcf7l2	Foxo4
Mybl2	Ppara	Rreb1	Hcfc1_Six5_S
Irx6_Irx2_Irx3	Hsf2	Ovol1	Nr4a3
Six6	Sox13	Hbp1	Mga
Nkx1.1_Nkx1.2	Prdm1	Sox17	Lhx8
Gata6	Grhl1	Pbx2	Dlx5_Dlx4

Table S3: Motifs correlating positively or negatively strongest with the first principal component for each tissue separately: their activity pattern is either increasing or decreasing with time. (see Figure S13)We use the subset for each tissue of the activity matrix to calculate the correlation to the corresponding PC1. Top ten motifs, according to their correlation value (negative and positive), are selected.

Name	Averaging Configuration
Mesenchymal-Hematopoietic	astrocyte, keratinocyte, fibroblast of dermis, osteoblast against reg T cell, T cell mem T cell, common myeloid progenitor, neutrophil.
osteoblast	osteoblast against keratinocyte, fibroblast of dermis and astrocyte
astrocyte	astrocyte against keratinocyte, fibroblast of dermis and osteoblast
fibroblast	fibroblast against keratinocyte, osteoblast and astrocyte
keratinocyte	keratinocyte against fibroblast of dermis, osteoblast and astrocyte
CLP-CMP	regulatory T cell, mem Tcell and T cell against CMP and neutrophil
reg T cell	reg T cell against T cell
neutrophil	neutrophil against CMP

Table S4: Averaging configuration for dataset 3.

GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
TGGAATTCTCGGGTGCCAAGG
GATCGGAAGAGCACACGTCTG
TCGTATGCCGTCTTCTGCTTG
CAAGCAGAAGACGGCATACGAGAT
AATGATACGGCGACCACCGAGATCTACAC
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

Table S5: Illumina Adapter sequences, selected from Illumina Adapter Sequences [34]

BIBLIOGRAPHY

[1] Phil Arnold, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen. Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics*, 28(4):487–494, 2011.

[2] Phil Arnold, Anne Schöler, Mikhail Pachkov, Piotr Balwierz, Helle Jørgensen, Michael B Stadler, Erik van Nimwegen, and Dirk Schübeler. Modeling of epigenome dynamics identifies transcription factors that mediate polycomb targeting. *Genome research*, pages gr–142661, 2012.

[3] Roberto Avellino and Ruud Delwel. Expression and regulation of c/ebpα in normal myelopoiesis and in malignant transformation. *Blood, The Journal of the American Society of Hematology*, 129(15):2083–2091, 2017.

[4] Pietro Baldelli and Jacopo Meldolesi. The transcription repressor rest in adult neurons: physiology, pathology, and diseases. *ENeuro*, 2(4), 2015.

[5] Piotr J Balwierz, Piero Carninci, Carsten O Daub, Jun Kawai, Yoshihide Hayashizaki, Werner Van Belle, Christian Beisel, and Erik van Nimwegen. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. *Genome biology*, 10(7):R79, 2009.

[6] Piotr J Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J Gruber, Mihaela Zavolan, and Erik van Nimwegen. Ismara: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome research*, 24(5):869–884, 2014.

[7] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823 – 837, 2007.

[8] J Berger and A Bird. Role of mbd2 in gene regulation and tumorigenesis, 2005.

[9] Severin Berger, Saeed Omidi, Mikhail Pachkov, Phil Arnold, Nicholas Kelley, Silvia Salatino, and Erik van Nimwegen. Crunch: Integrated processing and modeling of chip-seq data in terms of regulatory motifs. *bioRxiv*, 2018.

[10] Severin Berger, Mikhail Pachkov, Phil Arnold, Saeed Omidi, Nicholas Kelley, Silvia Salatino, and Erik van Nimwegen. Crunch: integrated processing and modeling of ChIP-seq data in terms of regulatory motifs. *Genome Research*, 29(7):1164–1177, July 2019.

[11] Eva Beuling, Nana Yaa A Baffour-Awuah, Kelly A Stapleton, Boaz E Aronson, Taeko K Noah, Noah F Shroyer, Stephen A Duncan, James C Fleet, and Stephen D Krasinski. Gata factors regulate proliferation, differentiation, and gene expression in small intestine of mature mice. *Gastroenterology*, 140(4):1219–1229, 2011.

[12] Gregory D Bowman and Michael G Poirier. Post-translational modifications of histones that influence nucleosome dynamics. *Chemical reviews*, 115(6):2274–2295, 2015.

[13] Sophie E Broughton, Timothy R Hercus, Matthew P Hardy, Barbara J McClure, Tracy L Nero, Mara Dottore, Huy Huynh, Hal Braley, Emma F Barry, Winnie L Kan, et al. Dual mechanism of interleukin-3 receptor blockade by an anti-cancer antibody. *Cell reports*, 8(2):410–419, 2014.

[14] Jeroen Bussmann, Jeroen Bakkers, and Stefan Schulte-Merker. Early endocardial morphogenesis requires scl/tal1. *PLoS genetics*, 3(8):e140, 2007.

[15] Wenqiang Cao, Jing Guo, Xiaofeng Wen, Li Miao, Feng Lin, Guanxin Xu, Ruoyu Ma, Shengxia Yin, Zhaoyuan Hui, Tingting Chen, et al. Cxxc finger protein 1 is critical for t-cell intrathymic development through regulating h3k4 trimethylation. *Nature communications*, 7(1):1–11, 2016.

[16] Tainã Figueiredo Cardoso, Raquel Quintanilla, Anna Castelló, Emilio Mármol-Sánchez, Maria Ballester, Jordi Jordana, and Marcel Amills. Analysing the expression of eight clock genes in five tissues from fasting and fed sows. *Frontiers in genetics*, 9:475, 2018.

[17] Josefa Castillo, Gerardo López-Rodas, and Luis Franco. Histone post-translational modifications and nucleosome organisation in transcriptional regulation: some open questions. In *Protein Reviews*, pages 65–92. Springer, 2017.

[18] Ali Z Chaudhry, Gary E Lyons, and Richard M Gronostajski. Expression patterns of the four nuclear factor i genes during mouse embryogenesis indicate a potential role in development. *Developmental dynamics: an official publication of the American Association of Anatomists*, 208(3):313–325, 1997.

[19] Jeannette Connerney, Dana Lau-Corona, Andy Rampersaud, and David J Waxman. Activation of male liver chromatin accessibility and stat5-dependent gene transcription by plasma growth hormone pulses. *Endocrinology*, 158(5):1386–1405, 2017.

[20] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[21] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2017.

[22] Benjamin Deneen, Ritchie Ho, Agnes Lukaszewicz, Christian J Hochstim, Richard M Gronostajski, and David J Anderson. The transcription factor nfia controls the onset of gliogenesis in the developing spinal cord. *Neuron*, 52(6):953–968, 2006.

[23] Cody Desjardins and Francisco Naya. The function of the mef2 family of transcription factors in cardiac development, cardiogenomics, and direct reprogramming. *Journal of cardiovascular development and disease*, 3(3):26, 2016.

[24] Brandt L Esplin, Tomoyuki Shimazu, Robert S Welner, Karla P Garrett, Lei Nie, Qingzhao Zhang, Mary Beth Humphrey, Qi Yang, Lisa A Borghesi, and Paul W Kincade. Chronic exposure to a tlr ligand injures hematopoietic stem cells. *The Journal of Immunology*, 186(9):5367–5375, 2011.

[25] Ben J. Greenwell, Alexandra J. Trott, Joshua R. Beytebiere, Shanny Pao, Alexander Bosley, Erin Beach, Patrick Finegan, Christopher Hernandez, and Jerome S. Menet. Rhythmic food intake drives rhythmic gene expression more potently than the hepatic circadian clock in mice. *Cell Reports*, 27(3):649 – 657.e5, 2019.

[26] Matthew G Guenther, Stuart S Levine, Laurie A Boyer, Rudolf Jaenisch, and Richard A Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007.

[27] Jorge Guzman-Lepe, Eduardo Cervantes-Alvarez, Alexandra Collin de l'Hortet, Yang Wang, Wendy M Mars, Yoshinao Oda, Yuki Bekki, Masahiro Shimokawa, Huanlin Wang, Tomoharu Yoshizumi, et al. Liver-enriched transcription factor expression relates to chronic hepatic failure in humans. *Hepatology communications*, 2(5):582–594, 2018.

[28] Carl Haasper, Michael Jagodzinski, Maren Drescher, Rupert Meller, Michael Wehmeier, Christian Krettek, and Eric Hesse. Cyclic strain induces fosb and initiates osteogenic differentiation of mesenchymal cells. *Experimental and Toxicologic Pathology*, 59(6):355–363, 2008.

[29] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–318, 2007.

[30] Brian Hendrich, Jacqueline Guy, Bernard Ramsahoye, Valerie A Wilson, and Adrian Bird. Closely related proteins mbd2 and mbd3 play distinctive but interacting roles in mouse development. *Genes & development*, 15(6):710–723, 2001.

[31] Walter Herzog. The multiple roles of titin in muscle contraction and force production. *Biophysical reviews*, 10(4):1187–1199, 2018.

[32] Biao Hu, Mehrnaz Gharaee-Kermani, Zhe Wu, and Sem H Phan. Essential role of mecp2 in the regulation of myofibroblast differentiation during pulmonary fibrosis. *The American journal of pathology*, 178(4):1500–1508, 2011.

[33] Ryosuke Ikeda, Yoshiki Tsuchiya, Nobuya Koike, Yasuhiro Umemura, Hitoshi Inokawa, Ryutaro Ono, Maho Inoue, Yuh Sasawaki, Tess Grieten, Naoki Okubo, Kazuya Ikoma, Hiroyoshi Fujiwara, Toshikazu Kubo, and Kazuhiro Yagita. REV-ERB and REV-ERB function as key factors regulating Mammalian Circadian Output. *Scientific Reports*, 9(1):10171, December 2019.

[34] Illumina. Adapter sequences. `https://eurofinsgenomics.eu/media/1610545/illumina-adapter-sequences.pdf`, 2016.

[35] Ari Itoh-Nakadai, Mitsuyo Matsumoto, Hiroki Kato, Junichi Sasaki, Yukihiro Uehara, Yuki Sato, Risa Ebina-Shibuya, Mizuho Morooka, Ryo Funayama, Keiko

Nakayama, et al. A bach2-cebp gene regulatory network for the commitment of multipotent hematopoietic progenitors. *Cell reports*, 18(10):2401–2414, 2017.

[36] Richard Kin Ting Kam, Yi Deng, Yonglong Chen, and Hui Zhao. Retinoic acid synthesis and functions in early embryonic development. *Cell & bioscience*, 2(1):1–14, 2012.

[37] Megan Keniry, Robert K. Dearth, Michael Persans, and Ramon Parsons. New Frontiers for the NFIL3 bZIP Transcription Factor in Cancer, Metabolism and Beyond. *Discoveries*, 2(2):e15, June 2014.

[38] Chang H Kim. Foxp3 and its role in the immune system. In *Forkhead Transcription Factors*, pages 17–29. Springer, 2009.

[39] Rasmus Kjøbsted, Janne R Hingst, Joachim Fentz, Marc Foretz, Maria-Nieves Sanz, Christian Pehmøller, Michael Shum, André Marette, Remi Mounier, Jonas T Treebak, et al. Ampk in skeletal muscle function and metabolism. *The FASEB journal*, 32(4):1741–1777, 2018.

[40] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[41] JY Lee and TL Orr-Weaver. Chromatin. *Encyclopedia of Genetics*, 2001.

[42] Ana Rita Lourenço, M Guy Roukens, Danielle Seinstra, Cynthia L Frederiks, Cornelieke E Pals, Stephin J Vervoort, Andreia S Margarido, Jacco van Rheenen, and Paul J Coffer. C/ebpa is crucial determinant of epithelial maintenance by preventing epithelial-to-mesenchymal transition. *Nature communications*, 11(1):1–18, 2020.

[43] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, 389(6648):251, 1997.

[44] Raphaëlle Luisier, Elif B. Unterberger, Jay I. Goodman, Michael Schwarz, Jonathan Moggs, Rémi Terranova, and Erik van Nimwegen. Computational modeling identifies key gene regulatory interactions underlying phenobarbital-mediated tumor promotion. *Nucleic Acids Research*, 42(7):4180–4195, apr 2014.

[45] Dario Magnani, Laurette Morlé, Kerstin Hasenpusch-Theil, Marie Paschaki, Monique Jacoby, Stéphane Schurmans, Bénédicte Durand, and Thomas Theil. The ciliogenic transcription factor rfx3 is required for the formation of the thalamocortical tract by regulating the patterning of prethalamus and ventral telencephalon. *Human molecular genetics*, 24(9):2578–2593, 2015.

[46] Ruslan Medzhitov and Tiffany Horng. Transcriptional control of the inflammatory response. *Nature Reviews Immunology*, 9(10):692–703, 2009.

[47] Anna V Molofsky, Robert Krenick, Erik Ullian, Hui-hsin Tsai, Benjamin Deneen, William D Richardson, Ben A Barres, and David H Rowitch. Astrocytes and disease: a neurodevelopmental perspective. *Genes & development*, 26(9):891–907, 2012.

[48] Masahito Nagaki and Hisataka Moriwaki. Transcription factor hnf and hepatocyte differentiation. *Hepatology Research*, 38(10):961–969, 2008.

[49] Mikhail Pachkov, Piotr J Balwierz, Phil Arnold, Andreas J Gruber, Mihaela Zavolan, and Erik van Nimwegen. Ismara: Completely automated inference of gene regulatory networks from high-throughput data. *PeerJ Preprints*, 5:e3328v1, 2017.

[50] Graziella Pellegrini, Elena Dellambra, Osvaldo Golisano, Enrica Martinelli, Ivana Fantozzi, Sergio Bondanza, Diego Ponzin, Frank McKeon, and Michele De Luca. p63 identifies keratinocyte stem cells. *Proceedings of the national academy of sciences*, 98(6):3156–3161, 2001.

[51] Joaquín Pérez-Schindler, Serge Summermatter, Silvia Salatino, Francesco Zorzato, Markus Beer, Piotr J Balwierz, Erik van Nimwegen, Jérôme N Feige, Johan Auwerx, and Christoph Handschin. The corepressor ncor1 antagonizes pgc-1α and estrogen-related receptor α in the regulation of skeletal muscle function and oxidative metabolism. *Molecular and cellular biology*, 32(24):4913–4924, 2012.

[52] Abraham B Roos and Magnus Nord. The emerging role of c/ebps in glucocorticoid signaling: lessons from the lung. *Journal of Endocrinology*, 212(3):291–305, 2012.

[53] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology*, 33(11):1165, 2015.

[54] Veit Rothhammer, Davis M. Borucki, Emily C. Tjon, Maisa C. Takenaka, Chun-Cheih Chao, Alberto Ardura-Fabregat, Kalil Alves de Lima, Cristina Gutiérrez-Vázquez, Patrick Hewson, Ori Staszewski, Manon Blain, Luke Healy, Tradite Neziraj, Matilde Borio, Michael Wheeler, Loic Lionel Dragin, David A. Laplaud, Jack Antel, Jorge Ivan Alvarez, Marco Prinz, and Francisco J. Quintana. Microglial control of astrocytes in response to microbial metabolites. *Nature*, 557(7707):724–728, may 2018.

[55] Stefan Schoenfelder and Peter Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, page 1, 2019.

[56] Mona D Shahbazian, Juan I Young, Lisa A Yuva-Paylor, Corinne M Spencer, Barbara A Antalffy, Jeffrey L Noebels, Dawna L Armstrong, Richard Paylor, and Huda Y Zoghbi. Mice with truncated mecp2 recapitulate many rett syndrome features and display hyperacetylation of histone h3. *Neuron*, 35(2):243–254, 2002.

[57] C Stirzaker, JZ Song, W Ng, Q Du, NJ Armstrong, WJ Locke, AL Statham, H French, R Pidsley, F Valdes-Mora, et al. Methyl-cpg-binding protein mbd2 plays a key role in maintenance and spread of dna methylation at cpg islands and shores in cancer. *Oncogene*, 36(10):1328, 2017.

[58] Debora Sugiaman-Trapman, Morana Vitezic, Eeva-Mari Jouhilahti, Anthony Mathelier, Gilbert Lauter, Sougat Misra, Carsten O Daub, Juha Kere, and Peter Swoboda. Characterization of the human rfx transcription factor family by regulatory and target gene analysis. *BMC genomics*, 19(1):181, 2018.

[59] Weiwei Tao, Jing Wu, Qian Zhang, Shan-Shan Lai, Shan Jiang, Chen Jiang, Ying Xu, Bin Xue, Jie Du, and Chao-Jun Li. EGR1 regulates hepatic clock gene amplitude by activating Per1 transcription. *Scientific Reports*, 5(1):15212, December 2015.

[60] Jean-Michel Terme, Sébastien Lemaire, Didier Auboeuf, Vincent Mocquet, and Pierre Jalinot. The proto-oncogenic protein tal1 controls tgf-β1 signaling through interaction with smad3. *Biochimie open*, 2:69–78, 2016.

[61] Yi Tian, Zhengcai Jia, Jun Wang, Zemin Huang, Jun Tang, Yanhua Zheng, Yan Tang, Qinghong Wang, Zhiqiang Tian, Di Yang, et al. Global mapping of h3k4me1 and h3k4me3 reveals the chromatin state-based cell type-specific gene regulation in human treg cells. *PloS one*, 6(11), 2011.

[62] Alexandra J Trott and Jerome S Menet. Regulation of circadian clock transcriptional output by clock: Bmal1. *PLoS genetics*, 14(1):e1007156, 2018.

[63] Alexandra J. Trott and Jerome S. Menet. Regulation of circadian clock transcriptional output by CLOCK:BMAL1. *PLOS Genetics*, 14(1):e1007156, January 2018.

[64] Amy B. Truong and Paul A. Khavari. Control of Keratinocyte Proliferation and Differentiation by p63. *Cell Cycle*, 6(3):295–299, feb 2007.

[65] Ghanshyam Upadhyay, Asif H Chowdhury, Bharat Vaidyanathan, David Kim, and Shireen Saleque. Antagonistic actions of rcor proteins regulate lsd1 activity and cellular differentiation. *Proceedings of the National Academy of Sciences*, 111(22):8071–8076, 2014.

[66] ER Vagapova, PV Spirin, TD Lebedev, and VS Prassolov. The role of tal1 in hematopoiesis and leukemogenesis. *Acta naturae*, 10(1):15, 2018.

[67] Erik van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC bioinformatics*, 8(S6):S4, 2007.

[68] Alex Vassilev, Kotaro J Kaneko, Hongjun Shu, Yingming Zhao, and Melvin L DePamphilis. Tead/tef transcription factors utilize the activation domain of yap65, a src/yes-associated protein localized in the cytoplasm. *Genes & development*, 15(10):1229–1241, 2001.

[69] Stephin J Vervoort, Ana Rita Lourenço, Ruben van Boxtel, and Paul J Coffer. Sox4 mediates tgf-β-induced expression of mesenchymal markers during mammary cell epithelial to mesenchymal transition. *PloS one*, 8(1), 2013.

[70] Ping Wang, Zhonghui Tang, Byoungkoo Lee, Jacqueline Jufen Zhu, Liuyang Cai, Przemyslaw Szalaj, Simon Zhongyuan Tian, Meizhen Zheng, Dariusz Plewczynski, Xiaoan Ruan, et al. Chromatin topology reorganization and transcription repression by pml-rarα in acute promyeloid leukemia. *Genome biology*, 21:1–21, 2020.

[71] Yuhao Wang, Zheng Kuang, Xiaofei Yu, Kelly A. Ruhn, Masato Kubo, and Lora V. Hooper. The intestinal microbiota regulates body composition through NFIL3 and the circadian clock. *Science*, 357(6354):912–916, September 2017.

[72] Hongtan Wu, Yubo Xiao, Shihao Zhang, Suyuan Ji, Luyao Wei, Fuqin Fan, Jing Geng, Jing Tian, Xiufeng Sun, Funiu Qin, et al. The ets transcription factor gabp is a component of the hippo pathway essential for growth and antioxidant defense. *Cell reports*, 3(5):1663–1677, 2013.

[73] Heng Zhang, Chen-Ying Liu, Zheng-Yu Zha, Bin Zhao, Jun Yao, Shimin Zhao, Yue Xiong, Qun-Ying Lei, and Kun-Liang Guan. Tead transcription factors mediate the function of taz in cell growth and epithelial-mesenchymal transition. *Journal of biological chemistry*, 284(20):13355–13362, 2009.

[74] Rena Zheng, Boris Rebolledo-Jaramillo, Yiwei Zong, Liqing Wang, Pierre Russo, Wayne Hancock, Ben Z Stanger, Ross C Hardison, and Gerd A Blobel. Function of gata factors in the adult mouse liver. *PLoS One*, 8(12), 2013.

[75] Yonghao Zhong, Hongyang Huang, Min Chen, Jinzhou Huang, Qingxia Wu, Guang-Rong Yan, and De Chen. Pou2f1 over-expression correlates with poor prognoses and promotes cell growth and epithelial-to-mesenchymal transition in hepatocellular carcinoma. *Oncotarget*, 8(27):44082, 2017.

5 ANALYSIS OF ACUTE EXERCISE RESPONSE USING SVD AND ISMARA

The lack of sufficient physical activity is one of the main causes of death worldwide. In actual fact, the medicine here is cheap and easy: regular exercise has been shown to improve and prevent most of the modern-society diseases and is also beneficial in chronic and age related diseases. One of the nodal regulators of exercise is Peroxisome proliferator-activated receptor gamma coactivator 1-alpha Pgc1α, a co-activator investigated thoroughly during the last decades. Here, we provide computational predictions of the gene regulatory network induced by an acute bout of exercise in mouse, which – in part – depends on Pgc1α: We sequenced the quadriceps of wild type and Pgc1α muscle knock out mice which were killed immediately, 4h, 6h, and 8h after an acute bout of exercise. Then we combined computational modeling to predict important regulators from RNAseq data with singular varlue decomposition to disentangle most important regulators and their activities during this post exercise time-course. We find that early immediate genes play an important role in kicking off the exercise response, followed by downstream autophagy and immune response pathways. Interestingly we find Pgc1α being not mandatory for this immediate response, however, in the long run wild type mice benefit from a sustained high level of Pgc1α induced gene programs.

Anne Krämer[1,2], Regula Furrer[1], Erik van Nimwegen[2], Christoph Handschin[1]

[1] University of Basel, Biozentrum, Basel, Switzerland

[2] Swiss Institute for Bioinformatics, Basel, Switzerland

5.1 INTRODUCTION

EXERCISE Individuals in our society often experience lifestyle-associated diseases such as heart disease and stroke, obesity and type 2 diabetes (T2D) and hypertension. One of the main causes of those diseases is the lack of appropriate amount of physical activity. About 1/3 of adults and even 4/5 of teens worldwide do not exercise enough, although regular training has been shown to dramatically improve and even prevent lifestyle-related diseases, other chronic diseases and age-related muscle waisting [32, 33].

For instance, just a short bout of exercise reversed the symptoms of metabolic disease [2]. In T2D and sarcopenia, exercise and a healthy diet outperforms any pharmacological treatment [3, 4]. In recent years, lots of efforts have been undertaken to disentangle the complex gene regulatory network induced by acute and chronic exercise, but we still lack full understanding of all the connections and regulators.

PGC1α One of the nodal regulators of exercise has been investigated for several years: Peroxisome proliferator-activated receptor gamma coactivator 1-alpha (Pgc1α). This co-activator is implied in multiple cellular processes, for instance in adaptive thermogenesis, fatty acid oxidation, gluconeogenesis and mitochondrial biogenesis [9] and the adaptation to endurance exercise. Pgc1α mRNA is found to be highly elevated after bouts of endurance exercise in rats and humans [6, 7]. Additionally, it is higher expressed in slow, oxidative fibers than in fast twitch fibers [5]. Only overexpression of Pgc1α in skeletal muscle is enough to improve exercise performance by improving mitochondria and mediating the switch of fast to slow twitch muscles [5, 8].

A single exercise bout changes the expression level of a myriad of genes, all tightly regulated by a network of transcription factors. These factors can enable or disable transcription by modifying chromatin structure and allocating the transcriptional machinery to the regulatory site. As the number of factors is far less than the number of responding genes, it reduces the dimensionality of the data tremendously if we assume that the observed gene expression is a result of transcription factor activity. We then can express the global gene expression in terms of regulators.

In this study, we aimed to elucidate the gene regulatory interactions following acute exercise and how they depend on Pgc1α. We sequenced the RNA of murine quadriceps muscle at different timepoints after the mice completed an acute exhaustion exercise protocol. This was equally done with two groups of mice, one wild type (WT) and one lacking Pgc1α in the muscle (KO). This provided us with an extensive dataset to investigate the temporal sequence of molecular mechanisms happening after an acute bout of exercise. As exercise is a strong inducer of gene expression, a multitude of genes was found to be differentially expressed directly and hours after the bout.

The key challenge thus was to explain the observed gene expression changes in terms of a few important regulators. Further, we needed to associate the factors with their potential target genes. This was done by using Motif Activity Response Analysis (ISMARA) [5]. ISMARA uses sophisticated statistical methods to infer the regulatory activity of transcription factors taking into account the measured gene expression pattern and predicted regulatory binding sites in promoters of the measured genes. ISMARA has been applied with great success in a number of studies previously [10, 54].

Here we used ISMARA combined with singular value decomposition (SVD) to investigate the regulatory mechanisms in exercise response. We were able to identify key regulatory factors responsible for the induction of stress response, growth related mechanisms and identify metabolic pathways as major difference between the genotypes. Our findings may lead to a better understanding of the temporal pattern of transcription factor activity and gene expression in exercise response.

5.2 RESULTS

TRANSCRIPTIONAL CHANGES FOLLOWING ACUTE EXERCISE To disentangle the contribution of Pgc1α in acute exercise response, we used WT and KO mice. Both groups performed an exercise protocol until exhaustion and were sacrificed at 0h, 4h, 6h and 8h after the exercise, with the control being sacrificed subsequently throughout 0h-8h time-course to account for eventual circadian fluctuations in gene expression. Tissue was taken from the quadriceps muscle, RNA extracted and sequenced (Figure 1A and Methods).

We performed differential gene expression (DE) analysis to find the differences be-

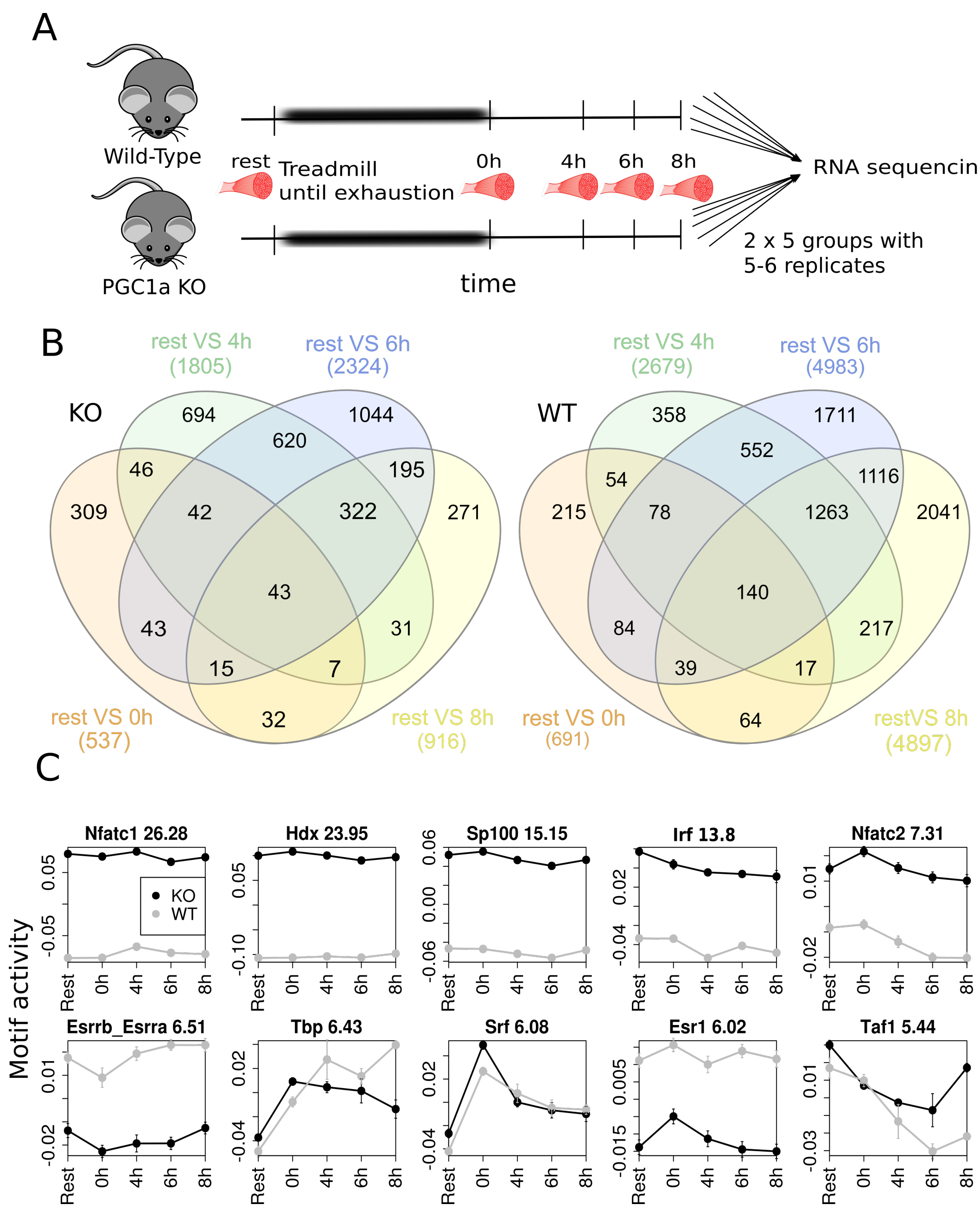


Figure 1: **The genotype difference dominates the ISMARA model across WT and KO animals.** A) Experimental set up B) Venn diagram of differential expressed genes for knockout (KO) and wild type (WT) animals. Comparison was always the mice which didn't perform exercise (rest). C) Top 10 motifs explaining the dataset, inferred by ISMARA. Numbers above the plots are the corresponding zScores

tween exercised and resting mice for all timepoints in both groups. The resulting gene expression changes were vast, in the WT group up to 5000 (FDR < 0.05) genes changed and in almost the whole Pgc1α mice, the number of DE genes exceeded 1000. Interestingly, the common set of genes which depicts all genes once expressed differentially compared to the WT mice is rather small, only 43 for the KO mice and 140 for the WT animals.

What could already be observed is that the number of total DE genes dropped in the

KO mice after the 6h timepoint. This indicates that most of the genes are induced at the beginning or directly after the exercise. However, when comparing just the genotypes, the largest difference in DE expressed genes is found at the 8h timepoint (see Figure 1B and S1). This could be explained by the long duration of transcription of the Pgc1α gene: It takes, for the gene around 5h to be transcribed into mRNA (taken into account the current release of NCBI its location is $chr5 : 51454249 - 52115853$, thus the length of $661604bp$ and a speed of the polymerase of $34bp/sec$ [14]). This means that Pgc1α was induced by the exercise bout, but it takes until $5h$ after the exercise that the mRNA is transcribed fully and even longer until the protein is translated and the targets are expressed. That is, where changes due to the difference in genotypes then become more apparent on the transcriptomic level.

DETECTION OF MOST IMPORTANT REGULATORS DRIVING THE VARIANCE IN THE WHOLE DATASET As it is very challenging to find important regulatory genes from this extensive set of DE genes, we used ISMARA on the whole dataset to infer transcription factor activities using the observed gene expression data and information on binding sites in the promoters. Notably, ISMARA infers the activity of a transcriptional regulator, which is not necessarily correlated with its mRNA expression. This is very useful as rapid gene induction often relies on ubiquitously expressed factors which only need to be activated, for instance by phosphorylation or translocation to the nucleus. ISMARA estimated the significance of inferred motifs by rigorously ranking the inferred transcriptional regulators then according to their variation across samples and consistency across target promoters (see zScore in Methods).

KNOWN MOTIFS ASSOCIATED PREVIOULSY WITH EXERCISE RESPONSE AND / OR PGC1α

We ran ISMARA on the whole dataset, including KO and WT animals. Figure 1C shows the 6 top ranked motifs and their corresponding zScores. Obviously, the difference in genotypes dominates the activity changes. Interestingly, while most of the pattern show higher activity in the KO animals, Errα (Esrrb,Esrra) is more active in the WT. We find several factors which have already been found with respect to exercise or Pgcα:

Nfatc1 / Nfatc2 (Nuclear factor of activated T-cells) is the most important motif driving the difference in gene expression between the genotypes. A study [1] showed that reduced Nfat activity resulted in reduced expression of slow-twitch muscle gene expression. However, in our hands, Nfatc1 is more active in the KO animals, suggesting Pgc1α may be downstream of Nfatc1, hence if it is abundant, the activity of Nfatc1 may shut down.

Irf (Interferon regulatory) factors are known to be stress sensors and act in immune response but also in metabolism related diseases. Irf6 was found to interact with Pgc1α as well [51].

Esrra / Esrrb (Estrogen related receptor alpha) is a well studied partner of Pgc1α, also its activity seems to be higher in the WT animals. It has been associated with controlling whole body lactate levels in exercise when being co-activated by Pgc1α [54]. It's activity was found to be higher in the WT compared to the KO.

Srf (Serum response factor) is a factor known to be involved in acute stress response, stress in our case is the exercise. Immediately after (or even during) exercise it increases rapidly, inducing several pathways [16], which is reflected in its activity pattern.

Esr1 (Estrogen receptor alpha): Pgc1α has been shown to co-activate Esr1 and induce antioxidant genes [53]. As Esrra, it is more active in WT.

PREDICTED MOTIFS NOT YET ASSOCIATED WITH EXERCISE OR PGC1α Tbp (TATA binding protein) is a TATA-Box associated element which has not been implicated with the response to exercise or Pgc1α before.
Hdx (Highly divergent homeobox) transcription factor, its function is not known.
Taf1 (The TATA-box binding protein associated factor 1) is a key unit of the transcription factor II D complex that has not been investigated deeply. One study found it to serve a vital function in embryogenesis in zebrafish [52].
Sp100 is part of the Promyelocytic leukemia protein (Pml) complex and was not investigated with respect to exercise or Pgc1α. Our analysis suggested an elevated activity of SP100 in KO mice.

IDENTIFICATION OF GENERAL PATTERN OF MOTIF ACTIVITY IN EXERCISE RESPONSE When we had expressed all DE genes in terms of $\approx$ 500 regulators, we still wanted to assess whether there are groups of motifs which could induce others or respond in a similar manner to the stimulus. To perform this in a completely unbiased manner, we applied SVD (see Methods and Figure 2B) on the activity table we had obtained previously (Figure 2A). Around 70% of the variance was explained by the first 3 singular vectors and each of them explained more than 10% of the data, such that we proceed our analysis with the first three singular vectors S2. The first singular vector captured the highest amount of variance, (referred to as PC1) clearly distinguished between WT and knockout animals, whereas PC2 (the second singular vector) showed more variation in time

DETECTION OF MOTIF GROUPS FOLLOWING THE SAME DYNAMICS How exactly did the pattern look like and how much different pattern existed in the dataset? In figure 2C we plotted the values of the first three components at each timepoint as outlined in the Methods section. Top correlating motifs (pearson correlation value > 0.8) are sorted by zScore and plotted in Figure 2D. Note that SVD is invariant under point reflection, meaning that the inverse pattern of the one shown here is equally important. This component was thus representative for the top motifs already shown in 1C. However, the second and third component depicted the exercise response, interestingly, those pattern were very similar for both KO and WT animals. Motifs correlated with PC2 may be downstream targets of motifs in PC3. Notably, most of the motifs correlated to PC3 were genes involved in the early immediate gene response, like Srf and Jun Proto-Oncogene (Jun). We thus looked into the predicted targets of motifs following the pattern in PC3.

MOTIFS FOLLOWING THE SECOND SINGULAR VECTOR The second component captured a time-dependent pattern which is quite similar for both genotypes: Motifs which change slowly during and after exercise to come back after 8h to the original level.
Nfya, Nfyb, Nfyc (Nuclear transcription factor Y): Factors of this group were found previously to be implicated in the exercise response [23].

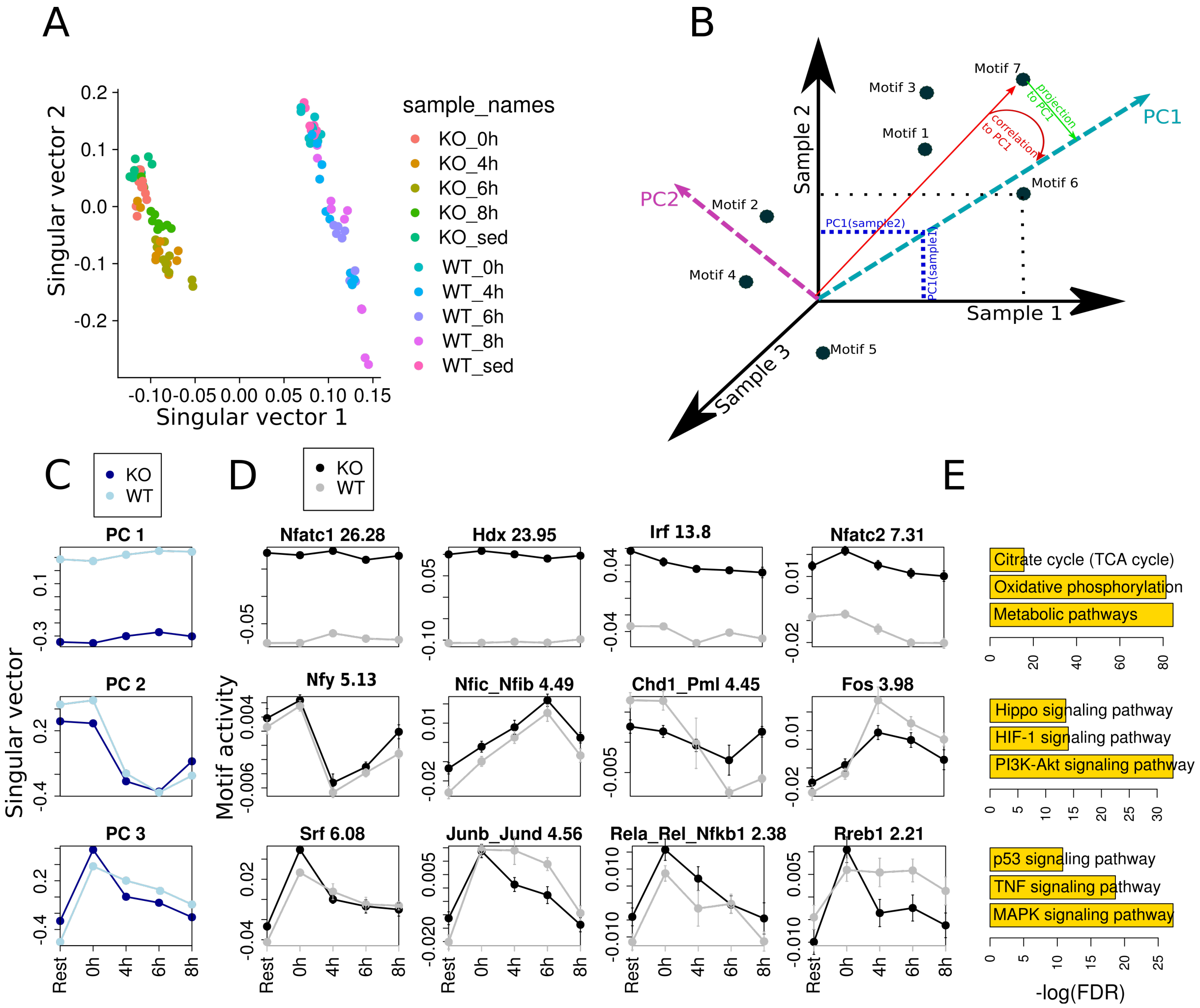


Figure 2: **Regulation of exercise response for wildtype (WT) and Pgc1α muscle knock out (KO) mice.** A) The first singular vector show clear separation of genotypes but also a time-induced pattern on singular vector 2. B) Schematic explanation of the , correlation and projection. C) The first three singular vectors plotted across time (blue and lightblue). Note that PCA is invariant under point reflection, which makes the presented pattern equally important to its inverse pattern. D) Motifs whose activity pattern correlates at least 80 % to the singular vectors. Sorted by the zScore inferred by ISMARA. E) KEGG categories associated with target genes of the motifs in each group characterized by the singular vectors.

Nfic Nfi (Nuclear Factor I)-factors are important factors in in development and differentiation of cells in the nervous system, lung and muscle [56].
Chd1, Pml (Chromodomain Helicase DNA Binding Protein 1) and Pml was shown to reduce acetylation of Pgc1α which increases its activity [24].
Fos (Fos Proto-Oncogene) is part of the Ap1 complex which has been associated with Pgc1α before [30]. It belongs to genes involved in the early immediate gene response.

MOTIFS FOLLOWING THE THIRD SINGULAR VECTOR PC3 showed an interesting pattern: A rapid increase in motif activity followed by a slower decrease. Here, we saw

that the response in KO animals is slightly dampened compared to the WT animals. The motifs behaving this way are:
<u>Srf</u> is known to be an activator of immediate early genes and its activity is induced rapidly in response to external stimuli [16]. It was found to be involved in many cellular programs, mostly related to muscle structure/function/generation and repair and cardiovascular development and maintenance, but also in a multitude of other organs [17].
<u>Junb and Jund</u> are part – as Fos too – of the Ap1 complex which has been found to be targeted by Pgc1α before [30]. Ap1, too, is responsible for the activation of early immediate genes, e.g. in response to cardiac hypertrophic stimuli [18, 19].
<u>Nfκb</u> (Nuclear factor κ B) plays critical roles in inflammation, cell proliferation, differentiation. Its activity relies on the degradation of its inhibitors (kappa B proteins). It has been implicated with binding to Pgc1α to inhibit its function in cardiac pathological processes [20].
<u>Rreb1</u> is associated with Ras-signaling and cancer, there has not been any profound association with exercise or even Pgc1α yet.

CLUSTERING THE MOTIFS FOR FUNCTIONAL ANNOTATION We grouped all motifs which correlate positively or negatively more than 80% to each singular vector, and extracted their top 20 target genes. This yielded 3 groups of genes which followed specific pattern. The advantage of using SVD to extract the clusters rather than using common clustering methods was that the principal components are mutually independent because they build an orthonormal basis. To get further information about the biological functions of the gene groups, we performed Gene Ontology (GO) for biological process and KEGG functional analysis. ISMARA reports, for each motif, promoters which contain a site for this motif (=target). The targets were sorted according to their target score (see Methods). We refer to the genes of targeted promoters by motif m from now on as target genes of motif m. We took the target genes of all motifs which correlate more than 80% to the singular vectors extracted the genes that were predicted to be regulated by these motifs. This yielded 3 groups of genes which each were regulated in the same manner. Figure 2E shows an excerpt of the top ten KEGG pathways.

METABOLIC PATHWAYS The first component depicted the genotype difference. Our target genes in the first group associated with KEGG pathways related to metabolic terms. In the GO analysis, we also observed terms related to energy derivation, cellular respiration and metabolic pathways (Figure S4). Pgc1α has been found previously to play a decisive role in numerous metabolic processes, e.g. oxidative phosphorylation was associated with Pgc1α before, as was shown in a study where overexpression of Pgc1α rescued OXPHOS in mitochondrial DNA (mtDNA) - defective cells [35] and several genes of the tricarboxylic acid (TCA) cycle, mitochondrial fatty acid beta oxidation and the krebs cycle were upregulated by Pgc1α [36, 37]. Notably, the expression of mitochondrial genes is mediated by co-activation of Errα and Nuclear receptor factor (Nrf1/2) by Pgc1α, the former was ranked among the top motifs in our ISMARA analysis.

TISSUE REGENERATION AND GROWTH The second component comprised motifs which get activated with and after the 4h timepoint. KEGG analysis yielded categories related to growth of the cell, like the PI3K-Akt-, HIF- or Hippo signaling pathway. The PI3K-Akt pathway for instance was directly activated following exercise [38]. Although it is known that Pgc1α plays a direct role in the Akt pathway, as it is phosphorylated and deactivated by Akt upon exposure to insulin, we did not see a significant difference between WT and KO mice in most pattern belonging to the second group of motifs. Especially the Hippo pathway and its implication in exercise adaptation has been investigated, as it is known to promote tissue growth. It has been postulated that the Hippo pathway plays a role in proliferation and renewal of myogenic cells [40], further it has even been proposed to be important for muscular adaptation i.e. hypertrophy to resistance exercise [39]. Hypoxia-inducible factor 1-alpha (Hif-1) enables the expression of genes involved in the hypoxia response of most mammalian cells and, Pgc1α has been found to stabilize Hif-1 and hence upregulate its target genes [41].

STRESS RESPONSE The third group contained genes targeted by the fastest responding motifs in our analyis. Among these is Srf, which is a known inducer of early immediate gene response and targets the promoters of several other factors involved in the rapid response to external stimuli. The corresponding KEGG pathway analysis revealed mainly stress-related and immune response terms such as Mitogen-activated protein kinase (Mapk) and p53. Top terms in the GO analysis corresponded to gene programs in response to compounds or external input (see Figure S4). Tumor protein p53 (p53) has been associated with exercise before and was found to eventually counteract cancer by the activation of tumor protein p21 (p21), Insulin like growth factor binding protein 3 (Igfbp-3), and Phosphatase and tensin homolog (Pten) [42]. The TNF pathway was suggested to provoke insulin resistance and dyslipidemia. Exercise in turn was observed to downregulate one of the key factors in this pathway, Tumor necrosis factor alpha (TNFα), which resulted in anti-inflammatory effects. [43] The Mapk pathway has been associated with growth and development [44] and is activated by cytokines, growth factors and cellular stress [45]. It is found also in the KEGG pathway analysis of the second group, however in the third group it is the top pathway, suggesting that it is turned on quickly but retains its activity for a longer period of time after the exercise bout. Taking all together, the first singular vector was responsible for the large amount of variance in our data and captured the genotypic differences and induces metabolic pathways. The second and third singular vectors denoted the time-dependent response to exercise (e.g. stress and immune response) which was only partly altered by the presence of Pgc1α.

RECONSTRUCTION OF GENE REGULATORY PATHWAYS IN EARLY RESPONSE TO EXERCISE USING ISMARA'S TARGET PREDICTIONS Looking at the component pattern with respect to the time-course nature of the data, we imagined that motifs whose activity follows PC3 could be downstream targets of motifs following pattern PC2. (see Figure S3 for an exemplary pathway inferred by motif activities and target genes). GO and KEGG analysis suggested that genes involved in stress response are upstream inducers of genes related to growth response in the second group. As one target gene can have multiple transcripts and thus promoters, we summed up expression pattern of transcripts belonging to the same gene. Srf was one of the top factors in our analysis whose activity pattern showed rather a time-dependent than a genotype dependent behavior. We thus start off with the construction of our network with Srf (Figure 3A). Interestingly, opposed to its highly significant activity profile, Srf mRNA was not changed during the time-course (see Figure S3), suggesting that it was ubiquitously expressed, waiting to be activated by an external stimuli. That confirms that Srf is indeed a regulator required to kick off genetic programs in the response to perturbations.

We checked the first top target genes which ISMARA predicted to be regulated by Srf: Egr1/2, Fos and Fosb are transcription factors themselves and their mRNA transcription and are rapidly initiated. However, transcription takes its time and the target genes of Egr1, Fos and Fosb were expressed time-delayed.

Figure 3C shows the expression level of the top ranking target genes for each of the motifs in Figure 3B. Targets shown here were manually chosen according to their ranking by target score and their expression level had to reach $\log_2(tpm) = 2$ at least.

Egr1 was denoted as the most important target of Srf. Among its targets were the Heat shock factor Hspa1b, which has been associated with exercise response earlier, its expression was found to correlate with the exercise intensity in rat soleus [46]. Interestingly, also Kruppel like factor 4 (Klf4) which was predicted to be downstream of Egr2 targets another heat shock protein - DnaJ Heat Shock Protein Family (Hsp40) Member B1 (Dnajb1). Fos1 is predicted to induce Ankrd1 and 1700101l11Rik, a lnc RNA close to GABA Type A Receptor Associated Protein Like 1 (Gabarpl1), which is associated with autophagy. Fosb itself activates presumably Irf5, which leads to the activation of immune-related pathways.

INFLUENCE OF THE GENOTYPE IN THE RESPONSE TO EXERCISE Concerning the differences in genotypes, we found that the rapid acute stress response is not affected by the lack of Pgc1α. This suggests that Pgc1α was not implicated directly in the early immediate gene response. However, it was found before that the lack of Pgc1α resulted in an overall reduction of oxidative phenotype in skeletal muscle [29]. We suggest that this superior adaptation is not depending on the rapid gene activation seen in our time-course.

From 4h on, we saw that the expression level of most target genes is more elevated in the WT animals. This holds for the top target genes of Srf (Egr1,2 and Fos, Fosb), as well as for most of their target genes.

Here, the WT animals sustained the level of mRNA expression of rapidly induced genes longer than the KO animals. We suggested that exactly these sustained mRNA levels induce the Pgc1α-mediated phenotype in response to exercise. What biological functions do those genes have?

As we assumed that the gene expression of the *sustained* genes is due to the sustained pattern of their regulators, we ran KEGG analysis on the target genes of the 4 factors depicted in Figure 3B.

HEAT SHOCK FACTOR GENE EXPRESSION IS DAMPENED IN PGC1α MUSCLE KNOCKOUT MICE A target gene that was ranked as highly important in Egr1 and Egr2, as well as in Srf, was Hspa1b. As outlined above, heat shock proteins are involved in the exercise response. Our data suggested that they are as well regulated or stabilized by Pgc1α - as we saw the mRNA expression of Hspa1b and Dnajb1 to be elevated in the WT for a longer period of time compared to the KO. As Heat shock proteins are thought to have cytoprotective functions, this could be one of the explanations why Pgc1α-lacking animals do not cope as well as WT animals with the stress induced

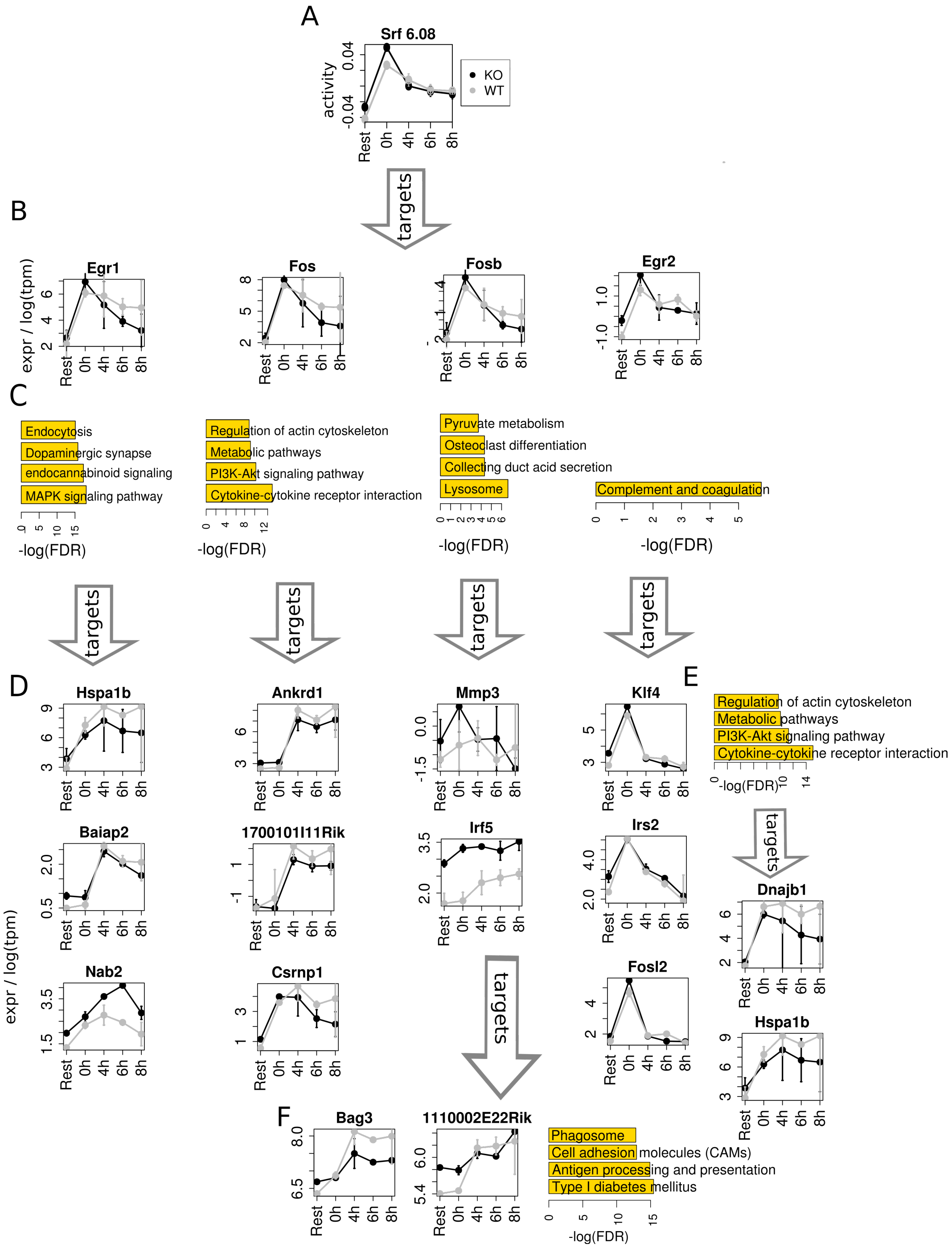


Figure 3: **Pathways induced by the early immediate gene program in response to actute exercise in wildtype (WT) and Pgc1α muscle knock-out (KO) mice.** A) Srf is induced immediately following exercise. B) Among its top targets are Early Growth Response 1/2 (Egr1/2), Fos, Fosb which are transcription factors themselves. C) Targets of the factors depicted in B) are genes highly induced 4h after the exercise bout. D) Some of the factors in B) were transcription factors themselves ant thus had predicted downstream targets, such as Egr2 which targets the promoter of Kruppel like factor 4 (Klf4) which targets DnaJ Heat Shock Protein Family (Hsp40) Member B1 (Dnjab1), Heat shock protein family a 1b (Hspa1b) and Fosb which targets Irf5 which in turn activates the transcription of other genes.

by exercise [47]. For instance, postischemic mice overexpressing Hspa1b showed enhanced contractile and metabolic myocardial recovery, underlining its protective function [48]. One suggestion is that Pgc1α can sustain the levels of this protein which superior exercise-adaptation of the organism.

AUTOPHAGY PATHWAY INDUCED BY FOS AND FOSB Fos and Fosb both showed a sustained mRNA expression from 4h after exercise intervention onwards. Among the KEGG pathways associated with their target genes were autophagy and growth pathways. Surprisingly, one of the top target genes of Fos was 1700101l11Rik, a lncRNA which starts in close vicinity to the GABA Type A Receptor Associated Protein Like 1 (Gabarapl1) gene. Gabarapl1 is a known autophagy marker but has only been studied sparsely, it has been associated with cell proliferation, invasion, and autophagic flux and accumulation of damaged mitochondria, though [49].
Also Irf5, a predicted target gene of Fosb was predicted to activate BAG Cochaperone 3 (Bag3), another autophagy related gene, involved in the protein quality control [57]. As autophagy is a catabolic process that provides the degradation of altered/damaged organelles through the fusion between autophagosomes and lysosomes, it is indispensable for skeletal muscle health [50]. Our data suggested here that Pgc1α played a role in maintaining high levels of autophagy after an acute bout of exercise.

5.3 DISCUSSION

This study was designed to answer two questions: 1) to disentangle the complex regulatory network of acute exercise response in WT mice, and 2) to gain insight into the the role of transcriptional co-activator Pgc1α during exercise response. Therefore, wild type and Pgc1α muscle knockout mice underwent an exhaustion exercise protocol and were killed at rest before the bout and then 0h, 4h, 6h and 8h afterwards. By RNA sequencing of their quadriceps muscle, we obtained an extensive dataset for the follow – up computational analysis. We used ISMARA, a linear model to express the highly dimensional dataset in terms of active transcription factors, which reduces the amount of variables drastically. We then used SVD to find the most important pattern of transcription factor activity. Taking into account the target predictions of ISMARA, we were able to reconstruct the parts of the regulatory network induced by our protocol, and found time-dependent induction of pathways. The genotype difference allowed to disentangle Pgc1αs role in the direct response to exercise.

We found that most of the variance in the dataset was driven by the genetic difference of the groups. Our model reported the activity of factors associated previously with Pgc1α, e.g. Errα to be dramatically reduced in KO animals, whereas other factors like Irf or Nfatc showed higher activity levels in the mice lacking Pgc1α.

Among the top active factors, two were changing their activity rather with time than because of the genotype: Tbp and Srf. While Tbp is rather unknown in the context of exercise response, Srf is a known regulator of early immediate gene response. Here, our model predicted Srf to act as first inducer of the following gene response by activating several factors, (Fos, Egr1/2 and Fosb) which - in turn start the transcription of their targets, including heat-shock proteins and immune-response related factors.

Further, we functionally associated the pattern of activity: The first pathways to happen were response to stimuli external stimuli and stress, followed by autophagy, immune response growth and development related pathways.

The lack of Pgc1α induced a strong phenotype which clearly separated the KO mice from the WT mice, mainly metabolic (e.g. OXPHOS, TCA cycle, Carbon metabolism) and tissue developmental were associated with the difference in transcription factor activity. In the direct response to exercise we didn't find any differences in the WT compared to the KO mice. The induction by Srf and the following activity of its top targets, Egr1/2, Fos and Fosb remained unchanged.

Notably, Pgc1α has been found to be dispensable for the exercise response earlier. Leick et al. [28] conducted a similar study using whole body Pgc1α knockout mice and WT mice and killed them immediately, 2h and 6h after the exercise bout. They claim that Pgc1α is not mandatory for exercise-induced adaptations in murine skeletal muscle directly after the exercise or that other mechanisms can take over its function if it is diminished. However, they still agree that it is needed at all to maintain normal RNA/protein expression levels in skeletal muscle.

In our hands, we found the same dispensability for Pgc1α in terms of immediate gene response: The KO animals responded in the same manner as the WT. However, Pgc1α was still necessary in the acute exercise response: Important pathways maintaining the sanity of the muscle were maintained at a higher level. This may help organisms to recover faster and ameliorate the adaptation of the muscle to exercise. The elevated response could be the result of positive feed-forward loops mediated by myocyte enhancer factor 2 (MEF2), as proposed in [55]. An important characteristic to check in future work is whether genes with sustained expression yield a MEF2-binding site in their promoter. Another very interesting follow up study could focus on how this sustained activity is achieved: It may be because the proteins of regulators themselves are stabilized and thus initiate the transcription much longer, or it may be because the regulatory region of their downstream targets are silenced later (e.g. by DNA methylation). A view on methylation and accessibility of regions (Methyl-seq and ATAC-seq) would be the appropriate answer to these questions, plus analyzing the same regions in the muscle of chronically trained mice. This would further give the possibility to investigate the *muscle memory*, an effect which facilitates retraining in trained athletes, but has not been understood profoundly to date.

5.4 METHODS

VIEW FROM THE REGULATOR SIDE: ISMARA ON THE RNA TIME-COURSE ISMARA models genome-wide gene expression pattern in terms of predicted functional Transcription Factor Binding Sites (TFBSs) in the respective gene's promoters. Promoter regions are either annotated or taken to be -500 +500 of the TSS defined by Cap analysis gene expression (CAGE) analysis. In the model, the expression of promoter p in sample s, E_{ps} is assumed to follow a linear function of the binding sites N_{pm} in promoter p and motif m times an unknown activity A_{ms} of a motif binding site m in sample s. Summing across all motifs gives the core ISMARA equation:

$$E_{ps} = \sum_{m} N_{pm} \cdot A_{ms} + c_p + c_s$$

whereas c_p and c_s account for the promoter related basal expression and for the sample-dependent normalization constant, respectively.
The matrix N_{pm} contains information on the binding sites in each promoter and has been inferred using the algorithm MotEvo, previously developed in our group [11]. MotEvo calculates the posterior probability of a site to occur in a promoter, depending on a background prior (which accounts for the site to be found randomly in any sequence) and the conservation of the site in the promoter. We have collected an extensive library of positional weight matrices (PWM) for around 600 motif binding sites in mouse, for which we predict binding sites. ISMARA also calculates, for each motif, a zScore which denotes the importance of the motif in explaining the observed gene expression data. In other words, a change in activity represents the change in the expression E_{ps} of promoter p in sample s, when motif m exactly in this promoter p would be removed. This means that the higher the activity, the higher the expression of genes having this motif in the promoter.

TARGET PREDICTIONS The target promoters p_m of motif m are genes that are expressed in the dataset and have a binding site for motif m in their sequence. To estimate the importance of each promoter, ISMARA calculates a target score which is the relative square deviation between the model without this binding site (mutated version) and the one with all binding sites (full version). which is the difference how worse the fit would be if the site for motif m in promoter p would be missing.

SVD ON THE MOTIF ACTIVITIES We applied SVD it to the activities of 503 motifs across WT and KO samples. Our matrix A_{ms} thus contains 10 x 503 (samples x motifs) entries. SVD decomposes the matrix as $A = U\Lambda V$, where V and U are the right and left singular vectors, respectively. Λ contains the singular values. Any motif activity pattern $(\vec{a}_m)_s$ can now be represented in a linear combination of the right singular vectors $\vec{v}_i$.

INTERPRETATION OF THE SINGULAR VECTORS The coordinates of the motifs in the are sample space thus depict the contribution of each sample to the motif activity pattern. As the singular vectors point in the direction of highest variance in the dataset, they capture consecutively the pattern contributing to the variation in the dataset, e.g. the first component captures the highest amount of variance, followed by the second component. We plot the coordinates for each component across the samples (=timepoints) to visualize the pattern.

inferring motifs following the principal components As now the motif activities are vectors in the principal component space, their projection on each of the axes (=principal components) indicates how strongly the motif overlaps with the bases, thus, how significantly it changes following the pattern. We calculate the projections according to $q_{mk} = \vec{a}_m \cdot \vec{v}_k$, and as, according to the SVD, $AV = U\Lambda$ it can be written in matrix multiplication $q_{mk} = (U\Lambda)_{mk}$ (see also suppl Figure S2) We are interested in the pearson correlation of the motif vectors to the principal components as they indicate how accurately the motif activity pattern follows the one depicted by the principal components. As the new orthogonal basis is mean centered, such as the motif activities, we can readily compute the correlation as: $p_{mk} = q_{mk}/\sqrt{\sum_k (q_{mk})^2}$. The Pearson correlation coefficients p_{mk} and the projection values q_{mk} do not overlap in information content. Motifs with large activity changes will have higher projections than a motif whose activity changes are minor, whereas they could have the same Pearson correlation coefficient. Here, only use the pearson correlation to decide whether a motif follows the pattern. We usually ask for a minimal value of 0.8. To evaluate its significance, we make use of the in-built calculation of the zScore which also gives, for each motif, information about how strongly it is changing its activity across samples.

gene ontology enrichment analysis We used the R package STRINGdb 1.24.0 [12] in R version 3.6.0 to compute functional categories for the target genes of motifs or motif groups. We used it directly on Ensemble Gene Ids, as a reference set we used all the genes expressed in the dataset in at least one timepoint and one genotype. We extracted GO categories for the biological process and KEGG category and filtered for terms with a FDR < 0.05.

differential gene expression analysis To detect differentially expressed genes, we mapped the reads using `kallisto` [15] and summed the transcript raw counts for each gene, followed by standard edge R 3.26.1 analysis. We used the sedentary group for both WT and KO as a control to detect differential genes at each timepoint after the exercise bout. We only use genes with an FDR < 0.05 for further analysis. To draw the venn diagram, we used all gene names of differentially expressed genes and calculated the overlap using `InteractiVenn` [13].

animals Male mice at the age of 20 weeks were used and housed in a 12h light/-dark cycle. They had ad libitum access to food and water. The Pgc1α muscle specific knockout (KO) mice were generated as depicted in [31], floxed littermates were used as control [58]. The mice were acclimatized to the treadmill with the following procedure:

day1		day2		day3		day4		day5	
time	speed	time	speed	time	speed	time	speed	time	speed
min	m/min	min	m/min	min	m/min	min	m/min	min	m/min
5	0	5	0	5	0	5	0	5	0
5	5	5	5	5	5	5	5	5	5
5	8	5	8	5	8	5	8	5	8
10	10	15	10	15	10	15	10	10	10
		5	12	5	12	5	12	5	12
				2	14	2	14	2	14

Table 1: **Acclimatization protocol for the mice undergoing the exhaustion test**. The treadmill had 5° inclination

EXERCISE PROTOCOL For the exhaustion exercise a treadmill was used (Columbus Instruments). The exercise protocol was

time	speed
5min	0m/min
5min	5m/min
5min	8m/min
every 15min	+2m/min
final	26m/min

Table 2: **Protocol for the final exhaustion treadmill training**. The treadmill was inclined by 5°.

The speed increased 2m/min every 15 min until a final velocity 26m/min was reached before exhaustion. Immediately (0h), 4h, 6h and 8h the mice were killed with CO2 and the quadriceps muscle was collected. The control mice were not exposed to any exercise.

PREPARATION OF THE LIBRARY AND SEQUENCING Then, RNA was extracted from quadriceps muscle and purfied with the Direct-zol RNA MiniPreo Kit (R2050). 1µg of purified RNA was used to construct the library with the Illumina TruSeq RNA library Prep Kit. Single end RNA sequencing was performed in 50 cycles with a High-Seq 2500 Illumina machine.

5.5 supplementary material

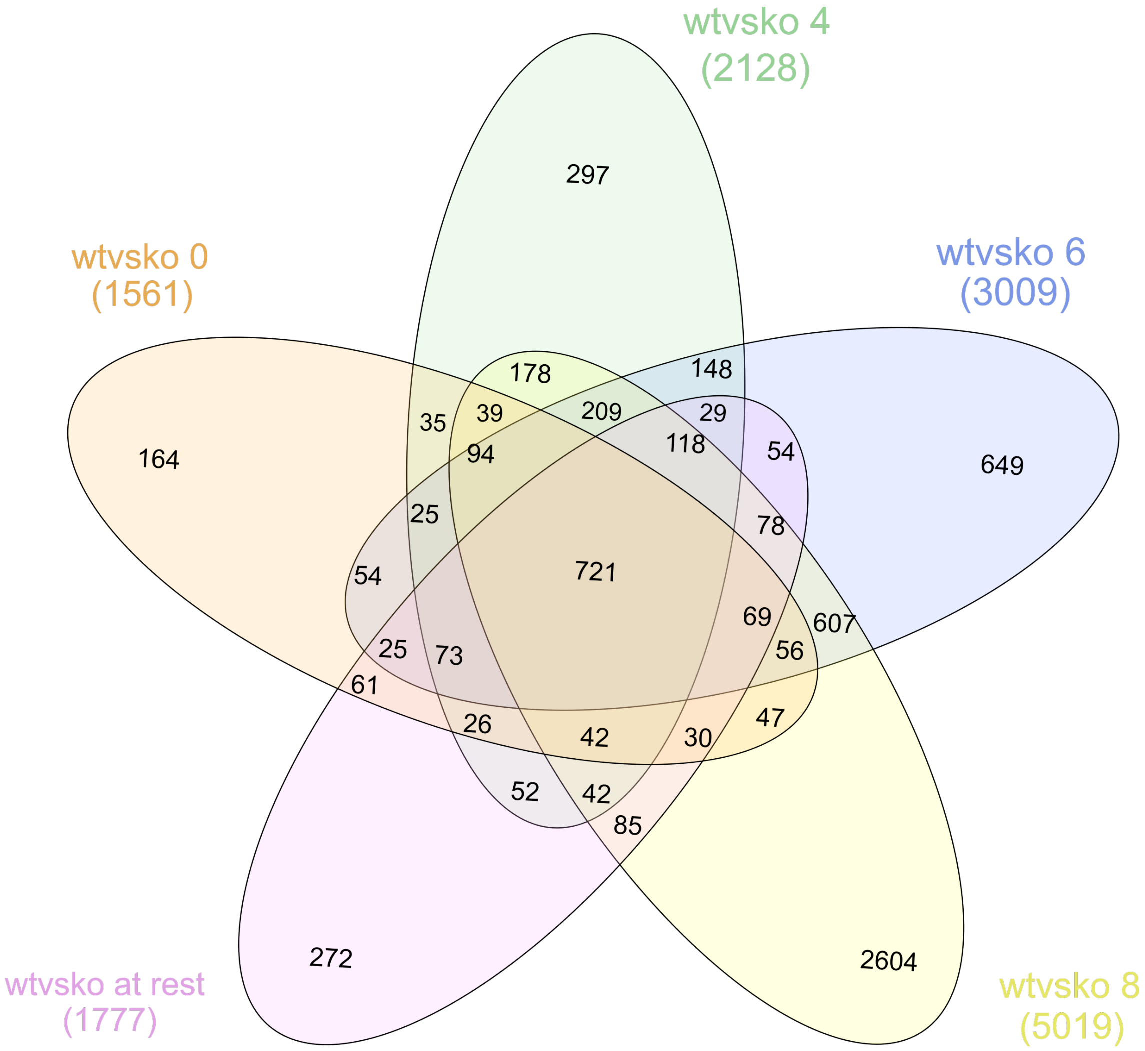


Figure S1: Differential Gene expression Analysis reveals large gene expression changes in response to exercise. The venn diagrams show the number of genes which are differentially (DE) expressed when comparing WT-KO at all the timepoints. Numbers in brackets show the overall DE expressed genes. We only report genes with a FDR < 0.05. A) Pgc1α muscle knock out (KO) animals. We compare the timepoints after exercise, 0h,4h,6h and 8h to the sedentary control (rest).

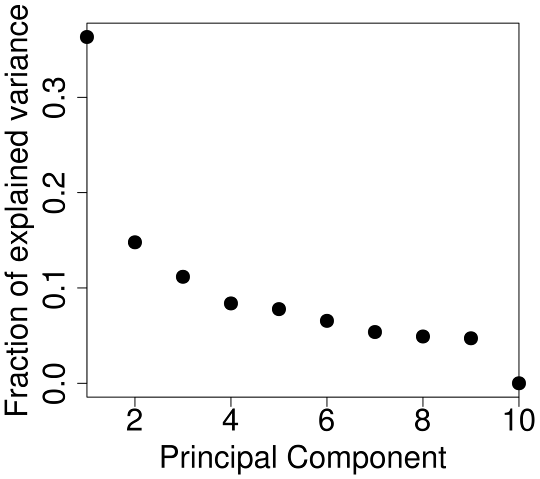
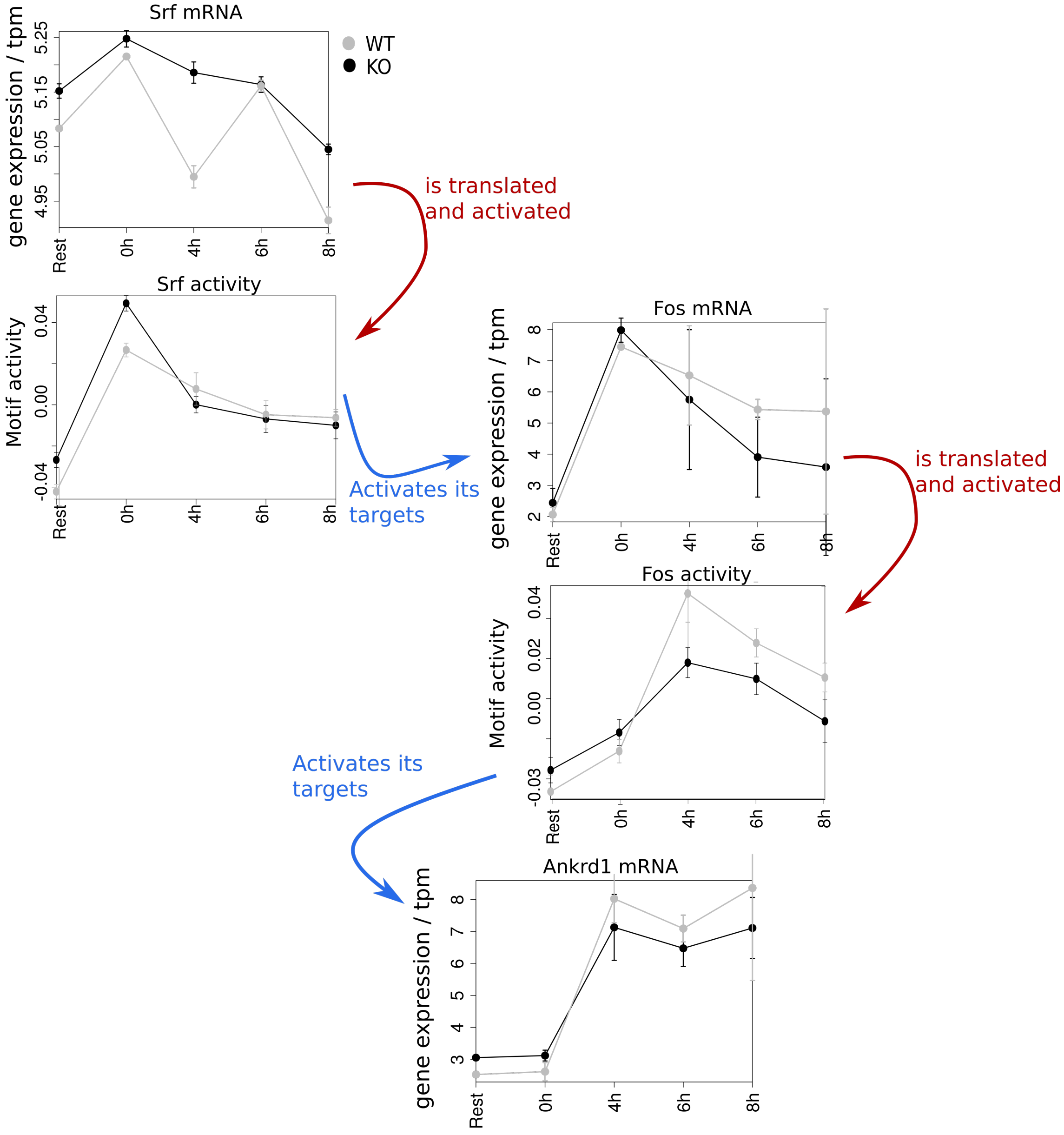


Figure S2: Fraction of explained variance in the SVD across all timepoints and both genotypes.



s

Figure S3: The output of ISMARA and reconstruction of regulatory networks (Srf-Fos-Ankrd1 as an example: Srf mRNA was constant across the time. Most likely its activity is induced by any posttranslational modification like phosphorylation (proposed in [15]) or gets activated by translocation into the nucleus. Its activity is highly induced by exercise, meaning that starts with the initation of transcription of its target genes. One of the target genes is Fos, whose mRNA responds in the same way as the activity pattern of SRF. However, looking at its activity, meaning the expression of its targets, it gets active one timepoint later. This is perfectly understandable, as the translation process of Fos and the transcription process of the target genes takes time. Here, the mRNA of Ankrd1 peaks at 4h, consistent with the activity of Fos.

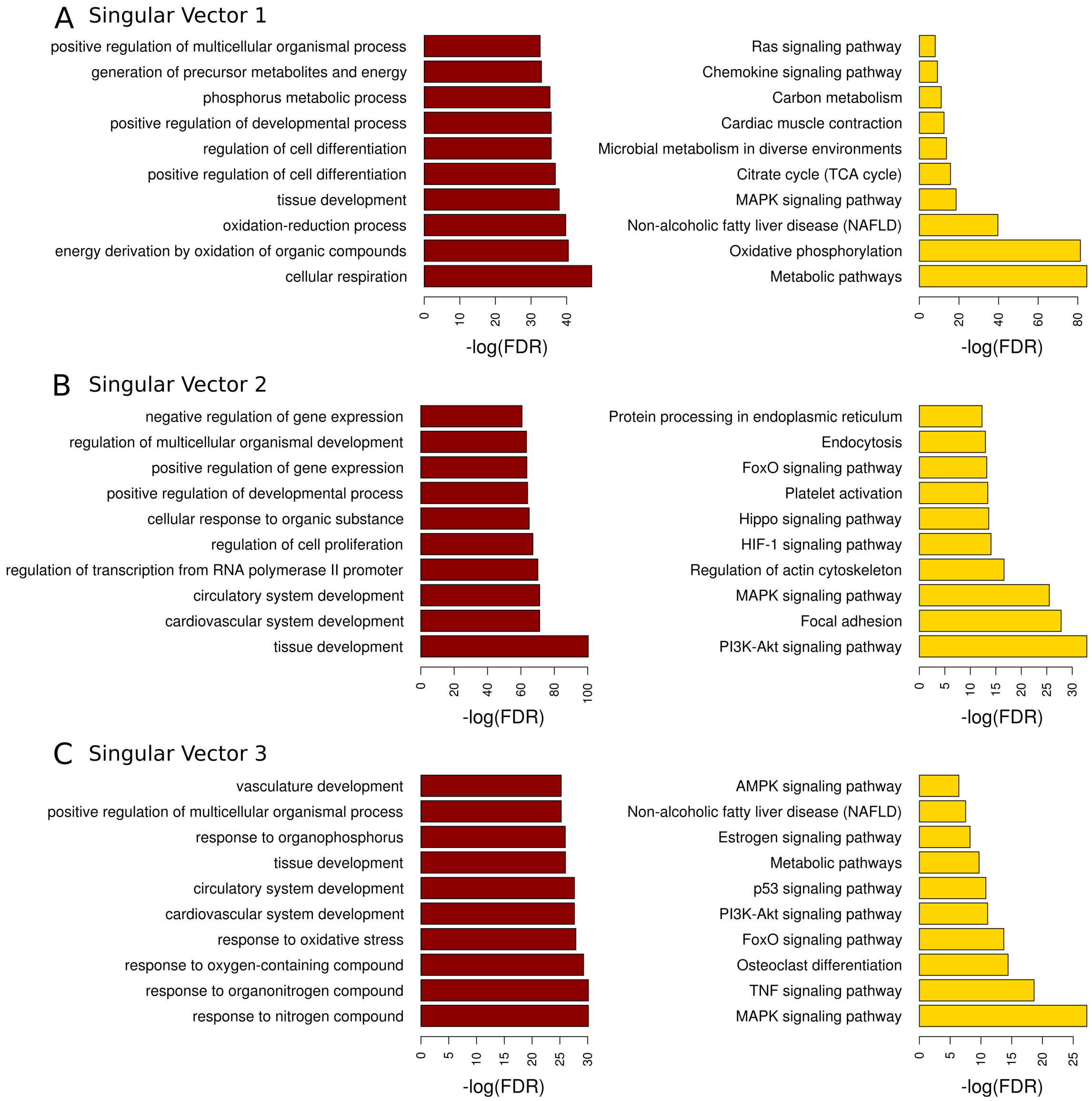


Figure S4: Gene Ontolgoy (red) and Kegg pathways (yellow) derived for associated genes of motifs correlating with singular vectors 1-3 (PC1-3).

BIBLIOGRAPHY

[1] Lira, V. A., Benton, C. R., Yan, Z., & Bonen, A. (2010). PGC-1alpha regulation by exercise training and its influences on muscle function and insulin sensitivity. American Journal of Physiology. Endocrinology and Metabolism, 299(2), E145-61. https://doi.org/10.1152/ajpendo.00755.2009

[1] Fitzsimons, Daniel P., et al. "Effects of endurance exercise on isomyosin patterns in fast-and slow-twitch skeletal muscles." Journal of Applied Physiology 68.5 (1990): 1950-1955.

[2] O'gorman, D. J., et al. "Exercise training increases insulin-stimulated glucose disposal and GLUT4 (SLC2A4) protein content in patients with type 2 diabetes." Diabetologia 49.12 (2006): 2983-2992.

[3] Knowler, William C., et al. "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin." The New England journal of medicine 346.6 (2002): 393-403.

[4] Borst, Stephen E. "Interventions for sarcopenia and muscle weakness in older people." Age and ageing 33.6 (2004): 548-555.

[5] Lin, Jiandie, et al. "Transcriptional co-activator PGC-1α drives the formation of slow-twitch muscle fibres." Nature 418.6899 (2002): 797-801.

[6] Pilegaard, Henriette, Bengt Saltin, and P. Darrell Neufer. "Exercise induces transient transcriptional activation of the PGC-1α gene in human skeletal muscle." The Journal of physiology 546.3 (2003): 851-858.

[7] Irrcher, Isabella, et al. "PPARγ co-activateor-1α expression during thyroid hormone-and contractile activity-induced mitochondrial adaptations." American Journal of Physiology-Cell Physiology 284.6 (2003): C1669-C1677.

[8] Wu, Zhidan, et al. "Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic co-activateor PGC-1." Cell 98.1 (1999): 115-124.

[9] Knutti, Darko, and Anastasia Kralli. "PGC-1, a versatile co-activateor." Trends in Endocrinology & Metabolism 12.8 (2001): 360-365.

[10] Luisier, Raphaelle, et al. "Computational modeling identifies key gene regulatory interactions underlying phenobarbital-mediated tumor promotion." Nucleic acids research 42.7 (2014): 4180-4195.

[11] Arnold, Phil, et al. "MotEvo: integrated Bayesian probabilistic methods for iPgc1$\alpha\alpha\alpha$rring regulatory sites and motifs on multiple alignments of DNA sequences." Bioinformatics 28.4 (2012): 487-494.

[12] Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

[13] Heberle, H.; Meirelles, G. V.; da Silva, F. R.; Telles, G. P.; Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 16:169 (2015).

[14] Bahar Halpern K. et al., Bursty gene expression in the intact mammalian liver. Mol Cell. 2015 Apr 2 58(1):147-56. doi: 10.1016/j.molcel.2015.01.027. p.150 left column top paragraph

[15] Rivera, Victor M., et al. "A growth factor-induced kinase phosphorylates the serum response factor at a site that regulates its DNA-binding activity." Molecular and cellular biology 13.10 (1993): 6260-6273.

[16] Lee, Seung-Min, Mansi Vasishtha, and Ron Prywes. "Activation and repression of cellular immediate early genes by serum response factor cofactors." Journal of Biological Chemistry 285.29 (2010): 22036-22049.

[17] Modak, Cristina, and Jianyuan Chai. "Serum response factor: look into the gut." World Journal of Gastroenterology: WJG 16.18 (2010): 2195.

[18] Parker, Thomas G., and Michael D. Schneider. "Growth factors proto-oncogenes and plasticity of the cardiac phenotype." Annual Review of Physiology 53.1 (1991): 179-200.

[19] Schneider, M. D., et al. "Growth factors, growth factor response elements, and the cardiac phenotype." Cardiac Adaptation in Heart Failure. Steinkopff, 1992. 33-48.

[20] Alvarez-Guardia, David, et al. "The p65 subunit of NF-κB binds to PGC-1α, linking inflammation and metabolic disturbances in cardiac cells." Cardiovascular research 87.3 (2010): 449-458.

[21] Wang, Lie, et al. "The zinc finger transcription factor Zbtb7b represses CD8-lineage gene expression in peripheral CD4+ T cells." Immunity 29.6 (2008):

[22] Ryoo, In-geun, and Mi-Kyoung Kwak. "Regulatory crosstalk between the oxidative stress-related transcription factor Pgc1ααα2l2/Nrf2 and mitochondria." Toxicology and applied pharmacology 359 (2018): 24-33.

[23] Ramachandran, Krithika, et al. "Dynamic enhancers control skeletal muscle identity and reprogramming." PLoS biology 17.10 (2019): e3000467.

[24] Carracedo, Arkaitz, et al. "A metabolic prosurvival role for PML in breast cancer." The Journal of clinical investigation 122.9 (2012): 3088-3100.

[25] Horak, Martin, et al. "Exercise-induced circulating microRNA changes in athletes in various training scenarios." PloS one 13.1 (2018).

[26] Noble, Earl G., and Garry X. Shen. "Impact of exercise and metabolic disorders on heat shock proteins and vascular inflammation." Autoimmune diseases 2012 (2012).

[27] Yan, Zhen. "Exercise, PGC-1α, and metabolic adaptation in skeletal muscle." Applied Physiology, Nutrition, and Metabolism 34.3 (2009): 424-427.

[28] Leick, Lotte, et al. "PGC-1α is not mandatory for exercise-and training-induced adaptive gene responses in mouse skeletal muscle." American journal of physiology-endocrinology and metabolism 294.2 (2008): E463-E474.

[29] Handschin, Christoph, et al. "Skeletal muscle fiber-type switching, exercise intolerance, and myopathy in PGC-1α muscle-specific knock-out animals." Journal of Biological Chemistry 282.41 (2007): 30014-30021.

[30] Baresic, Mario, et al. "Transcriptional network analysis in muscle reveals AP-1 as a partner of PGC-1α in the regulation of the hypoxic gene program." Molecular and cellular biology 34.16 (2014): 2996-3012.

[31] Handschin, Christoph, et al. "Skeletal muscle fiber-type switching, exercise intolerance, and myopathy in PGC-1α muscle-specific knock-out animals." Journal of Biological Chemistry 282.41 (2007): 30014-30021.

[32] Handschin, Christoph, and Bruce M. Spiegelman. "The role of exercise and Pgc1$\alpha\alpha\alpha$ in inflammation and chronic disease." Nature 454.7203 (2008): 463-469.

[33] Gill, Jonathan F., et al. "Peroxisome proliferator-activated receptor γ co-activateor 1α regulates mitochondrial calcium homeostasis, sarcoplasmic reticulum stress, and cell death to mitigate skeletal muscle aging." Aging cell 18.5 (2019): e12993.

[34] Handschin, C and Spiegelman, B Peroxisome proliferator-activated receptor-gamma co-activateor 1 alpha (PGC-1 alpha): transcriptional co-activateor and metabolic regulator, PNAS June 10, 2003 100 (12) 7111-7116

[35] Patti, Mary Elizabeth, et al. "Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of Pgc1$\alpha\alpha$ and NRF1." Proceedings of the National Academy of Sciences 100.14 (2003): 8466-8471.

[36] Hatazawa, Yukino, et al. "Metabolomic analysis of the skeletal muscle of mice overexpressing PGC-1α." PloS one 10.6 (2015).

[37] Handschin, Christoph. "Regulation of skeletal muscle cell plasticity by the peroxisome proliferator-activated receptor γ co-activateor 1α." Journal of receptors and signal transduction 30.6 (2010): 376-384.

[38] Wang, Lin Ru, and Seung-Soo Baek. "Treadmill exercise activates PI3K/Akt signaling pathway leading to GSK-3β inhibition in the social isolated rat pups." Journal of exercise rehabilitation 14.1 (2018): 4.

[39] Watt, Kevin I., et al. "The hippo signaling pathway in the regulation of skeletal muscle mass and function." Exercise and sport sciences reviews 46.2 (2018): 92-96.

[40] Yu, Fa-Xing, and Kun-Liang Guan. "The Hippo pathway: regulators and regulations." Genes & development 27.4 (2013): 355-371.

[41] O'Hagan, Kathleen A., et al. "PGC-1α is coupled to HIF-1α-dependent gene expression by increasing mitochondrial oxygen consumption in skeletal muscle cells." Proceedings of the National Academy of Sciences 106.7 (2009): 2188-2193.

[42] Yu, Miao, et al. "Exercise activates p53 and negatively regulates igf-1 pathway in epidermis within a skin cancer model." PloS one 11.8 (2016).

[43] Petersen, Anne Marie W., and Bente Klarlund Pedersen. "The anti-inflammatory effect of exercise." Journal of applied physiology 98.4 (2005): 1154-1162.

[44] Roux, Philippe P., and John Blenis. "ERK and p38 MAPK-activated protein kinases: a family of protein kinases with diverse biological functions." Microbiol. Mol. Biol. Rev. 68.2 (2004): 320-344.

[45] Kramer, Henning F., and Laurie J. Goodyear. "Exercise, MAPK, and NF-κB signaling in skeletal muscle." Journal of applied physiology 103.1 (2007): 388-395.

[46] Milne, Kevin J., and Earl G. Noble. "Exercise-induced elevation of HSP70 is intensity dependent." Journal of Applied Physiology 93.2 (2002): 561-568.

[47] Kiang, Juliann G., and George C. Tsokos. "Heat shock protein 70 kDa: molecular biology, biochemistry, and physiology." Pharmacology & therapeutics 80.2 (1998): 183-201.

[48] Marber, Michael S., et al. "Overexpression of the rat inducible 70-kD heat stress protein in a transgenic mouse increases the resistance of the heart to ischemic injury." The Journal of clinical investigation 95.4 (1995): 1446-1456.

[49] Boyer-Guittaut, Michaël, et al. "The role of GABARAPL1/GEC1 in autophagic flux and mitochondrial quality control in MDA-MB-436 breast cancer cells." Autophagy 10.6 (2014): 986-1003.

[50] Grumati, Paolo, et al. "Physical exercise stimulates autophagy in normal skeletal muscles but is detrimental for collagen VI-deficient muscles." Autophagy 7.12 (2011): 1415-1423.

[51] Leveille, Melissa, et al. "PGC-1α isoforms coordinate to balance hepatic metabolism and apoptosis in inflammatory environments." Molecular Metabolism (2020).

[52] Gudmundsson, Sanna, et al. "TAF1, associated with intellectual disability in humans, is essential for embryogenesis and regulates neurodevelopmental processes in zebrafish." Scientific reports 9.1 (2019): 1-11.

[53] Besse-Patin, Aurele, et al. "Estrogen signals through peroxisome proliferator activated Receptor gamma co-activateor 1α to reduce oxidative damage associated with diet-induced fatty liver disease." Gastroenterology 152.1 (2017): 243-256.

[54] Summermatter, Serge, et al. "Skeletal muscle PGC-1α controls whole-body lactate homeostasis through estrogen-related receptor α-dependent activation of LDH B and repression of LDH A." Proceedings of the National Academy of Sciences 110.21 (2013): 8738-8743.

[55] Handschin, Christoph, et al. "An autoregulatory loop controls peroxisome proliferator-activated receptor γ coactivator 1α expression in muscle." Proceedings of the national academy of sciences 100.12 (2003): 7111-7116.

[56] Harris, Lachlan, et al. "Nuclear factor one transcription factors: divergent functions in developmental versus adult stem cell populations." Developmental dynamics 244.3 (2015): 227-238.

[57] Stürner, Elisabeth, and Christian Behl. "The role of the multifunctional BAG3 protein in cellular protein quality control and in disease." Frontiers in molecular neuroscience 10 (2017): 177.

[58] Heim, Barbara, Molecular mechanisms of the transcriptional regulation by PGC-1α/β in skeletal muscle, PhD Thesis, University of Basel, 2018

6 PREDICTION OF THE RESPONSE TO IMMUNOTHERAPY IN CANCER

Immune checkpoint blockade (ICB) therapy has recently evolved to be one of the most promising therapies in a number of cancers. Patients undergoing this therapy experience a decrease or even the cure of their disease and a dramatically prolonged survival. Unfortunately, only a subset of patients respond to the therapy in a positive way, others might suffer from serious side effects caused by autoimmune responses. Given the versatile structure of most cancers is is challenging to find reliable genetic markers to predict the response of patient to the therapy. Thus researchers make use of the growing amount of sequencing data on different cancer types. However, even with computational tools and growing datasets with clinical annotation, it remains a huge challenge to disentangle the needed information in those complex datasets. In this study we use microarray and RNA-seq data of melanoma and neuroblastoma to predict the possibility of a patient to respond to ICB therapy in a positive way by testing and applying machine learning strategies.

Anne Krämer[1,2], Christoph Handschin[1], Erik van Nimwegen[1] and Jung Kyoon Choi[2]
[1] University of Basel, Basel, Switzerland
[2] Korea Advanced Institute of Science and Technology, Daejeon, South Korea

6.1 INTRODUCTION

SPONTANEOUS REGRESSION The sudden healing of cancer has been described since hundreds of years. Spontaneous (without any apparent cause) regression (decrease of the size of the tumor) was standardly defined as *the partial or complete disappearance of a malignant tumor in the absence of treatment or in the presence of therapy considered inadequate to exert a significant influence on the disease* in the 1960s [14, 15]. Spontaneous regression is caused by the human immune system which makes use of immune checkpoint regulatory pathways. In the case of cancer the auto defense fails as these pathways – when activated – dampen the immune response in order to prevent the immune system to arbitrarily kill cells. Cancer cells trick this system by activating important genes in the checkpoint regulatory pathway [16, 17]. Like this, the malignant cells escape the defense mechanism of the body, and even trigger immune system suppression. Otherwise, if the immune cells would recognize cancer cells as hostile, the biologic defense system against the tumor would be started and counteract or even stop the disease.

IMMUNE CHECKPOINT BLOCKADE Cancer cells use the ability of the innate immune system to dampen the immune response by activating so-called immune checkpoint blockades. Immune checkpoint therapy (ICB) is a very successful approach to target malignant cancer cells by usage of the cytotoxic potential of the immune system. As cancer cells use the ability of the innate immune system to dampen the immune response by activating so-called immune checkpoint blockades, the therapy targets the cytotoxic T lymphocyte antigen 4 (CTLA4) or the programmed cell death ligand 1 (PD-L1) to stop the blockade. ICB has had several successes across different cancer types. Numerous other targets for negative regulation between tumor cells and T-cells, or myeloid cells and T-cells are in clinical and preclinical study. However, only a small fraction of patients is likely to respond to the therapy, imposing a big challenge on clinicians to decide a priori which patient should be treated in which way. The potential factors discerning a responsive from a non-responsive patient range from patient-related parameters such as weight, age and sex, tumor-intrinsic parameters such as the host immune system and tumor-associated stroma and biological parameters, such as the gut microbiota [17, 11].

MACHINE LEARNING STRATEGIES RNA profiling followed by computational analysis is becoming increasingly important. In the treatment of cancer, getting information about the genetic signature of the cancer can help deciding which therapy to use. Naturally, these large transcriptomic datasets are highly dimensional and often very hard to interpret. Machine Learning (ML) algorithms use statistical and optimization techniques to *learn* from past examples and detect otherwise hard to discern patterns in unknown new data. This makes it especially applicable for cancer detection and classification. More recently, the field of personalized medicine uses ML to predict which therapy fits best to the patients and tumours genetic signature [18]. In this study, we apply a combination of self-written machine learning algorithms and published tools to detect the possibility that a patient would respond to immune blockade therapy.

NEUROBLASTOMA AND MELANOMA CELLS Here, we focused on melanoma, a cancer type that frequently shows spontaneous regression. We included neuroblastoma (NB) cells, as they have the same common origin as melanoma cells and show frequent spontaneous regression, especially in children [1, 19]. We made use of a public transcriptomic dataset with clinical annotation. Moreover, spontaneous regression is seen in NB patients, and very frequently in children less than 18 months of age [12]. Further, NB is the first type of cancer to be treated with ICB with approval of the US Food and Drug Administration. Thus, we assumed that building up a predictor on spontaneous regression in NB will be an exact predictor for the response to immune therapy.

6.2 RESULTS

OVERVIEW OF THE DATA We then followed loosely the approach of Auslander et al [1] to construct a powerful predictor of immune response (IMPRS) and test it on various other datsets. We used 4 published datasets on neuroblastoma and melanoma cancers:
1) An extensive microarray set of primary neuroblastoma taken from patients of different ages up to 24 years old [27] (NB).

2) A transcriptome data from pretreatment tumor samples of melanoma tumor biopsies from 40 patients [21] (vanAllen).
3) A dataset of whole exome sequencing samples of 68 patients pre- and during therapy [6] (Riaz) and
4) transcriptomes of responding (n = 15) and non-responding (n = 13) pretreatment melanoma tumors (total 27 of 28 pretreatment; 1 of 28 early on-treatment) [5] (Hugo).

All the datasets are publicly available. As the vanAllen and Hugo datasets were too sparse to perform an appropriate analysis, we merged those datasets with others: The vanAllen and Hugo Datasets were combined (anti-PD-L1 and anti-CTLA4 datasets) respectively, this dataset is referred to as vanAllen. The Hugo/Riaz datasets were combined (both anti PD-L1), this dataset is referred to as Hugo. The Riaz and the NB dataset included enough patients which makes the analysis possible on the single datasets (NB and Riaz).

NEUROBLASTOMA DATASET It was necessary to reduce the amount of samples in the dataset, as in total 489 samples were provided and we hypothesized that not all of them are equally important to find a signature for spontaneous regression and immune response. Therefore, we tried to pre-select samples which behave similarly.
The data was annotated with information on the age, sex and whether spontaneous regression took place. To discern important parameters shaping the structure of the data, we performed SVD and included the given annotation to see whether we could explain the formation of clusters.
Our SVD clearly showed a separation of patients along the first singular vector. To explain the clusters, we first looked at four different characteristics, age, sex, spontaneous regression and progression of the cancer. None of those factors clearly distinguished the prominent clusters along the first singular vector (Figure 1). Hence, we needed to perform another type of feature selection.
Different approaches for the selection of meaningful samples have been applied by Auslander et al. [1]. We followed their approach partwise.
First a SVD was performed on all samples. 2-3 clusters appeared in the 1st and 2nd component. Auslander et al. chose samples with PC2+PC3>0 to get a unbiased subset of samples. However, in our approach the association of SVD clusters with clinical features was not as clear as stated in the corresponding paper. Therefore we focused on the feature reduction using literature and clinical data.
Spontaneous regression has been almost exclusively observed in younger patients < 18 months. Older patients often show metastases or unresectable tumors which would confine the analysis ([1], suppl. Material). We thus chose only those samples. Further we classified the remaining samples into two groups: *Spontaneous Regression*: 'no high risk' and 'no cancer progression'; and *No Spontaneous Regression*: 'high risk' and 'cancer progression'. This results in our case in 236 samples (17 'high risk', 219 'not high risk').

FEATURE SELECTION As in [1], we focused on in total 26 immune checkpoint genes (BTLA, PDCD1, CD200, CD200R1, CD27, CD276, CD28, CD40, CD80, CD86, CEACAM1, CTLA4, HAVCR2, IDO1, IL2RB, LAG, PD1LG2, PVR, PVRL2, TIGIT, TNFRSF14, TNFRSF18, TNFRSF4, TNFRSF9, OX40L, CD137) (Table S1). which reduced the massive feature (=gene) space in the datasets and restricts the predictions to be

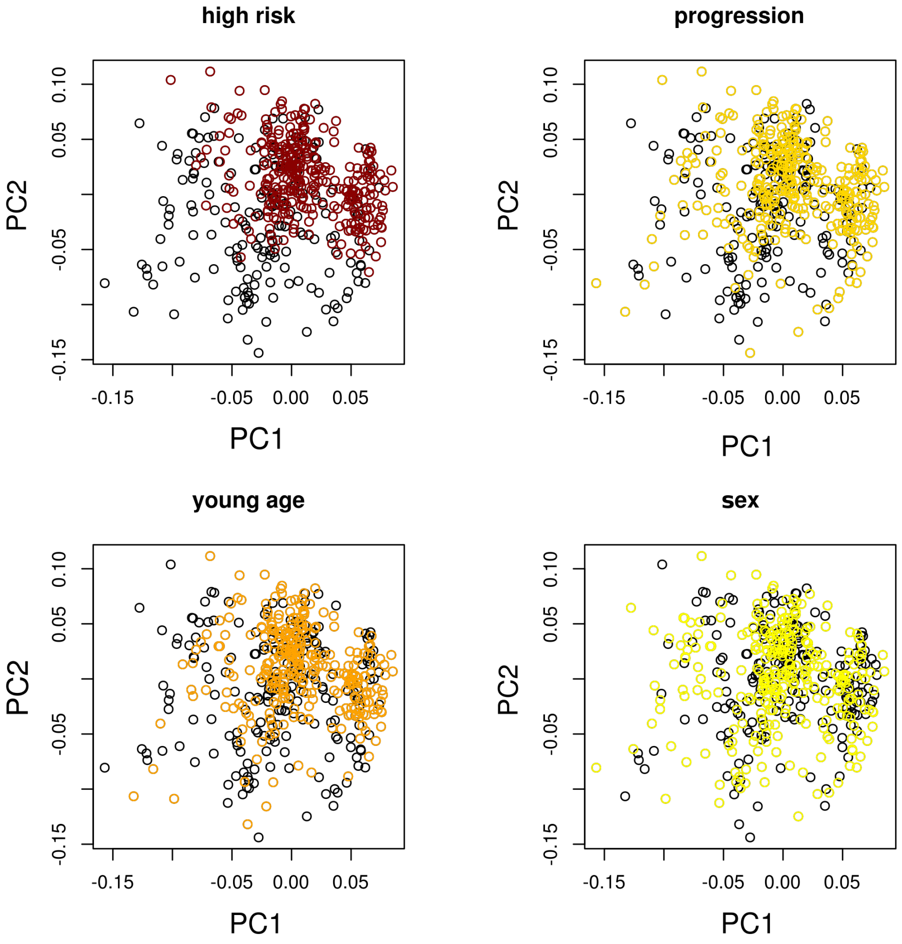


Figure 1: **SVD of the neuroblastoma (NB) data:** No obvious cluster was found to be defined by important clinical features.

done entirely in regard to the immune response.

As gene expression levels can vary across patients, we generalized the data in the following way: For each gene-pair out of the immune checkpoint genes above, we constructed a pairwise comparison matrix which took into account the expression levels of gene i and j across all genes:

$$F_{i,j} = \begin{cases} \exp_i(x) < \exp_j(x) \\ 0, \text{otherwise} \end{cases} \quad \text{(see Figure 3A,B)}$$

This principle of a binary feature table is used in all following approaches for all datasets.

PERFORMANCE OF DIFFERENT APPROACHES First we tested different already published tools to make our predictions on the datasets.

NEURALNET Neural networks have been applied successfully to cancer classification problems. The first attempts started about 20 years ago [7, 8].

In brief, neural networks first transform the input data into a feature space, consisting of h linear combinations and an activation function. Transformation can be done

in each layer differently. The output layer then transforms the data e.g. by regression back into the desired format, e.g. binary classification (see Figure S1B).
The essential steps of a neural network are:

- *Forward Propagation* The initial input (e.g. gene expression, microarray) propagates through the layers, taking into account the number of hidden nodes and the activation function which determines how much weight each node gets.

- *Backward Propagation* In the backward propagation, the likelihood of the outcome is maximized with respect to the weights each node gets.

- *Training* is the process of fore- and backpropagating through the layer while optimizing the layers each time until the defined stopping criterion is reached

- *Prediction* To classify yet unknown samples, the samples are fed forward through the network. If the true classes of the samples are known, the networks performance can be assessed by quantifying the dependence of true positive to false positive rate (ROC curve) or the area under the ROC curve (AUC).

We aimed to start from a very basic implementation of a neural network consisting of the steps mentioned above and used cross validation to quantify its performance. As this neural network yielded moderate performance when applied to our datasets (Figure S1A), only some iterations reached an AUC above 0.5. Hence, we tried a more sophisticated model provided by the R package `neuralnet`.
When applying `neuralnet` with the same parameters as used in the self-implemented model, it showed immediately higher AUC values while cross-validation (Figure 2A and Table 1) However, as seen previously in applying the ML approach to the datasets of Riaz, van Allen and Hugo et al, we saw weaker performance of `neuralnet` in those datasets.
We used the neural network with two hidden layers and changed the number of nodes in the layers. The number of used neurons in each layer is given on the right side of figure 2A.

SVM SVMs are used heavily in cancer classification tasks, the first successful application dates back to the early 2000s [9]. As a next test, we used an SVM proposed by Vapnik [13], which is implemented as R package `kernlab` and has been studied extensively for classification, regression and density estimation.
Briefly, SVMs find hyperplanes maximizing the distance to points of different classes which are closest to each other, which results in a quadratic programming problem (a complete description of the algorithm is beyond the scope of this report, for more information, refer to [2]).
We applied SVM to the 3 datasets to estimate its performance. To find out the best sets of parameters for the SVM, we performed grid search on the following parameters:

- *kernel*: A kernel function maps features in the initial space to a space which just depends on the dot product of the vectors of two features. `kernlab` allows the use of six different kernels: `rbfdot`, `polydot`, `laplacedot`, `besseldot`, `anovadot`, `vanilladot`.

- *cost*: The cost parameter is used to control for overfitting. Increasing its value will result in finding a separating hyperplane which will classify as many datapoints

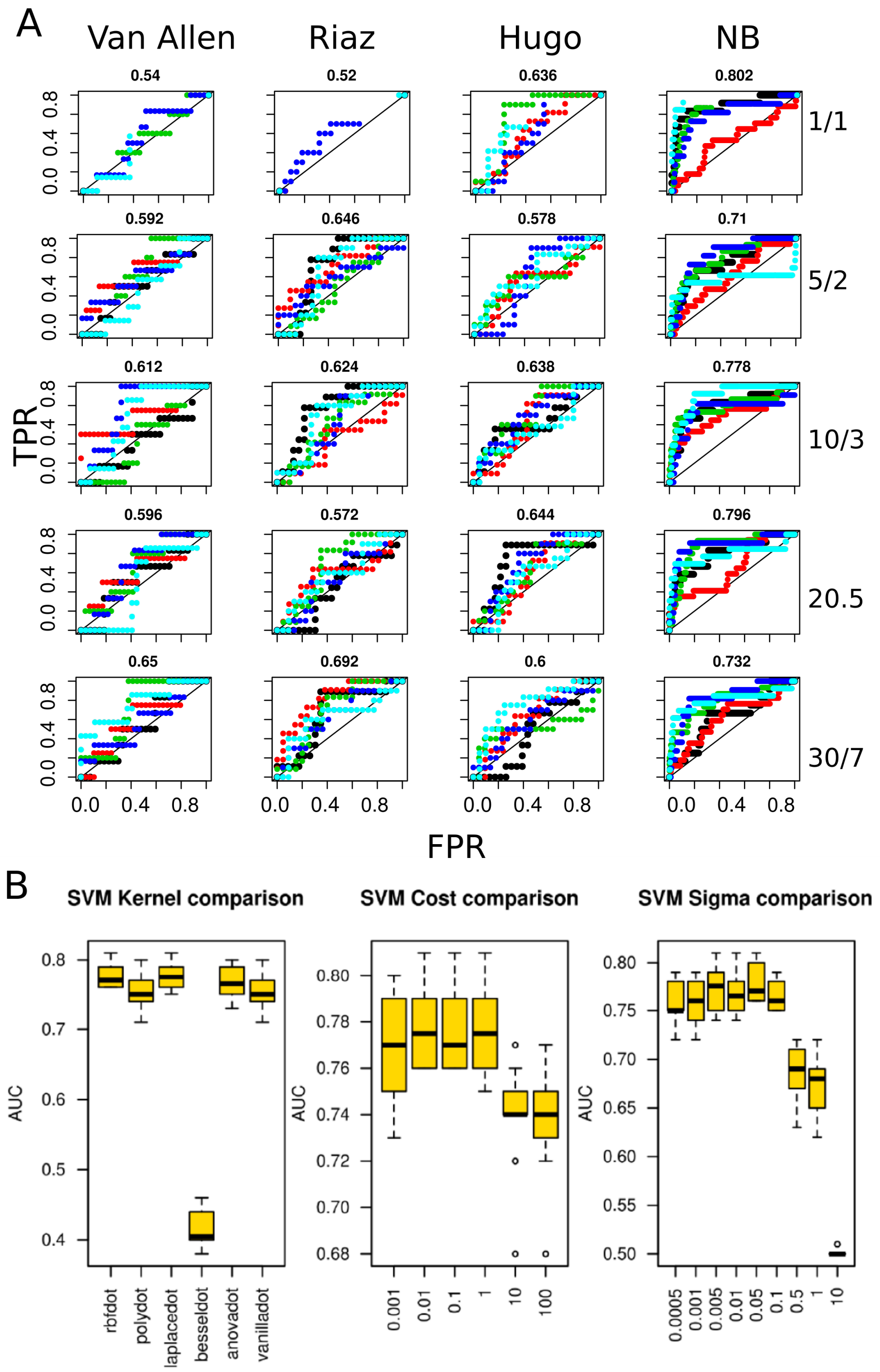


Figure 2: **Performance of standard approaches**. A) Neuralnet Performance on all datasets, numbers on the right denote the number of neurons used in each of the two layers. Numbers above are the mean AUC achieved across cross validation sets. B) Performance of the SVM. We first chose the right kernel (left), then we adjusted the cost parameter and then the sigma parameter.

as possible in the right way. Contrariwise, choosing a smaller value results in higher classification errors, but will be safer with regards to overfitting [10].

- *sigma*: As the best performing kernel for all the datasets was the `rbfdot` gaussian kernel, we can additionally adjust for σ which specifies the variance of our gaussian kernel. The highest AUC for each dataset is listed in table (Table 1 and Figure 2B).

THE IMPRS APPROACH As the above presented methods yielded moderate performance, we referred to a recently published [1] sophisticated approach developed specifically to predict the response of immune checkpoint therapy in metastatic melanoma. Briefly, we first find gene-pairs (=features) which are highly predictive for the outcome of the therapy.
Based on these features, the prediction of spontaneous regression of a tumor sample from its expression data is simply made by counting the number of predictive feature pairs that are fulfilled (true) in that sample given its transcriptomics data. This number, ranging from 0 to 15, denotes its IMPRS score, with higher scores predicting spontaneous regression. The resulting predictor obtains an value of 0.81 (in terms of the area under the receiver operator curve (AUC) in the NB dataset.

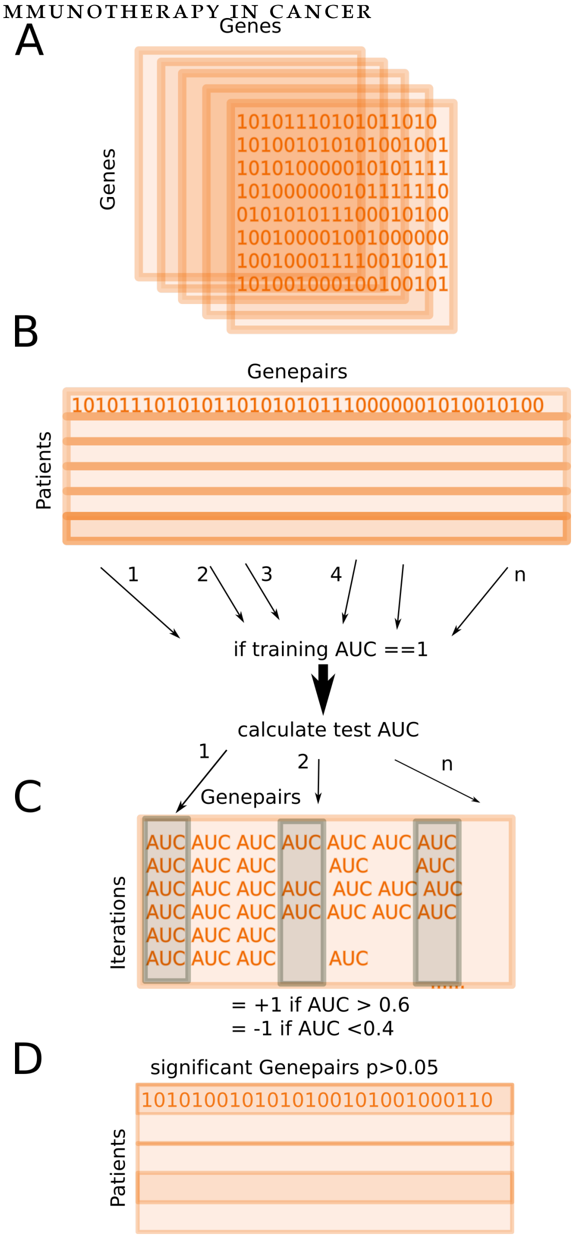


Figure 3: **Sketch of the Algorithm Design:** A Construction of the Binary Table: For each patient p we calculate the the genepairs gg and the table is in the shape of $ggxgg$. Then we reshape the table to have a $pxgg$ shape (B), n is set to 20 in our case. Then we calculate an AUC using a random sampling of genepairs as features. The feature maximizing our AUC will be taken into account for the final selected gene set. Genepairs in iterations with an AUC > 0.6 are counted as one, otherwise 0. We then use a threshold of $p < 0.05$ above which we take the genes as selected.

COMPUTATIONAL STEPS We started from the same generalized table containing as in the previous approaches. We then applied a hill climbing strategy in 500 iterations which consists of the following steps (Figure 3).

- initialize the current set of features to an empty list.
- select randomly 26 samples (13 in each class) for the $trainingSet$ and 6 (3 in each class) for the $testSet$.
- select randomly n features, initialize the variable $trainingAUC$ to 0

while the current $trainingAUC < 1$:

- Add 1 feature from the randomly selected features at a time to the $trainingSet$ set and calculate the the ROC curve and the corresponding AUC = $trainingAUC$
- after having added all the features select the one which had the highest AUC and add it to the current group of features

- update the current AUC with the new AUC

when $AUC == 1$ was reached, we calculated the AUC for the current group of features on the testSet and stored this value in a matrix in the row for this iteration and all the involved gene-pairs in the group of features (Figure 3C).

To extract patients with high vs low risk, we simply summed up the number of features that are true for every patient and extract the patients IMPRS score (Figure 3D) The most important features were extracted by calculating a score for each of the features: $score(f) = score_{+}(f) - score_{-}(f)$, where $score_{+}$ and $score_{-}$ is the number of successful iterations with $AUC > 0.6$ and unsuccessful iterations, $AUC < 0.4$, respectively, in which feature f was selected to be in the test group.
We extract the most important features using a p-value cutoff of 0.05. The features found by this procedure are listed in table S2. Note that some gene-pairs occured twice, meaning their relation was actually not important for the prediction - in contrast to what our model said for different training sets. To provide a more sophisticated solution to this, the algorithm should be adapted to treat redundant features as one feature. To extract patients with high vs low risk, we simply summed up the number of features that are true for every patient and extracted the patients IMPRS score (Figure 3D). Here we summed, every gene-pair which is selected and where the comparison is 1 (that means that $expression(X) > expression(Y)$ in gene-pair (X, Y)). This gave a patient score, the higher the score, the higher the patient was predicted to respond to the therapy.

AUC AND IMMUNE RESPONSE To validate the predictions made by our IMPRS implementation we included the clinical annotations of the samples. For the NB dataset we saw that patients with higher IMPRS score tend to show higher immune response than those with smaller IMPRS score. We achieve an AUC of 0.81 on the NB dataset with the predicted IMPRS score. Patients showing low immune response tend to have lower IMPRS scores. (see Figure 4B). The other datasets weren't performing best when using the predictor of the NB dataset. When computing the predictor for the Riaz (Figure 4C) and a combined dataset of vanAllen/Hugo (Figure 4E), the AUC improved in the cross validation test. We also computed the AUC of the prediction of our target genes. When running the IMPRS approach directly on the data itself, considerable high AUCs were achieved. (for the NB and Riaz dataset: Figure 4B and for the van Allen dataset: Figure 4E). Note that the van Allen and Hugo dataset were combined.
For the Riaz dataset, patients were tested twice, once before the anti-PD1 therapy started and once during therapy. We computed the IMPRS score for both groups of patients. Those on therapy showed higher IMPRS scores, which means that the gene expression changes significantly and immune response checkpoint genes change their expression patterns due to the therapy (Inlay Figure 4C). In conclusion, the IMPRS approach yielded higher AUCs as the other standard approaches.

SURVIVAL CURVES High IMPRS scores predict a good response to the immunotherapy, so we expected those samples to survive for a longer time. The clinical data provided access to the overall survival times for two of the datasets. In the Riaz dataset (Figure 4D), we observe a distinct survival for groups defined by "high" if IMPRS>median(IMPRS), and "low" if IMPRS<=median(IMPRS). The survival of Patients

Dataset	Neural net	`neuralnet`	`kernlab`	IMPRS
NB	0.5	0.76	0.77	0.81
Riaz	0.51	0.55	0.63	0.7
vanAllen	0.52	0.59	0.62	0.68
Hugo	0.52	0.64	0.67	0.71

Table 1: **Overview over the AUC values** obtained by 10-fold cross validation (1-3) and IMPRS prediction across the datasets

showing higher IMPRS scores differs significantly from those having lower scores. The survival rate in the van Allen dataset (Figure 4F) was slightly less significant, still there was a shift towards longer survival in patients with a higher IMPRS score. That underlines again the importance of immune therapy in cancer, which is about extending the lifespan.

6.3 DISCUSSION

We conducted a study on 4 different datasets to find out how to predict the response to immune therapy efficiently and robustly in two different cancer types. Algorithms off the shelf like the R packages `neuralnet` and `kernlab` provided considerable accurate predictions. We followed, the customized – self implemented approach proposed by Auslander et al [1] loosely which yielded higher AUC values for most of the datasets (Table 1).

We obtained AUC values above 0.7 with cross validation for the datasets that we based our predictor on. This predictor however didn't perform that well when applied to the other datasets as opposed to the study of Auslander et al [1], it was enough to base the predictor on the NB dataset, and still the AUC for other datasets was around 0.8. Certainly, they used a more sophisticated feature selection and used mRNA data to construct the predictor. In our case, it would be interesting computing the predictor on a combination of all 4 datasets and check the AUC when using this predictor on the single datasets. All in all, the approach is very powerful, considering that it achieved the highest AUC values across most datasets. With calculation of the IMPRS score, we were able to capture patients that would respond to a immune blockade therapy, and generally patients with higher IMPRS scores also survived longer. In conclusion, one guideline to predict the possibility of response to ICB is to look at the immune checkpoint genes we found. Further studies on sufficiently large datasets could be based on this approach to further ameliorate the predictive performance with respect to ICB.

6.4 METHODS

DATASETS We analyse four different publicly available datasets:

- Neuroblastoma dataset (NB) [3]: An extensive set of primary neuroblastoma taken from patients of ages up to 24 years old (GSE49710). Figure 1 shows the complexity of the dataset: None of the clinical parameters are clearly separating the first components from each other.

- van Allen et al [21]: transcriptome data from pretreatment tumor samples of melanoma tumor biopsies from 40 patients.
- Hugo et al (H) [5]: transcriptomes of responding (n = 15) and non-responding (n = 13) pretreatment melanoma tumors (total 27 of 28 pretreatment; 1 of 28 early on-treatment)
- Riaz et al [6]: whole exome sequencing of 68 patients pre- and during therapy

the log transformed data was quantile normalized before any analysis with the R package `preprocessCore` and the function `quantile.normalize`. Then we reduced the number of genes to in total 26 immune checkpoint genes (BTLA, PDCD1, CD200, CD200R1, CD27, CD276, CD28, CD40, CD80, CD86, CEACAM1, CTLA4, HAVCR2, IDO1, IL2RB, LAG, PD1LG2, PVR, PVRL2, TIGIT, TNFRSF14, TNFRSF18, TNFRSF4, TNFRSF9, OX40L, CD137L). which reduces the massive feature space in the datasets and restricts the predictions to be done entirely in regard to the immune response. The SVD is performed using the `svd` function of the R `base` package.

neural net We implemented a network following the instructions given by [20] and implemented it in R version 3.5.1. Then we used the R package `neuralnet` and texttt NeuralNetTools for visualization [22, 23, 24]. We used `neuralnet` for a sequence of different layers: 1,5,10,20,30 the function `neuralnet::neuralnet(mF,TRAINING, hidden=c(layer,(la` `threshold=0.01)`

imprs We used the the R packages `caret, stats` and `pROC`. We used the functions `caret::train` for training our training data set, `stats::predict` for the prediction on the test set, `caret::varImp` to calculate the most important feature and `pROC::auc` to draw the ROC curve.

survival curves We use the R package `survival` [25] to draw the survival curves with `survival::Surv` and `survival::survfit` and `survminer::ggsurvplot` [26] for drawing the curves.

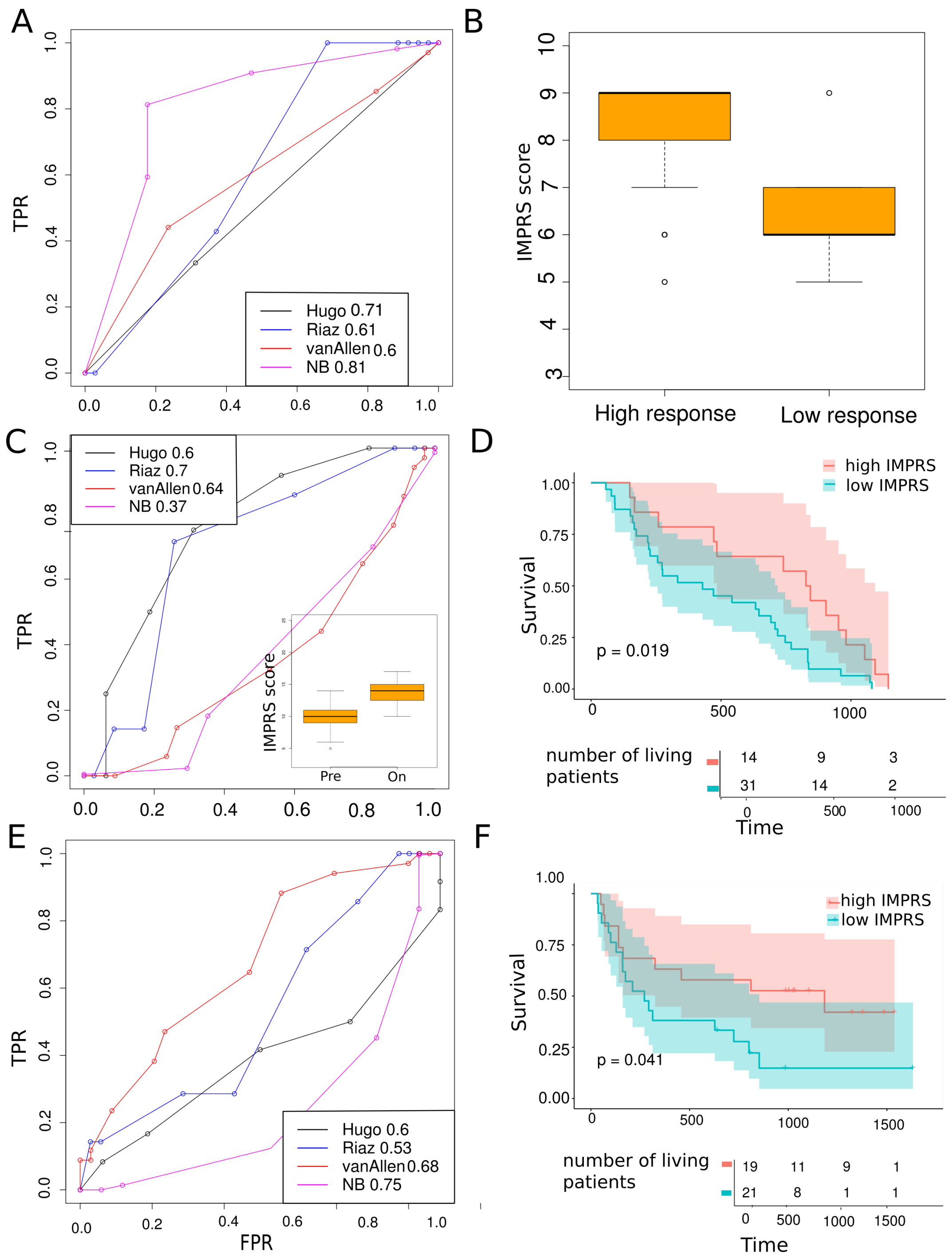


Figure 4: **Performance of the IMPRS approach across different datasets** A) Using the predictor trained on the neuroblastoma dataset (NB) B) IMPRS scores of patients with high response and low response to therapy in the NB dataset. C) Performance of the IMPRS approach using the predictor trained on the Riaz dataset. Inlay shows the IMPRS score of patients already on therapy and before the onset of therapy. D) Survival curve of patients in the Riaz dataset, 'high IMPRS' means patients with IMPRS score above the mean and 'low IMPRS' below the mean. Shaded area denotes the 95% confidence interval. E) Performance of the IMPRS approach using the predictor trained on the vanAllen/Hugo dataset. F) Survival curve for the vanAllen/Hugo dataset for high and low IMPRS scores, computed as in D).

6.5 supplementary material

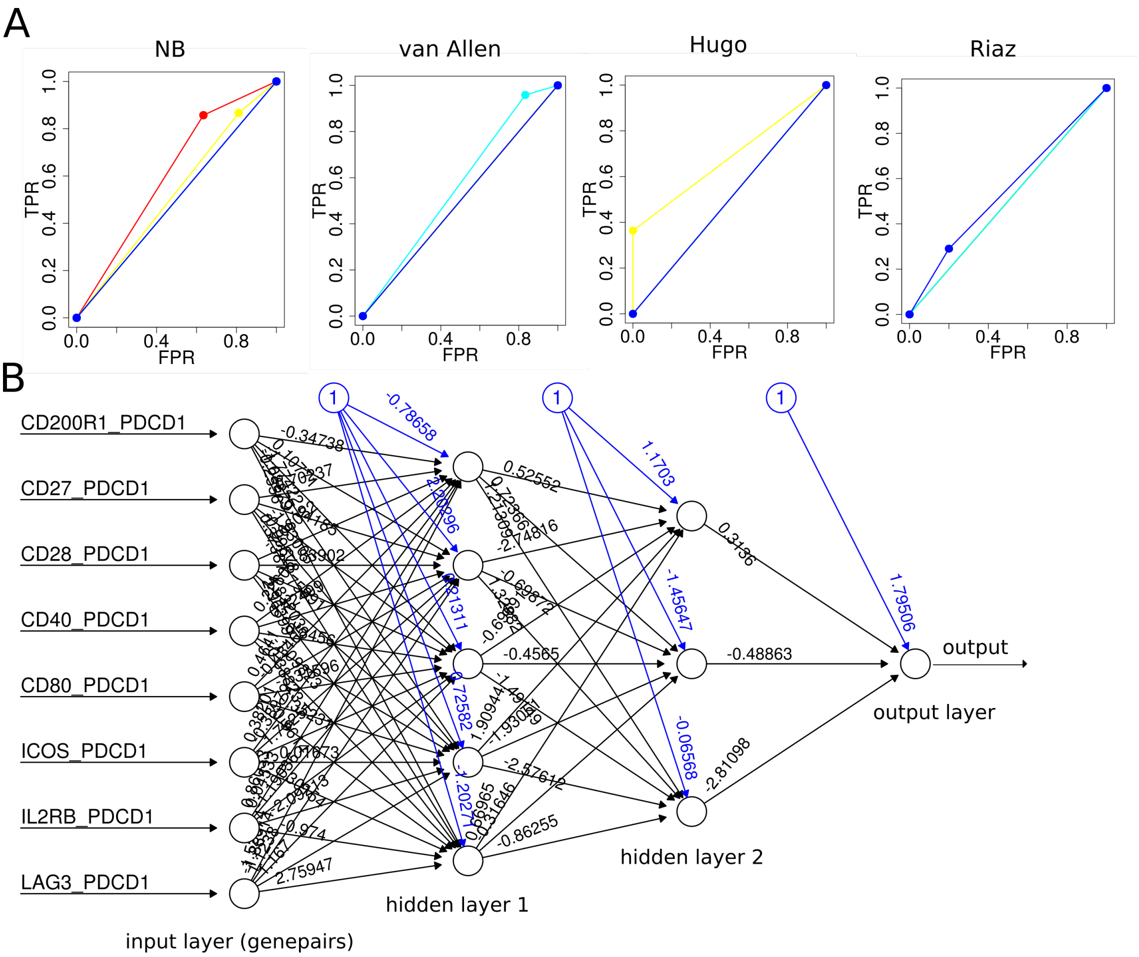


Figure S1: A) Neuralnet performance for all 4 datasets separately. B) Neuralnet Performance on all datasets, numbers on the right denote the number of neurons used in each of the two layers. Numbers above are the max AUC achieved across different training sets.

Gene	Publication
BTLA	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
PDCD1	https://www.ncbi.nlm.nih.gov/pubmed/29076134
CD200	http://www.cell.com/immunity/pdf/S1074-7613(16)30151-0.pdf
CD200R1	http://www.cell.com/immunity/pdf/S1074-7613(16)30151-0.pdf
CD27	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
CD276	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
CD28	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
CD40	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
CD80	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
CD86	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
CEACAM1	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4481276/
CTLA4	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
HAVR2	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
IDO1	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
IL2RB	http://www.cell.com/immunity/pdf/S1074-7613(16)30146-7.pdf
LAG3	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
PD1LG2	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
PVR	http://www.cell.com/immunity/pdf/S1074-7613(16)30146-7.pd
PVRL2	http://www.cell.com/immunity/pdf/S1074-7613(16)30146-7.pd
TIGIT	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
TNFRSF14	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
TNFRSF18	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
TNFRSF4	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
TNFRSF9	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
OX40L	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786574/
CD137L	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6607944/

Table S1: Immune checkpoint genes taken from literature for feature selection

NB	vanAllen Hugo	Hugo Riaz
CD200R1<CD27	BTLA<CD200R1	BTLA<CD80
CD200R1<PDCD1	BTLA<CTLA4	CD200<CD40
CD27<CD200R1	BTLA<HAVCR2	CD200<CD86
CD27<IL2RB	CD200<CD200R1	CD200<HAVCR2
CD40<CD200R1	CD200<CD40	CD200R1<BTLA
CTLA4<BTLA	CD200<HAVCR2	CD200R1<TNFRSF18
CTLA4<TNFRSF9	CD200<IL2RB	CD276<CD80
IL2RB<CD27	CD276<CD200R1	CD276<PDCD1
IL2RB<PDCD1	CD276<CEACAM1	CD28<IL2RB
IL2RB<CD27	CD40<TNFRSF18	CD40<PDCD1
PDCD1<IL2RB	CD80<BTLA	CD40<TNFRSF18
PDCD1<TNFRSF4	CD80<HAVCR2	CD80<CD200
TNFRSF18<CD200R1	CD80<PDCD1	CD80<IL2RB
TNFRSF4<PDCD1	CD80<TNFRSF14	CEACAM1<CD276
	CEACAM1<TNFRSF18	CEACAM1<IL2RB
	CTLA4<CD86	CEACAM1<TNFRSF18
	CTLA4<PVR	CTLA4<CD276
	CTLA4<TNFRSF14	HAVCR2<CEACAM1
	HAVCR2<BTLA	IDO1<BTLA
	HAVCR2<TNFRSF18	IDO1<CD28
	PVR<CD200	IDO1<CD86
	IL2RB<CEACAM1	IL2RB<CD80
	IL2RB<TNFRSF18	PVR<CTLA4
	PDCD1<HAVCR2	
	PVR<PDCD1	
	TIGIT<CD28	
	TNFRSF14<CD200	
	TNFRSF18<CD200	
	TNFRSF18<HAVCR2	
	TNFRSF18<TNFRSF14	
	TNFRSF4<CD40	

Table S2: Selected Genes for IMPRS computed on a combination of Hugo/vanAllen and Hugo/Riaz Dataset $p < 0.1$. Note that sometimes the features are redundant. We propose an additional superior approach which just takes into account single gene pairs.

IMMUNE CHECKPOINT GENES FROM LITERATURE

BIBLIOGRAPHY

[1] Auslander, N., Zhang, G., Lee, J. S., Frederick, D. T., Miao, B., Moll, T., ... Boland, G. (2018). Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nature medicine, 1.

[2] Sweilam, N. H., Tharwat, A. A., & Moniem, N. A. (2010). Support vector machine for diagnosis cancer disease: a comparative study. Egyptian Informatics Journal, 11(2), 81-92.

[3] Su, Z., Fang, H., Hong, H., Shi, L., Zhang, W., Zhang, W., ... & Yang, X. (2014). An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. Genome biology, 15(12), 1.

[4] Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., ... & Utikal, J. (2015). Genomic correlates of response to CTLA4 blockade in metastatic melanoma. Science, aad0095.

[5] Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., ... & Seja, E. (2016). Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell, 165(1), 35-44.

[6] Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., ... & Bhatia, S. (2017). Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell, 171(4), 934-949.

[7] Cicchetti, D. V. (1992). Neural networks and diagnosis in the clinical laboratory: state of the art. Clinical chemistry, 38(1), 9-10.

[8] Simes, R. J. (1985). Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. Journal of chronic diseases, 38(2), 171-186.

[9] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[10] Phan, J., Moffitt, R., Dale, J., Petros, J., Young, A., & Wang, M. (2006, January). Improvement of SVM algorithm for microarray analysis using intelligent parameter selection. In 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference (pp. 4838-4841). IEEE.

[11] Kalbasi, Anusha, and Antoni Ribas. "Tumour-intrinsic resistance to immune checkpoint blockade." Nature Reviews Immunology (2019): 1-15.

[12] Diede, Scott J. "Spontaneous regression of metastatic cancer: learning from neuroblastoma." Nature Reviews Cancer 14.2 (2014): 71-72.

[13] Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). “kernlab – An S4 Package for Kernel Methods in R.” Journal of Statistical Software, 11(9), 1–20. http://www.jstatsoft.org/v11/i09/.

[14] Jessy, Thomas. "Immunity over inability: The spontaneous regression of cancer." Journal of natural science, biology, and medicine 2.1 (2011): 43.

[15] Ogawa, Ryoko, et al. "Lung cancer with spontaneous regression of primary and metastatic sites: A case report." Oncology letters 10.1 (2015): 550-552.

[16] Salman, Tarik, Journal of Oncological Science Volume 2, Issue 1, April 2016, Pages 1-4

[17] Pardoll, Drew M. "The blockade of immune checkpoints in cancer immunotherapy." Nature Reviews Cancer 12.4 (2012): 252-264.

[18] Kourou, Konstantina, et al. "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

[19] Morandi, Fabio, et al. "Novel immunotherapeutic approaches for neuroblastoma and malignant melanoma." Journal of immunology research 2018 (2018).

[20] David Selby (9 January 2018),Building a neural network from scratch in R, retrieved from `https://selbydavid.com/2018/01/09/neural-network/`

[21] Van Allen, Eliezer M., et al. "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma." Science 350.6257 (2015): 207-211.

[22] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[23] Beck MW (2018). "NeuralNetTools: Visualization and Analysis Tools for Neural Networks." Journal of Statistical Software, 85(11), 1–20. doi: 10.18637/jss.v085.i11.

[24] Beck MW (2018). "NeuralNetTools: Visualization and Analysis Tools for Neural Networks." Journal of Statistical Software, 85(11), 1–20. doi: 10.18637/jss.v085.i11.

[25] Terry M. Therneau, Patricia M. Grambsch (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York. ISBN 0-387-98784-3.

[26] Alboukadel Kassambara, Marcin Kosinski, Przemyslaw Biecek, Scheipl Fabian, survminer, R package, downloaded 2018

[27] Zhang, Wenqian, et al. "Comparison of RNA-seq and microarray-based models for clinical endpoint prediction." Genome biology 16.1 (2015): 133.

7 CONCLUSION, DISCUSSION AND OUTLOOK

7.1 CONCLUSION

High throughput sequencing techniques have opened up a whole new field of research. Now it is possible to get a complete snapshot of the regulatory state of cells or tissues by sequencing the DNA or RNA, which can even be supplemented by screening for accessible regions, transcription factor binding or modification of histones. Thus, during the last decade, the amount of researchers using high throughput sequencing techniques has increased drastically. As a result, a multitude of computational tools have been developed, dealing with all the different steps of the analysis. Yet those tools rely on a number of parameters that are individual for each dataset and that users have to supply. This results in an increased problem of reproducibility, when the exact documentation of parameters, version control and architecture of the operating system is missing. Moreover, after the successful analysis, most researchers investigating genome wide chromatin changes are left with highly dimensional data containing information about single regions where accessibility or binding changes. It is very hard to solely determine potential interesting candidates for downstream analysis from this kind of data.

In this thesis, we applied three different strategies to analyse high-throughput sequencing data. The focus is our CREMA tool, which provides the first complete pipeline to analyze and compare genome wide chromatin state across samples by concretely modeling the data in terms of key regulators being active in the system (chapter 4). We combined sophisticated statistical models for the determination of regulatory relevant regions, CRUNCH [2], with a powerful linear modeling approach, ISMARA [1]. Our method allows to express the genome wide changes in DNA accessibility or transcription factor binding across any number of datasets in terms of regulatory key players. The highly dimensional data becomes thus much easier to interpret by narrowing down the list of potential key players driving the observed system.
To our understanding, no similar approach has yet been performed. We then applied different computational strategies to two other important questions: We use ISMARA [1] and SVD to predict the temporal response of exercise invention on mice (chapter 5) and contemporary machine learning techniques to dive into the classification of cancer patients (chapter 6). Our different approaches show the variety of applications of computational tools in biology. I strongly believe that this is only the advent of a whole new field of science and that it will open our eyes to numerous new observations and enhance our understanding of biological processes and diseases tremendously.

7.2 DISCUSSION

The novelty of our CREMA tool is clearly the genome-wide scanning for regulatory relevant regions across a non-restricted number of samples and further setting them into context by finding binding sites for transcription factors and predicting targeted genes and associated pathways. The genome wide approach allows the inclusion of unknown, yet important enhancer elements to rigorously explain the underlying system. CREMA successfully infers known regulators in already investigated systems and provides an extensive collection of potential new regulatory key players and circuits in proximal and distal regions. We believe that the application of CREMA to the increasing amount of genomic sequencing data can contribute substantially to our understanding of gene regulatory networks, epigenetic regulation and the roles of enhancer elements.

The quantitative comparison of sequencing data across samples is always a very challenging task, given that some systems are much harder to treat as others and thus are more likely to show biases due to experimental errors or sampling issues. Given the complexity of the underlying systems, it is often very hard to disentangle which fluctuations relate to *real* biological differences and which ones are due to experimental errors or biases in sequencing. Our normalization strategy includes three steps throughout the pipeline and we hope to get rid of most biases introduced at the experimental or sequencing level. Especially in ChIP seq, differing efficiency of the antibody and real biological binding strength changes cannot be uncoupled by modeling so far. An experimental validation to keep the level of efficiency approximately at the same level would help to get rid of this uncertainty and help interpreting the data. As the amount of paired end-sequencing is increasing drastically, a future refinement of the algorithm would be to use the actual fragment size uniquely for shifting the corresponding reads. As the observed regions are usually larger than the expected in fragment length across the reads for each sample, we don't see this as ultimately necessary. Moreover, we find a high reproducibility of the model when changing e.g. the sliding window size or even taking fixed region lengths for all our CREs.

For our predictions, we make use of a set of selected weight matrices for mouse and human. It is estimated that roughly 1000-1500 transcription factors exist in mammals and up to 3000 in human [3, 4]. Hence, we are fully aware, that our collection of ≈ 600 transcription factor binding sites cannot represent the whole regulatory apparatus. Note that it is not our intention to model and predict regulatory circuits in detail, our tool can be applied to get an overview over the general mechanisms explaining parts of the data to facilitate picking potential interesting candidates for an experimental follow up analysis. The linear model is easily solvable but of course limited in terms of complexity of the predictions. However, a true advantage of this 'simple' model is its robustness. As we sum over a large amount of regions, small fluctuations in potentially false positively predicted CRE regions have only a minor effect on the final predictions. This is true in the region size as well: enlarging the regions or even setting them all to the same length yields only in slight aberrations from the initial results.

To assess the performance of our tool, we applied it to several published datasets:
Our results show that our algorithm successfully infers activities of transcription factors, that have been known to steer regulatory processes in the examined conditions. By analysis of published ATAC seq of liver samples taken from mice left in darkness for several days, we are able, firstly, to reconstruct important key players in the circadian oscillation of the liver transcriptome and, secondly, to predict new potential circadian regulators. Fitting a harmonic curve underlines the current hypothesis, that

the liver clock is uncoupled from the optical input of light and darkness, because most regulators continue cycling. We are aware, that inclusion of more data points (e.g. two 24h periods) would drastically increase our confidence in predicted factors.
In embryonic development, a process highly regulated by chromatin state, we find some transcription factors almost exclusively in few tissues, whereas others seem to take a general function in development and increase their activity with time in almost all tissues. However, we find that most motifs either increase or decrease their activity (almost monotonically), these pattern explain almost half of the variance across the dataset. With integration of mRNA seq data we were able to compare the resulting activity pattern between ATAC seq and RNA seq, and could underline the importance of looking at gene regulation genome wide: We explicitly find motifs occurring predominantly at distal regions, whereas others seem to drive transcription from promoters and do not rely on opening or closing of the DNA.
Also our last dataset encourages looking at distal regulatory regions: ChIPped regions for the histone mark H3K4me1 were much more variable across cell types, suggesting a large portion of the cell-specific regulation happens at enhancers. Consistent with previous findings, H3K4me1 marked regions lay further away than promoter regions marked by H3K4me3.
These results confirm the various possibilities of applying CREMA to experimental data to find and quantify all the regulatory elements that drive the system.
One limit here, of course, is that some of the motif sequences are quite similar or very abundant, what could lead to a false positive prediction. One way to counteract this from the beginning would be to include only factors that have been found to by expressed by RNAseq for exactly the same experiment (which needs to be done very thoughtful, as a change in activity doesn't necessarily imply a change in mRNA expression). An experimental validation of the predictions would be essential to confidently claim the involvement of this factor in the regulation of the underlying system.
Next, we applied computational strategies to answer two very different questions. First, we were interested in the temporal regulation of exercise response after an acute training. We let mice perform an intensive exercise bout and sequenced the quadriceps mRNA at different timepoints after. Using ISMARA [1] and SVD, we find three dominant pattern across our timepoints. Further, our cohort consisted of wild type animals and PGC1a knockout animals. PGC1a is known to be a key node in exercise adaptation, so we expected differences between the WT and KO mice following acute exercise. However, we didn't find PGC1a to be mandatory for the very first step of gene regulation but, we saw that PGC1a is able to keep up important factors involved in immune response and autophagy at a higher level than the KO at later timepoints. This may be one explanation of the inferior ability of mice lacking PGC1a to adapt to exercise.
A very interesting follow up study could focus on how this sustained activity is achieved: It may be because the proteins of regulators themselves are stabilized and thus initiate the transcription much longer, or it may be because the regulatory region of their downstream targets are silenced later (e.g. by DNA methylation). We could even suggest to tackle the question of how *muscle memory* works, an effect that facilitates retraining in trained athletes which has not been understood until now.

In our last study, we applied machine learning techniques to four different patient datasets on melanoma and neuroblastoma. The data was clinically annotated and the goal was to predict whether a patient will respond to immune checkpoint blockade therapy. We built an algorithm to calculate a *score* for each patient and our predictions were mostly right, but the AUC depended strongly on the dataset that was used for training. Also, our features (marker gene pairs) were sometimes redundant. A further approach could improve this by just taking one genepair at once in consideration. Lastly, with the amount of data rising, it would be very interesting to train the algorithm on a larger set of data and make predictions accordingly.

7.3 OUTLOOK

We believe that the fully automated implementation of CREMA will encourage experimental groups to use our predictions to analyse their datasets and that this will help in uncovering new parts of the implication of chromatin state in the regulation of gene expression. Integration of ATAC seq and RNA seq would yield high potential in classifying functions of transcription factors. Our tool could be interesting for the pharma industry as well: Knowing key regulators in certain diseases would be the first step in reconstructing causal dependencies which shape the pathologic phenotype – and with this knowledge – construct follow up studies and finally a therapy.

With new techniques arising constantly, we are confident that the understanding of gene regulation will increase further. We have shown that the chromatin state is highly variable across tissues, timepoints and cell types. However, our analysis has been based on the assumption that these tissues are homogeneous. Yet, multicellular eukaryotic tissues are often a convolute of many different cell types with different physiological functions, morphological features and molecular markers [5]. Especially in cancer and stem cell development, it is important to capture subgroups of cells. For example, a cancerous tumor is made up of diverse cells, including malignant, immune, and stromal subsets, whose precise characterization is masked by bulk genomic methods [6]. By scRNA seq, those cells can be annotated based on the level of gene expression of certain marker genes along with point mutations, and fusion proteins. Current research already adjusts the ISMARA model to function with scRNA seq. Very likely, the difference measured in gene expression is governed by the chromatin state of these cells. Therefore, methods to measure chromatin accessibility in single cells have been developed [7]. We strongly think that the adaptation of or method CREMA would be very insightful for the analysis and regulatory circuits of single cells, and even help classifying the underlying cell types according to their chromatin state.

Our approach could further be refined by having exact information on enhancer regions and associated genes. Even though sole determination of enhancers is nowadays possible by measuring histone modifications, enhancer RNA (eRNA) or intraspecies sequence conservation, no reliable general database exists. Plus, although several new techniques and experiments are constantly emerging [9, 10, 11], the functional association of enhancers to genes imposes a big challenge. Using our model already supplies the researcher with possible functions of an enhancer by reporting the binding sites for specific transcription factors whose function is – sometimes – well studied. Given our finding that distal regions are much more variable in terms of regulatory activity, it would add a lot of precision to our predictions and deepen our understanding of the underlying systems.

Taken together, our novel ways of analysis of high - throughput data brings us one step closer to understanding the complex and versatile structure of gene regulatory networks. Given recent insights that regulation happens genome wide, even in distal regions, we analyse the whole genome and predict which transcription factors may be responsible for the observed chromatin state. Given the increasing amount of next generation sequencing data, and the easily accessible interface, we can provide researchers with a dedicated tool to investigate their samples. Hopefully, this will add a important piece to the highly dimensional puzzle of gene regulation.

BIBLIOGRAPHY

[1] Balwierz, Piotr J., et al. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs." Genome research 24.5 (2014): 869-884.

[2] Berger, Severin, et al. "Crunch: integrated processing and modeling of ChIP-seq data in terms of regulatory motifs." Genome research 29.7 (2019): 1164-1177.

[3] Babu, M. Madan, et al. "Structure and evolution of transcriptional regulatory networks." Current opinion in structural biology 14.3 (2004): 283-291.

[4] Zhou, Quan, et al. "A mouse tissue transcription factor atlas." Nature communications 8.1 (2017): 1-15.

[5] Amamoto, Ryoji, et al. "Probe-Seq enables transcriptional profiling of specific cell types from heterogeneous tissue by RNA-based isolation." eLife 8 (2019).

[6] Suvà, Mario L., and Itay Tirosh. "Single-cell RNA sequencing in cancer: lessons learned and emerging challenges." Molecular cell 75.1 (2019): 7-12.

[7] Shema, Efrat, Bradley E. Bernstein, and Jason D. Buenrostro. "Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution." Nature genetics 51.1 (2019): 19-25.

[8] Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. "Transcriptional enhancers: from properties to genome-wide predictions." Nature Reviews Genetics 15.4 (2014): 272-286.

[9] Mifsud, Borbala, et al. "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C." Nature genetics 47.6 (2015): 598.

[10] Andersson, Robin, et al. "An atlas of active enhancers across human cell types and tissues." Nature 507.7493 (2014): 455-461.

[11] Whalen, Sean, Rebecca M. Truty, and Katherine S. Pollard. "Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin." Nature genetics 48.5 (2016): 488.

ACKNOWLEDGMENTS

Life is like riding a bicycle.
To keep your balance, you must keep moving.
Albert Einstein

... What Albert Einstein said is true for my PhD "journey" too: Here I want to thank all the people who shared their knowledge with me, motivated, supported and guided me. Without you, I wouldn't have kept moving!

First of all, I would like to thank Prof. Erik van Nimwegen and Prof. Christoph Handschin for their constant supervision and motivation throughout the process of my PhD. Both found a way to encourage me to continue working on my projects and keep progressing, in good and tough times. Thank you for sharing and explaining so much knowledge with me. I would like to especially thank Erik for his good and useful input and guidance while working on our computational tool and paper, and Christoph for motivating me even if projects didn't work out the way we wanted and the guidance and help writing our Review.

A special thank you goes to Prof. Dirk Schübeler, who agreed on very short notice to be part of my PhD Committee.

Thanks to Dr. Mikhail Pachkov for immediate and competent help with everything related to computational questions and implementation of our tool. Even I had lots of questions throughout the entire process, your patience was just unbreakable.

A special thanks also goes to Arantxa, Luca and Athos from the van Nimwegen Group, for always being open for discussion scientific and personal topics and finding motivating words to keep me up in tough times. I will never forget the retreats and activities we spent together. I thank especially Dany and Jeremie for his advice on the CREMA paper. Of course, I thank all the current and former members of the van Nimwegen Group for being always helpful and answering all kinds of questions.

I thank Martin, Geraldine, Julien, Fabienne and Laura and all other people from the Handschin Group. It really enjoyed your company throughout my PhD.

Lastly, my family and friends were very important to me: Danke vor allem an meine Eltern, die mir diesen Weg ermöglicht haben. Danke and meinen Vater, danke dass du immer für mich da warst und mich unterstützt hast. Deine unvoreingenomme und faire Sichtweise hat mich oft auf den Boden der Tatsachen geholt und mir sehr bei vielen Entscheidungen geholfen.
Danke an meine Mutter, danke für deine uneingeschränkte Unterstützung, viele ermutigende Telefonate und großes Verständnis.
Natürlich danke ich auch Andreas, der immer an meiner Seite, war und auch meine

schlechten Launen stoisch ertragen hat. Ich danke dir sehr für deine unendliche Geduld und den Glauben an uns.

LIST OF FIGURES

LIST OF TABLES

APPENDIX

1) Anne I Krämer and Christoph Handschin: *How Epigenetic Modifications Drive the Expression and Mediate the Action of PGC1α in the Regulation of Metabolism*, Int. J. Mol. Sci. 2019, 20(21), 5449

International Journal of
Molecular Sciences

MDPI

Review

How Epigenetic Modifications Drive the Expression and Mediate the Action of PGC-1α in the Regulation of Metabolism

Anne I. Krämer [1,2] **and Christoph Handschin** [1,*]

[1] Biozentrum, University of Basel, 4056 Basel, Switzerland; anne.kraemer@unibas.ch
[2] Swiss Institute of Bioinformatics, 4056 Basel, Switzerland
[*] Correspondence: christoph.handschin@unibas.ch; Tel.: +41612072378

Received: 30 September 2019; Accepted: 29 October 2019; Published: 31 October 2019

Abstract: Epigenetic changes are a hallmark of short- and long-term transcriptional regulation, and hence instrumental in the control of cellular identity and plasticity. Epigenetic mechanisms leading to changes in chromatin structure, accessibility for recruitment of transcriptional complexes, and interaction of enhancers and promoters all contribute to acute and chronic adaptations of cells, tissues and organs to internal and external perturbations. Similarly, the peroxisome proliferator-activated receptor γ coactivator 1α (PGC-1α) is activated by stimuli that alter the cellular energetic demand, and subsequently controls complex transcriptional networks responsible for cellular plasticity. It thus is of no surprise that PGC-1α is under the control of epigenetic mechanisms, and constitutes a mediator of epigenetic changes in various tissues and contexts. In this review, we summarize the current knowledge of the link between epigenetics and PGC-1α in health and disease.

Keywords: peroxisome proliferator-activated receptor γ coactivator 1α (PGC-1α); exercise; metabolism; epigenetics; histone modification; DNA methylation; micro RNA; gene regulation; thermogenesis; metabolic diseases

1. Introduction

The term epigenetics originally described how phenotypic traits could be inherited without alterations in the DNA sequence of the genome [1,2]. In recent years, this term has been expanded and used in a more inclusive way to include non-heritable, even short-term plastic events. Often, the latter are triggered by changes in the environment and drive the adaptations to external stimuli, e.g., those exerted by exercise, fasting or high-fat diet [3–5]. In fact, in many of these contexts, epigenetic changes are integral to an adequate transcriptional response, and dysregulation of such changes have been linked to the etiology and/or pathology of various diseases. The peroxisome proliferator-activated receptor γ coactivator 1α (PPARGC1A, also called PGC-1α) is a central regulator of mitochondrial function and cellular metabolism, important for the adaptation of different tissues to increased energetic demand [6,7]. Accordingly, the gene expression of PGC-1α is strongly regulated when phenotypic changes of an organ require an increased production of ATP. Once activated, PGC-1α coordinates complex and tissue-specific transcriptional networks that mediate cellular plasticity. Soon after its discovery, epigenetic mechanisms have been linked to the action of PGC-1α as a transcriptional coactivator [8–10]. More recently, epigenetic changes have been identified to control the gene expression of PGC-1α in physiological and pathological contexts [11–13]. In this review, we summarize the current understanding of the epigenetic regulation of PGC-1α gene expression, and the epigenetic contribution to the activity of the PGC-1α-containing transcriptional complexes in health and disease.

2. Epigenetic Mechanisms

Epigenetic regulation has originally been defined as heritable changes in gene expression that do not involve DNA sequence alterations, hence mostly focused on DNA methylation and histone protein modifications [1,2,14]. However, more recent work has clearly demonstrated that these and other epigenetic changes can also occur short-term and in a transient manner. Thus, other mechanisms, for example microRNAs (miRNAs), mRNA modifications, long non-coding RNAs (lncRNAs) or nucleosome positioning are now included under the umbrella term epigenetics [15,16]. For many of these, both stable as well as transient effects have now been demonstrated. Of note, many of the recent insights have been made possible by the breakthrough advances in next generation sequencing techniques.

2.1. Histone Modifications and Nucleosome Positioning

DNA strands are compacted in several layers into chromosomes, with the nucleosomes, the wrapping of the DNA around eight core histones, as the first layer [17]. A condensed packaging is intrinsically repressive in regard to the binding of transcription factors, and thereby prevents unwanted transcriptional activity. Histone proteins can be posttranslationally modified at various residues, leading to changes in the chromatin structure [18,19]. The integration of the consequences of methylation, acetylation, phosphorylation and/or ubiquitination of histones thereby determines DNA accessibility for transcription factors, the degree of condensation of the chromatin, or long-range interactions between distal regulatory elements. Histone modifications can be stable as well as transient, the latter being an obligatory event in transcriptional regulation of gene expression. Many of the histone modifying enzymes have been identified, in particular those involved in histone acetylation (histone acetyl transferases, HATs) and methylation. Histone acetylation events have been linked to relaxation of chromatin packing, and thus facilitation of transcription factor and RNA polymerase binding [20]. The functional outcome of histone methylation is more complex and dependent on the modification of specific sites [21]. Histone lysine residues can be mono-, di- or tri-methylated, and act as activating or repressing marks. For example, mono-methylation of lysine 9 or lysine 27 of histone 3 (H3K9 and H3K27) is generally associated with transcriptional activation, di- or tri-methylated H3K4me2/3 with transcription factor binding regions and increased gene expression, whereas mono-methylated H3K3me1 often marks enhancer regions, and H3K27me3 or H3K9me3 are repressive marks [22,23]. For many of the known histone modifications, the exact consequence is still unclear, and additional mechanisms have been proposed, e.g., regulation of splicing or priming of promoters. Finally, histone modifications and DNA methylation events can act in a cooperative manner, e.g., DNA methylation-promoted methylation of H3K9 [21].

Even though the nucleosome is a stable DNA-protein complex, nucleosomes can reposition on the genomic DNA, a process called nucleosome sliding, which is independent of histone complex disruption [24]. The CCCTC-binding factor (CTCF) anchors nucleosome positions and thereby affects large transactivation domains (TADs). Moreover, nucleosome sliding is controlled by various ATP-dependent chromatin remodeling proteins, for example the SWItch/Sucrose Non-Fermentable (SWI/SNF) complex [25], leading to transcriptional activation such as large scale expression of tissue-specific genes.

2.2. DNA Methylation

Most often, DNA methylation has been linked to silencing of transcription [26,27]. Methylation events have primarily been described on the cytosine nucleotide, resulting in the formation of 5-methylcytosine (5mC) [27]. Hydroxymethylation of cytosines (5hmC) has been considered as an intermediate step towards demethylation. However, 5hmC marks are now recognized as an epigenetic marker [28]. Recently, methylation of adenosine, as originally observed in bacterial genomes, has also been found and attributed to functional outcomes in eukaryotic cells, potentially counteracting the effects of cytosine methylation [29]. Whole genome bisulfite sequencing has revealed that specific elements and regions exhibit marked differences in methylation events. For example, transposon-derived sequences are highly methylated in the human genome, presumably as a mechanism to silence these elements. In contrast, regions with a high CpG content, called CpG islands, can by hypomethylated, in particular when found in promoters or first exons. CpG islands in intergenic regions may act as distal regulatory elements, or, in particular when found in repeat regions, be important for chromosome stability [21,26,27]. Finally, CpG islands in gene bodies can affect differential promoter usage, transcription elongation or splicing. The methylation event on cytosines is mediated by a group of enzymes called DNA methyltransferases (DNMTs) [30]. Transcriptional silencing is subsequently achieved by preventing transcription factor binding and the recruitment of 5mC binding proteins, which in turn sequester histone deacetylases (HDACs). Inversely, DNA de-methylation is exerted by Ten-eleven translocation methylcytosine dioxygenases (TETs), which play an important role in the spatiotemporal control of opening genomic regions, e.g., in embryonic development [31].

2.3. miRNAs, lncRNAs, mRNA Modifications

Epigenetic changes might also be conferred by different types of RNAs [32]. miRNAs are small RNAs, of around 22 nucleotides in length, which can interact with mRNAs and thus modulate the activity of their targets in a posttranscriptional manner [33]. Long non-coding RNAs (lncRNAs) affect cellular functions in a number of different ways, for example by affecting promoter activity or mRNA translation [34]. Both types of RNAs not only act intracellularly, but are also delivered to other cells via exosomal transport [35]. Moreover, an overlap between RNA activity and other epigenetic mechanisms exists. In Arabidopsis, the miRNAs mir165 and mir166 are involved in the regulation of DNA methylation [36]. Similarly, DNMT1, -3 and -3a are all predicted targets of miRNAs [37], while miR-140 affects HDAC4 [38]. Furthermore, miR-132 fine-tunes circadian gene expression by modulation of chromatin remodeling and protein translation [39]. Finally, mRNAs are also targets for methylation events [40]. For example, the fat mass and obesity-associated protein (FTO) has been strongly associated with human obesity, and acts as an N6-methyladenosine demethylase on mRNAs, thereby affecting RNA metabolism and hence protein expression [41].

3. The Transcriptional Coactivator PGC-1α

PGC-1α is a transcriptional coactivator that was initially identified in an interaction screen with the nuclear receptor peroxisome proliferator-activated receptor γ (PPARγ) [42]. However, it is now clear that PGC-1α binds to and coactivates a large number of different transcription factors, both of the nuclear receptor superfamily as well as non-nuclear receptor-type of DNA binding proteins [6,43]. PGC-1α is the founding member of a small family of similar coactivator proteins, which also includes PGC-1β and the PGC-1-related coactivator (PRC) [44]. The PGC-1α gene is transcribed from two different promoters, and several transcript variants have been described, even though their exact regulation and function remains to be elucidated [7]. In higher mammals, PGC-1α is expressed in all tissue with a high energetic demand, e.g., brain, kidney, cardiac and skeletal muscle, brown adipose tissue and liver [45]. In most of these organs, PGC-1α gene expression and post-translational modifications are strongly regulated in a context-dependent manner, resulting in higher PGC-1α levels and activity upon internal and external stimuli that evoke an increased ATP demand, such as fasting in the liver, physical activity in cardiac and skeletal muscle, or cold exposure in brown adipose tissue [44,46]. Once activated, PGC-1α controls complex transcriptional networks that control cellular plasticity, resulting in tissue-specific gene programs controlling hepatic gluconeogenesis, thermogenesis in brown adipose tissue, or endurance exercise adaptation in skeletal muscle [6]. However, the core function of PGC-1α consists of the strong promotion of mitochondrial biogenesis and function, coupled to enhanced oxidative phosphorylation of energy substrates [47,48].

As a transcriptional coactivator, PGC-1α contains no discernable DNA binding domain. Moreover, no enzymatic activity has been attributed to this protein. Thus, mechanistically, PGC-1α relies on selective interaction with transcription factors to be recruited to target genes, and then serves as a protein docking platform to recruit other complexes. For example, via *N*-terminal interaction, PGC-1α binds to HAT complexes by interacting with p300/cAMP-responsive element binding protein (CREB), binding protein (CBP) and the sterol-receptor coactivator 1 (SRC-1) [8]. The ensuing acetylation of histones contributes significantly to the transcriptional activation of PGC-1α target genes. Similarly, recruitment of the thyroid hormone receptor-associated protein (TRAP)/vitamin D receptor interacting protein (DRIP)/mediator complex to the C-terminus of PGC-1α facilitates the interaction of the PGC-1α transcriptional complex with RNA polymerase II [9]. Moreover, the direct interaction between PGC-1α and the PPARγ interacting mediator subunit TRAP220 facilitates preinitiation complex formation and function. Finally, PGC-1α binds to the BRG1-associated factor 60A (Baf60a), and thereby promotes nucleosome remodeling and chromatin opening via SWI/SNF activity [10]. The recruitment of these different complexes are linked. For example, a mutant version of PGC-1α lacking the C-terminal domain not only lacks binding to the mediator complex, but also fails to enhance p300/CBP-dependent transcription via the still intact *N*-terminus [9].

The strong transcriptional regulation of PGC-1α gene expression, and the recruitment of several protein complexes that exert effects on histones and chromatin hint at a strong epigenetic control of PGC-1α expression and action. In the following paragraphs, we have summarized the current knowledge about the epigenetic regulation of PGC-1α in different physiological and pathophysiological contexts.

4. Regulation of Physiological PGC-1α Expression and Action by Epigenetic Mechanisms

4.1. Skeletal Muscle and Exercise

PGC-1α gene expression is strongly induced by multiple signaling pathways and stimuli in the contracting muscle fiber (Figure 1) [6]. Interestingly, PGC-1α induces its own transcription in a positive autoregulatory loop by coactivating myocyte enhancer factors 2 (MEF2) binding in the proximal promoter region [49]. However, the PGC-1α-mediated recruitment of HATs, and the resulting acetylation of histones, competes in the absence of active protein kinase D (PKD) with binding of HDAC5 to MEF2, which then mediates deacetylation of histones and transcriptional repression [50,51]. Indeed, different histone marks have been linked to the transcriptional activity of PPARGC1A—the gene encoding PGC-1α—in skeletal muscle after exercise. For example, the expression of transcript isoforms that are initiated from the distal promoter coincides with the deposition of the activation mark H3K4me3 one hour after training in murine quadriceps muscle [52]. Similarly, elevated acetylation of histone 3 was reported at the proximal promoter of rat PGC-1α in a muscle fiber type-dependent manner [53]. PGC-1α promoter activity furthermore is strongly influenced by DNA methylation events. In ex vivo stimulation experiments of mouse soleus muscle, enhanced expression of PGC-1α after 180 minutes was preceded by a decrease in DNA methylation at the promoter already after 45 minutes of stimulation [12]. In skeletal muscle in vivo, a similar reduction in promoter methylation of the PGC-1α gene was associated with elevated transcription [12]. Finally, a combination of H3K4me3 and H3K27me3 was found at the distal promoter, indicative of a poised promoter ready for rapid transcriptional activation in skeletal muscle, suggestive of the usage of poised promoters for isoform and tissue-specific expression of PGC-1α [52]. Then, the changes in DNA methylation in the PGC-1α promoter have been associated with nucleosome repositioning in this locus. Thus, after an acute endurance exercise bout, the –1 nucleosome in the PGC-1α promoter is repositioned away from the transcriptional start site by exercise and hypomethylation of the –260 nucleotide, leading to increased transcription of the PGC-1α gene [54]. Importantly, this mechanism has been linked to decreased ectopic lipid deposition in muscle, but only in high responders in regard to PGC-1α induction by exercise. Finally, the levels of muscle PGC-1α are affected by different RNAs. For example, miR-23, a putative repressor of PGC-1α, is strongly downregulated after 90 min of acute exercise in mouse muscle [55]. In chronically trained and casted mice, the expression of miR-696 and PGC-1α negatively correlated with higher and lower expression of PGC-1α in training and unloading, respectively [56]. The repressive effect of miR-696 on PGC-1α was subsequently confirmed in cultured myocytes. Furthermore, the presence of an upstream open reading frame (uORF) in the 5′ untranslated region of PGC-1α mediates translational repression in an evolutionary conserved manner [57]. Absence of a functional uORF in the genome of the Atlantic bluefin tuna correlates with high abundance of muscle mitochondria, slow-twitch, oxidative muscle fibers, and an exceptionally high endurance.

In addition to the effects on PGC-1α gene expression, epigenetic mechanisms are involved in modulating the activity of the PGC-1α protein in this tissue. For example, the coactivation of the nuclear receptor estrogen-related receptor α (ERRα) by PGC-1α correlates with the relative GC and CpG content of ERRα binding sites in PGC-1α target genes, implying a potential role of DNA methylation in controlling the interaction between these two partners in the regulation of PGC-1α-dependent metabolic gene expression [58]. Second, as described above, by recruiting HAT, mediator and SWI/SNF protein complexes, PGC-1α promotes various epigenetic changes to regulate a complex transcriptional network [59]. Then, the nuclear receptor corepressor 1 (NCoR1) competes with PGC-1α for binding to ERRα, and represses PGC-1α target gene expression by recruiting HDAC complexes to the respective regulatory sites [60]. Finally, the activity of PGC-1α is activated and repressed by deacetylation by sirtuin 1 (SIRT1) and acetylation by K(lysine) acetyltransferase 2A (Kat2a/Gcn5) [61], which are also involved in the acetylation and, in the case of Kat2a, succinylation of histones. However, whether and how posttranslational modifications of PGC-1α and histones by these enzymes are coordinated is unknown. Of note, while many of these mechanisms up- and downstream of PGC-1α have been studied and described in skeletal muscle, they might also be important for PGC-1α action in other tissues.

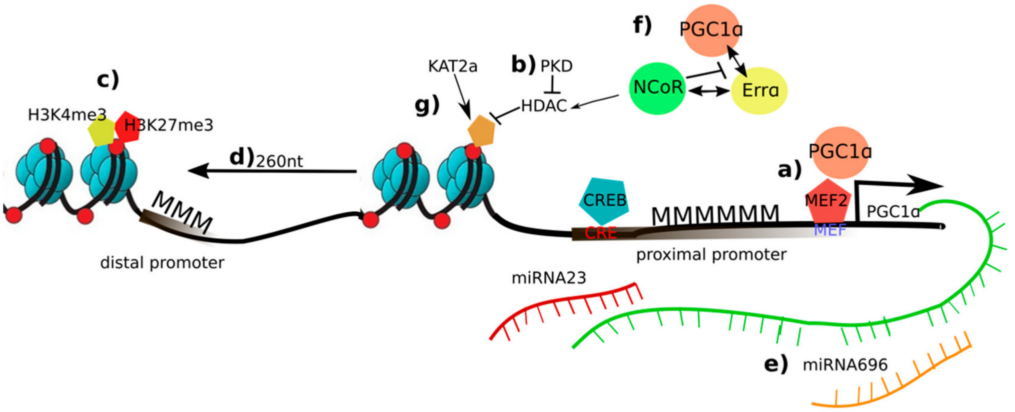


Figure 1. Overview of epigenetic changes on the peroxisome proliferator-activated receptor γ coactivator 1α (PGC-1α) in skeletal muscle and exercise: a) In an inactive state, the promoter of PGC-1α is methylated (MMM). PGC-1α induces its own transcription in a positive autoregulatory loop by coactivating the myocyte enhancer factor 2 (MEF2); b) Protein kinase d (PKD) represses histone deacetylase (HDAC) and retains the acetylation marks and elevation of PGC-1α transcription; (c) a combination of trimethylation of histone 3 at lysine 4 (H3K4me3) and H3K27me3 is deposited at the distal promoter of PGC-1α suggesting a fast switch of gene programs if necessary; d) nucleosome repositioning enhances PGC-1α transcription; e) Micro RNA (miR)-696 and miR-23 are putative repressors of PGC-1α; f) NCoR1 competes with PGC-1α for binding to estrogen-related receptor α (ERRα), to repress PGC-1α target gene expression; g) the activity of PGC-1α is activated and repressed by deacetylation by sirtuin 1 (SIRT1) and acetylation by K(lysine) acetyltransferase 2a (KAT2a).

4.2. *Brown Adipose Tissue and Thermogenesis*

Numerous studies with gain- and loss-of-function have underlined the central role of PGC-1α in controlling non-shivering thermogenesis in brown adipose tissue (Figure 2) [62]. Besides creatine cycling, mitochondrial uncoupling is the major mechanism by which thermogenesis in brown adipose tissue is achieved. Upon stimulation by β-adrenergic signaling, the expression and activity of the uncoupling protein 1 (UCP-1) is upregulated, which then produces heat by uncoupling the proton gradient across the inner mitochondrial membrane from ATP production [63]. PGC-1α gene expression is stimulated by β-adrenergic signaling in brown adipocytes, and PGC-1α subsequently coactivates PPARγ and recruits SRC-1/p300 in regulatory elements of the UCP-1 gene to induce transcription [8,42].

The regulation of PGC-1α gene expression in this context is mediated by different mechanisms. First, the transcription factor ATF-2 is recruited to cAMP-responsive elements (CRE) in the PGC-1α promoter upon phosphorylation by the p38 mitogen-activated protein kinase [64]. Second, in response to β-adrenergic signaling, HDAC1 association with the CRE element in the PGC-1α promoter is reduced and replaced by binding of the H3K27 lysine-specific demethylase 6A (KDM6A) together with the HAT CBP, leading to lower methylation and higher acetylation of H3K27 and subsequently enhanced PGC-1α gene expression [65].

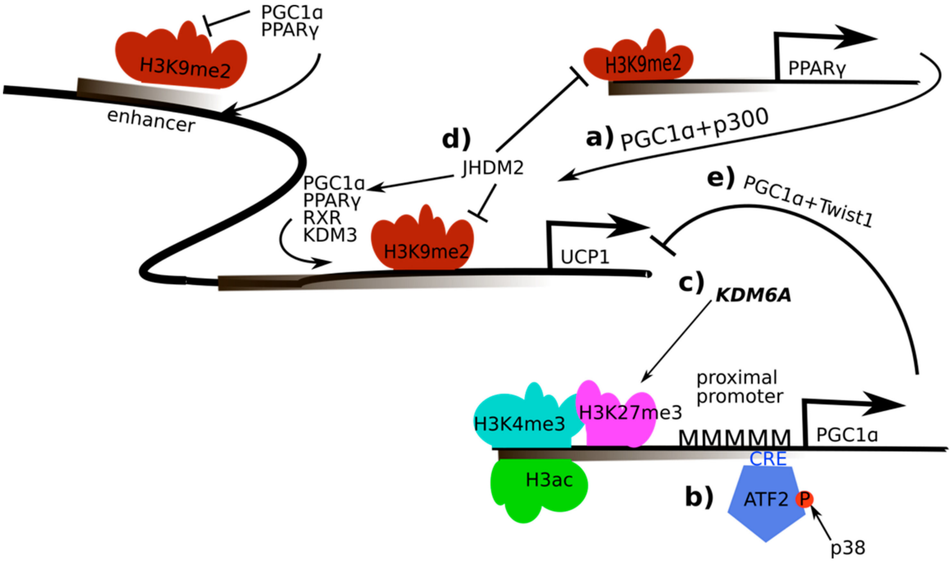


Figure 2. Regulation and activity of PGC-1α in the regulation of UCP-1 in brown adipose tissue and thermogenesis: a) PGC-1α recruits peroxisome proliferator-activated receptor γ (PPARγ) and sterol-receptor coactivator 1/E1A binding protein (SRC-1/p300) to regulatory elements of the uncoupling protein 1 (*UCP-1*) gene; b) AMP-dependent transcription factor (ATF-2) is recruited to cAMP response element (CRE) elements in the PGC-1α promoter upon phosphorylation by the p38 mitogen-activated protein kinase which enables PGC-1α transcription; c) Histone 3 lysine 27 (H3K27) is demethylated by H3K27 lysine-specific demethylase 6A (KDM6A), higher acetylation of H3K27 subsequently leads to enhanced PGC-1α gene expression; d) interaction of PGC-1α with the JmjC domain-containing histone demethylase 2 (JHDM2) affects the recruitment of the PPARγ complex containing retinoid X receptor α (RXRα), PGC-1α, p300 and SRC-1 to the PPAR-response elements in the UCP-1 promoter; e) interaction of twist-related protein 1 (TWIST1) and PGC-1α represses UCP-1 expression.

In addition to the regulation of PGC-1α gene expression in brown adipocytes, different epigenetic mechanisms have been implied in the PGC-1α-dependent regulation of UCP-1 expression in thermogenesis [62]. First, PGC-1α interacts with the H3K9 JmjC domain-containing histone demethylase 2 (JHDM2), which affects the recruitment of the PPARγ complex containing the heterodimerization partner retinoid X receptor α (RXRα), PGC-1α, p300 and SRC-1 to the PPAR-response elements in the UCP-1 promoter [66]. Consistently, JHDM2 knockout mice accumulate fat in adulthood and fail to adapt to cold exposure, lacking adequate regulation of UCP-1 in brown fat tissue. PGC-1α-mediated induction of UCP-1 is also influenced by the twist-related protein 1 (TWIST1) [67]. While both proteins are recruited to the UCP-1 promoter, TWIST1 associates with HDAC5, reduces PGC-1α-induced histone 3 acetylation and thereby represses the expression of UCP-1 and other target genes of PGC-1α. Interestingly, TWIST1 transcription is positively regulated by PPARβ/δ, a transcription factor binding partner for PGC-1α in the control of mitochondrial and other metabolic genes, and thereby exerts a negative feedback loop on PGC-1α activity in brown adipose tissue.

5. PGC-1α and Epigenetic Mechanisms in Disease

Many diseases are characterized by wide-spread epigenetic changes that could either contribute to, or be a consequence of the pathological changes [68]. Similarly, dysregulation of mitochondria is observed in numerous pathologies, often associated with changes in PGC-1α expression and/or activity [69]. In the following sections, we have therefore summarized the current knowledge about epigenetic mechanisms that control PGC-1α in different diseases (Figure 3).

5.1. Obesity

In skeletal muscle, obesity results in an altered gene expression profile that is associated with wide-spread changes in DNA methylation events [13]. As one of these genes, the promoter of PGC-1α is hypermethylated in obese subjects, and the methylation pattern is restored after gastric bypass surgery, comparable to that observed in lean individuals. Similar methylation changes of almost half of the CpG sites in the PGC-1α promoter could be triggered by short-term overfeeding of young, healthy men with a high fat diet in skeletal muscle [70], or of low-birthweight individuals in white adipose tissue [70]. In the latter cohort, PGC-1α gene expression was restored after insulin injection. Changes in the methylation status of the PGC-1α promoter were furthermore described in cultured human primary myocytes exposed to fatty acids, in a DNMT3B-dependent manner [11]. A link between fatty acid oxidation and PGC-1α promoter methylation was likewise proposed by the effect of decreased flavine adenine dinucleotide (FAD) levels leading to a loss of histone 3 acetylation and H3K3me2/3 deposition near the PGC-1α gene [71]. Of note, methylation of four specific CpG loci in the PGC-1α promoter in blood of children was predictive of adiposity later in life, independent of sex, age, pubertal timing, and activity [72].

5.2. Type II Diabetes

Hypermethylation of non-CpG sites at the PGC-1α promoter negatively correlated with PGC-1α expression in skeletal muscle of type 2 diabetic subjects compared to glucose-tolerant individuals [11]. This reduction was linked to DNMT3b activity in cultured myotubes treated with tumor necrosis factor α (TNFα) or free fatty acids, both leading to hypermethylation of the PGC-1α promoter. In particular, the methylation site at –260 nucleotide location was responsible for the transcriptional repression in that context. Moreover, a study in monozygotic twins showed higher methylation levels in the PGC-1α promoter in skeletal muscle and adipose tissue in type 2 diabetic subjects [73]. Similarly, a 2-fold increase in PGC-1α promoter methylation was described in human pancreatic islet cells of type 2 diabetic individuals compared to normal individuals [74]. Finally, placental PGC-1α promoter methylation correlated both with maternal hyperglycemia and newborn leptin levels [75].

5.3. Non-Alcoholic Fatty Liver Disease (NAFLD)

A comprehensive DNA methylation profiling of liver biopsies of morbidly obese patients with NAFLD revealed broad changes in the methylation pattern compared to healthy individuals [76]. Motif prediction implied an enrichment in methylation changes in DNA regions of PGC-1α recruitment. Moreover, bariatric surgery reversed some of the NAFLD-associated methylation changes, with a high enrichment of predicted binding sites for ERRα, a strong interaction partner for PGC-1α. However, whether methylation changes modifying predicted PGC-1α and ERRα recruitment sites really contribute to the degree of NAFLD remains to be shown. In line with this hypothesis, NAFLD-related insulin resistance is correlated positively with PGC-1α promoter methylation, and negatively with PGC-1α gene expression [77].

5.4. Parkinson's Disease

Adequate PGC-1α levels are indispensable for mitochondrial activity in the brain, and loss-of-function of PGC-1α promotes neurodegenerative events in this organ [78,79]. In an extensive study incorporating

322 samples from the brain and 88 samples from blood, non-canonical cytosine methylation of the PGC-1α gene was found to be significantly increased in Parkinson's patients compared to controls [80]. In line, treatment of mouse primary cortical neurons, microglia and astrocytes with palmitate caused PGC-1α promoter methylation at non-canonical cytosines. Likewise, the intracerebroventricular injection of palmitate into mice with transgenic expression of human α-synuclein triggered increased PGC-1α promoter methylation, reduced expression of PGC-1α and diminished the mitochondrial number in the substantia nigra. Moreover, PGC-1α promoter methylation correlated with increased endoplasmatic reticulum (ER) stress and inflammatory signaling.

5.5. Kidney Diseases

The lncRNA taurine-upregulated gene 1 (Tug1) interacts with PGC-1α in the kidney, and promotes the binding of PGC-1α to its own promoter [81]. Activation of this mechanism in podocytes improves mitochondrial function and reduces apoptosis as well as endoplasmic reticulum stress in diabetic nephropathy [81,82]. In acute kidney injury, the TNF-related weak inducer of apoptosis (TWEAK) stimulates HDAC recruitment to nuclear factor κB (NF-κB) on the PGC-1α promoter, resulting in histone deacetylation and repression of PGC-1α gene transcription [83]. Thereby, an inflammatory response is boosted while mitochondrial function is repressed in this pathological context.

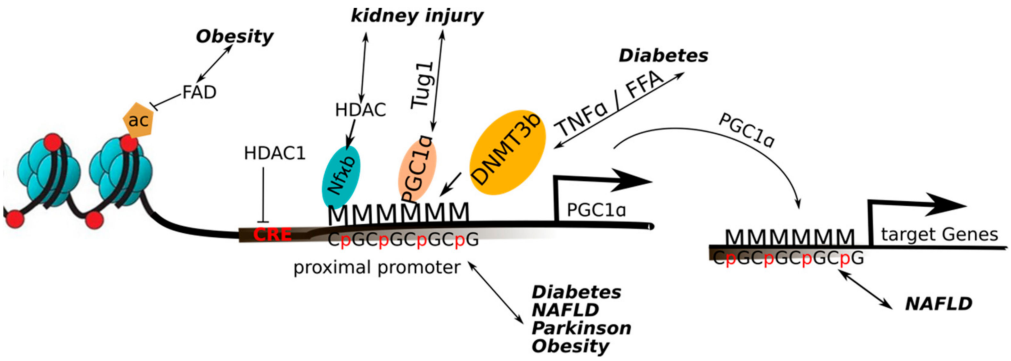


Figure 3. Overview of the epigenetic changes on the PGC-1α in a pathological context: Increased methylation of the PGC-1α promoter has been found to occur in obesity, diabetes, non-alcoholic fatty liver disease (NAFLD) and Parkinson's disease. Obesity and decreased flavin adenine dinucleotide (FAD) levels lead to a loss of histone 3 acetylation and thus a decreased gene expression of PGC-1α. Exposure to TNFα or FFA (free fatty acids) leads to a hypermethylation of the PGC-1α promoter by the activation of DNA methyltransferase 3b (DNMT3b). In NAFLD, a decreased expression of PGC1α target genes was associated with higher methylation of the respective promoters. In kidney diseases, the micro RNA taurine upregulated gene (TUG1) promotes the binding of PGC-1α to its own promoter. In acute kidney injury, histone deacetylase (HDAC) recruitment to nuclear factor κB (NF-κB) on the PGC-1α promoter promotes deacetylation and thus repression of PGC1α. Increased methylation of the PGC-1α promoter has been found to occur in diabetes, NAFLD and Parkinson's disease.

6. Conclusions and Perspectives

With the inclusion of transient, short-term changes, the traditional distinction between epigenetics and transcriptional regulation becomes blurry. It is thus of little surprise that a strong transcriptional regulator such as PGC-1α is not only controlled by, but also uses various epigenetic mechanisms to modulate complex transcriptional networks in acute settings. The more persistent changes in PGC-1α promoter methylation in numerous diseases however hint at a more long-term control of PGC-1α to be important for health and disease. Future studies will hopefully aim at elucidating these effects not only in the pathological, but also physiological context. For example, even though clear evidence exists, the hereditary aspects of exercise training remain enigmatic [5,84]. Intriguingly, the selection of high- and low-capacity runners of rats demonstrated the heritability of treadmill exercise, and was associated with higher PGC-1α protein levels in the muscles of high- compared to low-capacity runners [85]. It will be interesting to study whether epigenetic regulation of PGC-1α underlies this effect. These and other similar studies will ultimately help to understand cell plasticity over different time scales in health and disease.

Funding: The work in the laboratory of C.H. was funded by the Swiss National Science Foundation, the European Research Council (ERC) Consolidator grant 616830-MUSCLE_NET, Swiss Cancer Research grant KFS-3733-08-2015, the Swiss Society for Research on Muscle Diseases (SSEM), SystemsX.ch, the Novartis Stiftung für Medizinisch-Biologische Forschung and the University of Basel.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Cavalli, G.; Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **2019**, *571*, 489–499. [CrossRef] [PubMed]
2. Huang, B.; Jiang, C.; Zhang, R. Epigenetics: The language of the cell? *Epigenomics* **2014**, *6*, 73–88. [CrossRef] [PubMed]
3. Ling, C.; Ronn, T. Epigenetics in human obesity and type 2 diabetes. *Cell Metab.* **2019**, *29*, 1028–1044. [CrossRef] [PubMed]
4. Maniyadath, B.; Shukla, N.; Kolthur-Seetharam, U. Gene expression, epigenetics and ageing. *Subcell. Biochem.* **2018**, *90*, 471–504. [PubMed]
5. McGee, S.L.; Hargreaves, M. Epigenetics and exercise. *Trends Endocrinol. Metab.* **2019**, *30*, 636–645. [CrossRef] [PubMed]
6. Kupr, B.; Handschin, C. Complex coordination of cell plasticity by a pgc-1alpha-controlled transcriptional network in skeletal muscle. *Front. Physiol.* **2015**, *6*, 325. [CrossRef] [PubMed]
7. Martinez-Redondo, V.; Pettersson, A.T.; Ruas, J.L. The hitchhiker's guide to pgc-1alpha isoform structure and biological functions. *Diabetologia* **2015**, *58*, 1969–1977. [CrossRef]
8. Puigserver, P.; Adelmant, G.; Wu, Z.; Fan, M.; Xu, J.; O'Malley, B.; Spiegelman, B.M. Activation of ppargamma coactivator-1 through transcription factor docking. *Science* **1999**, *286*, 1368–1371. [CrossRef]
9. Wallberg, A.E.; Yamamura, S.; Malik, S.; Spiegelman, B.M.; Roeder, R.G. Coordination of p300-mediated chromatin remodeling and trap/mediator function through coactivator pgc-1alpha. *Mol. Cell* **2003**, *12*, 1137–1149. [CrossRef]
10. Li, S.; Liu, C.; Li, N.; Hao, T.; Han, T.; Hill, D.E.; Vidal, M.; Lin, J.D. Genome-wide coactivation analysis of pgc-1alpha identifies baf60a as a regulator of hepatic lipid metabolism. *Cell Metab.* **2008**, *8*, 105–117. [CrossRef]
11. Barres, R.; Osler, M.E.; Yan, J.; Rune, A.; Fritz, T.; Caidahl, K.; Krook, A.; Zierath, J.R. Non-cpg methylation of the pgc-1alpha promoter through dnmt3b controls mitochondrial density. *Cell Metab.* **2009**, *10*, 189–198. [CrossRef] [PubMed]
12. Barres, R.; Yan, J.; Egan, B.; Treebak, J.T.; Rasmussen, M.; Fritz, T.; Caidahl, K.; Krook, A.; O'Gorman, D.J.; Zierath, J.R. Acute exercise remodels promoter methylation in human skeletal muscle. *Cell Metab.* **2012**, *15*, 405–411. [CrossRef] [PubMed]
13. Barres, R.; Kirchner, H.; Rasmussen, M.; Yan, J.; Kantor, F.R.; Krook, A.; Naslund, E.; Zierath, J.R. Weight loss after gastric bypass surgery in human obesity remodels promoter methylation. *Cell Rep.* **2013**, *3*, 1020–1027. [CrossRef] [PubMed]
14. Hughes, T.R.; Lambert, S.A. Transcription factors read epigenetics. *Science* **2017**, *356*, 489–490. [CrossRef] [PubMed]
15. Wu, C.; Morris, J.R. Genes, genetics, and epigenetics: A correspondence. *Science* **2001**, *293*, 1103–1105. [CrossRef] [PubMed]
16. Dupont, C.; Armant, D.R.; Brenner, C.A. Epigenetics: Definition, mechanisms and clinical perspective. *Semin. Reprod. Med.* **2009**, *27*, 351–357. [CrossRef]
17. Zhou, B.R.; Bai, Y. Chromatin structures condensed by linker histones. *Essays Biochem.* **2019**, *63*, 75–87.
18. Wang, Y.; Yuan, Q.; Xie, L. Histone modifications in aging: The underlying mechanisms and implications. *Curr. Stem Cell Res. Ther.* **2018**, *13*, 125–135. [CrossRef]

19. Molina-Serrano, D.; Kyriakou, D.; Kirmizis, A. Histone modifications as an intersection between diet and longevity. *Front. Genet.* **2019**, *10*, 192. [CrossRef]
20. Barnes, C.E.; English, D.M.; Cowley, S.M. Acetylation & co: An expanding repertoire of histone acylations regulates chromatin and transcription. *Essays Biochem.* **2019**, *63*, 97–107.
21. Rose, N.R.; Klose, R.J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim. Biophys. Acta* **2014**, *1839*, 1362–1372. [CrossRef] [PubMed]
22. Hyun, K.; Jeon, J.; Park, K.; Kim, J. Writing, erasing and reading histone lysine methylations. *Exp. Mol. Med.* **2017**, *49*, e324. [CrossRef] [PubMed]
23. Dong, X.; Weng, Z. The correlation between histone modifications and gene expression. *Epigenomics* **2013**, *5*, 113–116. [CrossRef] [PubMed]
24. Mueller-Planitz, F.; Klinker, H.; Becker, P.B. Nucleosome sliding mechanisms: New twists in a looped history. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1026–1032. [CrossRef] [PubMed]
25. Bowman, G.D. Mechanisms of atp-dependent nucleosome sliding. *Curr. Opin. Struct. Biol.* **2010**, *20*, 73–81. [CrossRef] [PubMed]
26. Greenberg, M.V.C.; Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 590–607. [CrossRef] [PubMed]
27. Schmitz, R.J.; Lewis, Z.A.; Goll, M.G. DNA methylation: Shared and divergent features across eukaryotes. *Trends Genet.* **2019**. [CrossRef]
28. Richa, R.; Sinha, R.P. Hydroxymethylation of DNA: An epigenetic marker. *EXCLI J.* **2014**, *13*, 592–610.
29. Iyer, L.M.; Zhang, D.; Aravind, L. Adenine methylation in eukaryotes: Apprehending the complex evolutionary history and functional potential of an epigenetic modification. *Bioessays* **2016**, *38*, 27–40. [CrossRef]
30. Castillo-Aguilera, O.; Depreux, P.; Halby, L.; Arimondo, P.B.; Goossens, L. DNA methylation targeting: The dnmt/hmt crosstalk challenge. *Biomolecules* **2017**, *7*, 3. [CrossRef]
31. Williams, K.; Christensen, J.; Helin, K. DNA methylation: Tet proteins-guardians of cpg islands? *EMBO Rep.* **2011**, *13*, 28–35. [CrossRef] [PubMed]
32. Tammen, S.A.; Friso, S.; Choi, S.W. Epigenetics: The link between nature and nurture. *Mol. Asp. Med.* **2013**, *34*, 753–764. [CrossRef] [PubMed]
33. Pasquinelli, A.E. Micrornas: Heralds of the noncoding rna revolution. *RNA* **2015**, *21*, 709–710. [CrossRef] [PubMed]
34. Kopp, F.; Mendell, J.T. Functional classification and experimental dissection of long noncoding rnas. *Cell* **2018**, *172*, 393–407. [CrossRef]
35. Barile, L.; Vassalli, G. Exosomes: Therapy delivery tools and biomarkers of diseases. *Pharmacol. Ther.* **2017**, *174*, 63–78. [CrossRef]
36. Bao, N.; Lye, K.W.; Barton, M.K. Microrna binding sites in arabidopsis class iii hd-zip mrnas are required for methylation of the template chromosome. *Dev. Cell* **2004**, *7*, 653–662. [CrossRef]
37. Rajewsky, N. Microrna target predictions in animals. *Nat. Genet.* **2006**, *38* (Suppl. 6), S8–S13. [CrossRef]
38. Tuddenham, L.; Wheeler, G.; Ntounia-Fousara, S.; Waters, J.; Hajihosseini, M.K.; Clark, I.; Dalmay, T. The cartilage specific microrna-140 targets histone deacetylase 4 in mouse cells. *FEBS Lett.* **2006**, *580*, 4214–4217. [CrossRef]
39. Alvarez-Saavedra, M.; Antoun, G.; Yanagiya, A.; Oliva-Hernandez, R.; Cornejo-Palma, D.; Perez-Iratxeta, C.; Sonenberg, N.; Cheng, H.Y. Mirna-132 orchestrates chromatin remodeling and translational control of the circadian clock. *Hum. Mol. Genet.* **2011**, *20*, 731–751. [CrossRef]
40. Shi, H.; Wei, J.; He, C. Where, when, and how: Context-dependent functions of rna methylation writers, readers, and erasers. *Mol. Cell* **2019**, *74*, 640–650. [CrossRef]
41. Zhao, X.; Yang, Y.; Sun, B.F.; Zhao, Y.L.; Yang, Y.G. Fto and obesity: Mechanisms of association. *Curr. Diabetes Rep.* **2014**, *14*, 486. [CrossRef] [PubMed]
42. Puigserver, P.; Wu, Z.; Park, C.W.; Graves, R.; Wright, M.; Spiegelman, B.M. A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. *Cell* **1998**, *92*, 829–839. [CrossRef]
43. Vandenbeek, R.; Khan, N.P.; Estall, J.L. Linking metabolic disease with the pgc-1alpha gly482ser polymorphism. *Endocrinology* **2018**, *159*, 853–865. [CrossRef] [PubMed]
44. Lin, J.; Handschin, C.; Spiegelman, B.M. Metabolic control through the pgc-1 family of transcription coactivators. *Cell Metab.* **2005**, *1*, 361–370. [CrossRef]

45. Cheng, C.F.; Ku, H.C.; Lin, H. Pgc-1alpha as a pivotal factor in lipid and metabolic regulation. *Int. J. Mol. Sci.* **2018**, *19*, 3447. [CrossRef]
46. Handschin, C. The biology of pgc-1alpha and its therapeutic potential. *Trends Pharmacol. Sci.* **2009**, *30*, 322–329. [CrossRef]
47. Mootha, V.K.; Handschin, C.; Arlow, D.; Xie, X.; St Pierre, J.; Sihag, S.; Yang, W.; Altshuler, D.; Puigserver, P.; Patterson, N.; et al. Erralpha and gabpa/b specify pgc-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 6570–6575. [CrossRef]
48. Wu, Z.; Puigserver, P.; Andersson, U.; Zhang, C.; Adelmant, G.; Mootha, V.; Troy, A.; Cinti, S.; Lowell, B.; Scarpulla, R.C.; et al. Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator pgc-1. *Cell* **1999**, *98*, 115–124. [CrossRef]
49. Handschin, C.; Rhee, J.; Lin, J.; Tarr, P.T.; Spiegelman, B.M. An autoregulatory loop controls peroxisome proliferator-activated receptor gamma coactivator 1alpha expression in muscle. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 7111–7116. [CrossRef]
50. Czubryt, M.P.; McAnally, J.; Fishman, G.I.; Olson, E.N. Regulation of peroxisome proliferator-activated receptor gamma coactivator 1 alpha (pgc-1 alpha) and mitochondrial function by mef2 and hdac5. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1711–1716. [CrossRef]
51. Akimoto, T.; Li, P.; Yan, Z. Functional interaction of regulatory factors with the pgc-1alpha promoter in response to exercise by in vivo imaging. *Am. J. Physiol. Cell Physiol.* **2008**, *295*, C288–C292. [CrossRef] [PubMed]
52. Lochmann, T.L.; Thomas, R.R.; Bennett, J.P., Jr.; Taylor, S.M. Epigenetic modifications of the pgc-1alpha promoter during exercise induced expression in mice. *PLoS ONE* **2015**, *10*, e0129647. [CrossRef] [PubMed]
53. Masuzawa, R.; Konno, R.; Ohsawa, I.; Watanabe, A.; Kawano, F. Muscle type-specific rna polymerase ii recruitment during pgc-1alpha gene transcription after acute exercise in adult rats. *J. Appl. Physiol.* **2018**, *125*, 1238–1245. [CrossRef] [PubMed]
54. Bajpeyi, S.; Covington, J.D.; Taylor, E.M.; Stewart, L.K.; Galgani, J.E.; Henagan, T.M. Skeletal muscle pgc1alpha-1 nucleosome position and -260 nt DNA methylation determine exercise response and prevent ectopic lipid accumulation in men. *Endocrinology* **2017**, *158*, 2190–2199. [CrossRef] [PubMed]
55. Safdar, A.; Abadi, A.; Akhtar, M.; Hettinga, B.P.; Tarnopolsky, M.A. Mirna in the regulation of skeletal muscle adaptation to acute endurance exercise in c57bl/6j male mice. *PLoS ONE* **2009**, *4*, e5610. [CrossRef]
56. Aoi, W.; Naito, Y.; Mizushima, K.; Takanami, Y.; Kawai, Y.; Ichikawa, H.; Yoshikawa, T. The microrna mir-696 regulates pgc-1{alpha} in mouse skeletal muscle in response to physical activity. *Am. J. Physiol. Endocrinol. Metab.* **2010**, *298*, E799–E806. [CrossRef]
57. Dumesic, P.A.; Egan, D.F.; Gut, P.; Tran, M.T.; Parisi, A.; Chatterjee, N.; Jedrychowski, M.; Paschini, M.; Kazak, L.; Wilensky, S.E.; et al. An evolutionarily conserved uorf regulates pgc1alpha and oxidative metabolism in mice, flies, and bluefin tuna. *Cell Metab.* **2019**, *30*, 190–200. [CrossRef]
58. Salatino, S.; Kupr, B.; Baresic, M.; Omidi, S.; van Nimwegen, E.; Handschin, C. The genomic context and corecruitment of sp1 affect erralpha coactivation by pgc-1alpha in muscle cells. *Mol. Endocrinol.* **2016**, *30*, 809–825. [CrossRef]
59. Baresic, M.; Salatino, S.; Kupr, B.; van Nimwegen, E.; Handschin, C. Transcriptional network analysis in muscle reveals ap-1 as a partner of pgc-1alpha in the regulation of the hypoxic gene program. *Mol. Cell. Biol.* **2014**, *34*, 2996–3012. [CrossRef]
60. Perez-Schindler, J.; Summermatter, S.; Salatino, S.; Zorzato, F.; Beer, M.; Balwierz, P.J.; van Nimwegen, E.; Feige, J.N.; Auwerx, J.; Handschin, C. The corepressor ncor1 antagonizes pgc-1alpha and estrogen-related receptor alpha in the regulation of skeletal muscle function and oxidative metabolism. *Mol. Cell. Biol.* **2012**, *32*, 4913–4924. [CrossRef]
61. Dominy, J.E., Jr.; Lee, Y.; Gerhart-Hines, Z.; Puigserver, P. Nutrient-dependent regulation of pgc-1alpha's acetylation state and metabolic function through the enzymatic activities of sirt1/gcn5. *Biochim. Biophys. Acta* **2010**, *1804*, 1676–1683. [CrossRef] [PubMed]
62. Gill, J.A.; La Merrill, M.A. An emerging role for epigenetic regulation of pgc-1alpha expression in environmentally stimulated brown adipose thermogenesis. *Environ. Epigenet.* **2017**, *3*, dvx009. [CrossRef] [PubMed]
63. Villarroya, F.; Peyrou, M.; Giralt, M. Transcriptional regulation of the uncoupling protein-1 gene. *Biochimie* **2017**, *134*, 86–92. [CrossRef]

64. Cao, W.; Daniel, K.W.; Robidoux, J.; Puigserver, P.; Medvedev, A.V.; Bai, X.; Floering, L.M.; Spiegelman, B.M.; Collins, S. P38 mitogen-activated protein kinase is the central regulator of cyclic amp-dependent transcription of the brown fat uncoupling protein 1 gene. *Mol. Cell. Biol.* **2004**, *24*, 3057–3067. [CrossRef] [PubMed]
65. Galmozzi, A.; Mitro, N.; Ferrari, A.; Gers, E.; Gilardi, F.; Godio, C.; Cermenati, G.; Gualerzi, A.; Donetti, E.; Rotili, D.; et al. Inhibition of class i histone deacetylases unveils a mitochondrial signature and enhances oxidative metabolism in skeletal muscle and adipose tissue. *Diabetes* **2013**, *62*, 732–742. [CrossRef]
66. Tateishi, K.; Okada, Y.; Kallin, E.M.; Zhang, Y. Role of jhdm2a in regulating metabolic gene expression and obesity resistance. *Nature* **2009**, *458*, 757–761. [CrossRef]
67. Pan, D.; Fujimoto, M.; Lopes, A.; Wang, Y.X. Twist-1 is a ppardelta-inducible, negative-feedback regulator of pgc-1alpha in brown fat metabolism. *Cell* **2009**, *137*, 73–86. [CrossRef]
68. Portela, A.; Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol.* **2010**, *28*, 1057–1068. [CrossRef]
69. Sorrentino, V.; Menzies, K.J.; Auwerx, J. Repairing mitochondrial dysfunction in disease. *Annu. Rev. Pharmacol. Toxicol.* **2018**, *58*, 353–389. [CrossRef]
70. Jacobsen, S.C.; Brons, C.; Bork-Jensen, J.; Ribel-Madsen, R.; Yang, B.; Lara, E.; Hall, E.; Calvanese, V.; Nilsson, E.; Jorgensen, S.W.; et al. Effects of short-term high-fat overfeeding on genome-wide DNA methylation in the skeletal muscle of healthy young men. *Diabetologia* **2012**, *55*, 3341–3349. [CrossRef]
71. Hino, S.; Sakamoto, A.; Nagaoka, K.; Anan, K.; Wang, Y.; Mimasu, S.; Umehara, T.; Yokoyama, S.; Kosai, K.; Nakao, M. Fad-dependent lysine-specific demethylase-1 regulates cellular energy expenditure. *Nat. Commun.* **2012**, *3*, 758. [CrossRef] [PubMed]
72. Clarke-Harris, R.; Wilkin, T.J.; Hosking, J.; Pinkney, J.; Jeffery, A.N.; Metcalf, B.S.; Godfrey, K.M.; Voss, L.D.; Lillycrop, K.A.; Burdge, G.C. Pgc1alpha promoter methylation in blood at 5-7 years predicts adiposity from 9 to 14 years (earlybird 50). *Diabetes* **2014**, *63*, 2528–2537. [CrossRef] [PubMed]
73. Ribel-Madsen, R.; Fraga, M.F.; Jacobsen, S.; Bork-Jensen, J.; Lara, E.; Calvanese, V.; Fernandez, A.F.; Friedrichsen, M.; Vind, B.F.; Hojlund, K.; et al. Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes. *PLoS ONE* **2012**, *7*, e51302. [CrossRef] [PubMed]
74. Ling, C.; Del Guerra, S.; Lupi, R.; Ronn, T.; Granhall, C.; Luthman, H.; Masiello, P.; Marchetti, P.; Groop, L.; Del Prato, S. Epigenetic regulation of ppargc1a in human type 2 diabetic islets and effect on insulin secretion. *Diabetologia* **2008**, *51*, 615–622. [CrossRef] [PubMed]
75. Cote, S.; Gagne-Ouellet, V.; Guay, S.P.; Allard, C.; Houde, A.A.; Perron, P.; Baillargeon, J.P.; Gaudet, D.; Guerin, R.; Brisson, D.; et al. Ppargc1alpha gene DNA methylation variations in human placenta mediate the link between maternal hyperglycemia and leptin levels in newborns. *Clin. Epigenet.* **2016**, *8*, 72. [CrossRef] [PubMed]
76. Ahrens, M.; Ammerpohl, O.; von Schonfels, W.; Kolarova, J.; Bens, S.; Itzel, T.; Teufel, A.; Herrmann, A.; Brosch, M.; Hinrichsen, H.; et al. DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery. *Cell Metab.* **2013**, *18*, 296–302. [CrossRef]
77. Sookoian, S.; Rosselli, M.S.; Gemma, C.; Burgueno, A.L.; Fernandez Gianotti, T.; Castano, G.O.; Pirola, C.J. Epigenetic regulation of insulin resistance in nonalcoholic fatty liver disease: Impact of liver methylation of the peroxisome proliferator-activated receptor gamma coactivator 1alpha promoter. *Hepatology* **2010**, *52*, 1992–2000. [CrossRef]
78. Lin, J.; Wu, P.H.; Tarr, P.T.; Lindenberg, K.S.; St-Pierre, J.; Zhang, C.Y.; Mootha, V.K.; Jager, S.; Vianna, C.R.; Reznick, R.M.; et al. Defects in adaptive energy metabolism with cns-linked hyperactivity in pgc-1alpha null mice. *Cell* **2004**, *119*, 121–135. [CrossRef]
79. St-Pierre, J.; Drori, S.; Uldry, M.; Silvaggi, J.M.; Rhee, J.; Jager, S.; Handschin, C.; Zheng, K.; Lin, J.; Yang, W.; et al. Suppression of reactive oxygen species and neurodegeneration by the pgc-1 transcriptional coactivators. *Cell* **2006**, *127*, 397–408. [CrossRef]
80. Su, X.; Chu, Y.; Kordower, J.H.; Li, B.; Cao, H.; Huang, L.; Nishida, M.; Song, L.; Wang, D.; Federoff, H.J. Pgc-1alpha promoter methylation in parkinson's disease. *PLoS ONE* **2015**, *10*, e0134087. [CrossRef]
81. Long, J.; Badal, S.S.; Ye, Z.; Wang, Y.; Ayanga, B.A.; Galvan, D.L.; Green, N.H.; Chang, B.H.; Overbeek, P.A.; Danesh, F.R. Long noncoding rna tug1 regulates mitochondrial bioenergetics in diabetic nephropathy. *J. Clin. Investig.* **2016**, *126*, 4205–4218. [CrossRef] [PubMed]

82. Shen, H.; Ming, Y.; Xu, C.; Xu, Y.; Zhao, S.; Zhang, Q. Deregulation of long noncoding rna (tug1) contributes to excessive podocytes apoptosis by activating endoplasmic reticulum stress in the development of diabetic nephropathy. *J. Cell. Physiol.* **2019**. [CrossRef] [PubMed]
83. Ruiz-Andres, O.; Suarez-Alvarez, B.; Sanchez-Ramos, C.; Monsalve, M.; Sanchez-Nino, M.D.; Ruiz-Ortega, M.; Egido, J.; Ortiz, A.; Sanz, A.B. The inflammatory cytokine tweak decreases pgc-1alpha expression and mitochondrial function in acute kidney injury. *Kidney Int.* **2016**, *89*, 399–410. [CrossRef] [PubMed]
84. McGee, S.L.; Walder, K.R. Exercise and the skeletal muscle epigenome. *Cold Spring Harb. Perspect. Med.* **2017**, *7*, a029876. [CrossRef]
85. Wisloff, U.; Najjar, S.M.; Ellingsen, O.; Haram, P.M.; Swoap, S.; Al-Share, Q.; Fernstrom, M.; Rezaei, K.; Lee, S.J.; Koch, L.G.; et al. Cardiovascular risk factors emerge after artificial selection for low aerobic capacity. *Science* **2005**, *307*, 418–420. [CrossRef]